Effects of Adaptive Conversational User Interfaces on Enjoyment and Engagement while assessing Wellbeing

Charlotte Eijkelkamp
Supervisors: Willem van der Maden, Garrett Allen, Ujwal Gadiraju, Derek Lomas.
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

A decrease in wellbeing worldwide due to the COVID-19 pandemic called for ways to assess wellbeing in a scalable and adequate manner. Conversational User Interfaces (CUIs) seem suitable, however, applying them optimally in certain contexts remains a challenge. This study aims to find ways to make CUIs more engaging and have a better experience by making them adaptive. A 3x2 between-subjects experiment is designed in which the effects of avatar presence, gender, and an empathic conversational style are researched. A chatbot was created in telegram, and the visual design and conversational style were altered to measure the effects on Questionnaire Experience (QX), Enjoyment, and Empathy. In total 30 participants chatted with a randomly assigned chatbot and filled in a survey about their experiences. There is no statistical preference for avatar presence or conversational style. Male gendered chatbots score higher on QX, but female chatbots are perceived as more empathic when comparing gender.

1 Introduction

As the COVID-19 pandemic negatively affected wellbeing, universities started prioritizing staff and student wellbeing more [1]. Managing wellbeing at big organizations like universities calls for an accessible and scalable approach. Part of prioritizing is properly assessing wellbeing, as this can give a foundation for further actions tailored to the needs that come forward.

For more than 50 years now, systems with an ability to converse have been used in (mental) health-related contexts, such as ELIZA [2]. Even though ELIZA processed language using pattern matching and therefore had no actual intelligence, some users still thought they were talking to a real human being [2]. As of now, these systems able to mimic human conversations are known as Conversational User Interfaces (CUIs) or Conversational Agents [3]. The technologies used in CUIs are no longer only based on pattern matching but tend to be more AI-oriented [4].

This increased focus on well-being combined with an ongoing advancement in technologies such as Conversational User Interfaces (CUIs) and Artificial Intelligence creates opportunities for improved and scalable wellbeing assessment.

Previous research has shown that Conversational Agents are a suitable tool for assessing wellbeing [5] and that they are a proper way to collect self-reported wellbeing-related data [6]. Potentially beneficial adaptations for mental health support found in literature as yet are user personality [7] and cultural aspects [8]. Next to that [9] argues that Conversational Agents can be used in psychiatric treatment as well. All in all Conversational User Interfaces have proven to be of help in mental health care. However, how to properly use a CUI for wellbeing assessment is yet to be researched. Similarly, finding out what adaptations work best, in this context, for certain demographics is an unfamiliar factor as well.

Consequently, this research aims at finding ways that boost engagement and enjoyment while assessing wellbeing. To provide a foundation for effective wellbeing assessment the following research question is to be assessed:

RQ: *To what extent does adaptability affect enjoyment and engagement while using Conversational User Interfaces for wellbeing assessment?*.

The rest of the paper will be organized as follows. section 2 discusses the existing literature and explains the formed hypotheses. The methodology is found in section 3 and section 4 discusses the results. section 5 is about responsible research. The discussion and conclusion can be found in section 6 and section 7 respectively.

2 Related Work and Hypotheses

2.1 Design categories

A scoping review of virtual health assistants mentions five design categories in which conversational agents can differ. [10]. For chatbots, only three of them are relevant, namely Visual Design, Conversational Styles, and Cultural affiliation.

Visual Design

The presence of an avatar and its' appearance entails Visual Design. Different target groups do have different preferences regarding avatar representation [11]. Younger people seem to care less about the appearance of the chatbot as compared to elderly people [11]. This is likely because younger people are more experienced with technology and are familiar with different representations of conversational agents. However, other research shows that having an avatar, as compared to having only a textual view, positively impacts the effectiveness of a CUI [12].

Since the experiment will be performed with only students, who are experienced with technology and are used to chatting as a primary means of communication the following hypothesis is formed:

H1: *The presence of an avatar will have no effects on questionnaire experience when compared to the absence of an avatar.*

Conversational Styles

Conversational Styles are ways a CUI expresses itself. These are both non-verbal (gestures, facial expressions) or verbal. In terms of conversational styles, positive effects on user satisfaction are found while using empathy [12, 13]. The verbal strategies used for mimicking empathy by [13] are small talk, politeness, acknowledging mood changes and emotional states, and sympathy for less positive feelings. Next to that, the self-disclosure behavior of users is positively affected by the self-disclosing of chatbots [14]. Furthermore, people sharing personal information react more positively to validating responses rather than invalidating responses [15]. All in all, posing an empathic and self-disclosing conversational style seems to have a positive impact on the overall chatting experience. Consequently, the following hypothesis is formed:

H2: *A chatbot with an empathic conversational style, is overall preferred over a chatbot with no empathic conversational style.*

Cultural affiliation

Adapting the CUI culturally is done by changing both the visual and conversational elements. Research shows that there is a slightly bigger likelihood to be persuaded or trusting an agent presenting as the same culture [16]. Next to that, it shows that language has more impact on the perceived culture than other visual cues. Similarly, the appearance of the agent has no impact on feelings of resemblance [17]. Nonetheless [17] still presents that feelings of similarity do have an overall positive impact. So visual cues have fewer effects on perceived culture and thus perceived similarity, than the conversational elements.

Many of these researches are done in a Western context, so this does give a one-sided view. On the other hand, research done in India argues that there is a preference for a more culturally similar, as opposed to a western, CUI in terms of effectiveness [8]. This all combined suggests that matching cultures is more effective if it matches the culture of the current place of residence, rather than matching to certain individuals. I.e. an 'Indian' chatbot is more effective for an Indian citizen than it is for an American citizen with Indian roots. For the latter, an 'American' chatbot will likely provide the same satisfaction and effectiveness.

2.2 Chatbots in other fields

Chatbots are used in various fields and with different levels of advancement. This subsection will not include research on chatbots used for educational purposes. This is because they focus less on creating an optimal engaging chatbot, but rather on boosting effectiveness for acquiring skills.

Marketing

Chatbots are a popular and cheap way to manage customer service. In marketing-related contexts, the authenticity of the chatbot can be beneficial for user engagement in certain parts of the service area [18]. This authenticity is bigger when conversing with a female [18]. Next to authenticity, humanness also plays a role in the acceptance of the chatbot [19]. Chatbots are generally perceived as more humane and warmhearted when they are gendered female [19]. Next to this, female chatbots are generally forgiven more than male chatbots when making mistakes [20].

Many of these outcomes seem to have their origin in stereotypes around gender. Besides the stereotypical traits, like warmheartedness and being more humane, which have positive effects in marketing-related contexts, the female gender is often linked to being a caregiver. Linking this back to wellbeing assessment, it seems that a female chatbot is more likely to be favorable to a male counterpart. Therefore the following hypotheses are formulated:

H3: *Chatting with a female-gendered chatbot will be preferred over a male-gendered chatbot on enjoyment.*

H4: *Chatting with a female-gendered chatbot will be preferred over a male-gendered chatbot on perceived empathy.*

3 Methodology

3.1 Study Design

To test the effects of the different trait changes an in-between study was conducted. This was done to minimize the learning effects of the participating users. As three different avatars and two different conversation styles are used this yields the 3x2 table depicted in table 1

	Non-Empathic	**Empathic**
No gender	NG NE (Control Group)	NG E
Female	F NE	F E
Male	M NE	M E

Table 1: All experiment conditions and corresponding chatbot abbreviations

As seen in table 1 the no-gendered, non-empathic chatbot will function as the control group.

3.2 Chatbot creation

The chatbot is created to work with telegram. As all six chatbots have the same functionality, asking wellbeing questions, all chatbots are based on the same baseline chatbot. This chatbot goes by the name 'MyBot' and has no profile picture. The baseline bot sends the questions based on the *My Wellness Check* [1] survey and will not send any responses other than the questions. The conversation script can be found in appendix A

3.3 Implementing Empathy

Mimicking Empathy is multifaceted, since "empathy can be defined as the ability to identify, understand and react to others' thoughts, feelings, behavior and experiences" [21, p.1365]. So to implement an empathic conversation style, a way to identify and understand and a way to properly react is found. An overview of the identification and reaction creation process can be found in appendix B.

To identify and understand the users' feelings and thoughts the textual input is used. Since the chatbot mostly asks wellbeing questions, answers to these questions give usable input for identification and understanding. Especially numerical input simplifies this process, as this can be used to straightforwardly differentiate between negative, neutral, and positive feelings. For instance a question asking the user to rate their current mood from 0 to 10, provides this numerical input.

After identification, a suitable reaction must be formed. First, the appropriate reaction styles are formulated, secondly, a fitting reaction is made. After a general set-up of empathic reactions, these are tweaked in a few cooperative evaluation cycles. These cooperative evaluation cycles are executed with end-users, which are students from Delft University of Technology (TU Delft). An example of these reactions can be seen in a screenshot of a conversation shown in fig. 1.

Apart from adapting according to the answers to the wellbeing questions, some messages to create personal interest are added. This is done by asking for a bit of personal information and reacting to this. For example, asking someone's

name and introducing themselves or asking for their study progress and commenting on it.

Figure 1: Screenshot of a conversation with a chatbot with an empathic conversational style.

3.4 Mimicking Gender

To mimic a gender, the visual design of the bot is changed. The characteristics that are changed are the name and profile picture of the bot. The names for the female and male were changed to Jane and Jacob respectively. Next to that, the profile pictures were changed into the avatars shown in fig. 2. The avatars are made in a similar style and have similar characteristics to make sure other effects, such as age or race, do not play a role.

3.5 Controlled Experiment

A controlled experiment is executed with end-users to measure the effects on questionnaire experience, enjoyment, and perceived empathy. Participants get assigned a chatbot to talk to at random and after having a conversation with the chatbot the participant is asked to fill out a survey about the chatting experience.

Survey creation

As the effects on three components are tested a survey consisting of three parts was created. To test the questionnaire

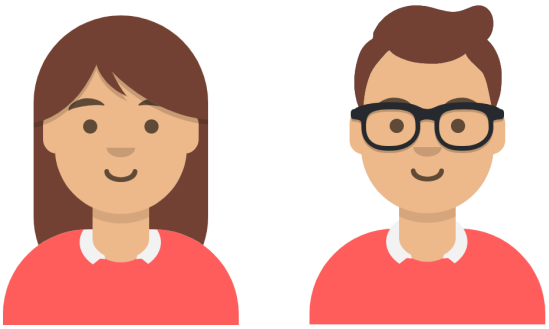

Figure 2: Avatars of Jane (left) and Jacob (right).

experience, five questions are asked, based on the NASA Task Load Index (NASA-TLX) [22] and Questionnaire Experience (QX) [23]. Secondly, three questions are asked about overall enjoyment. Lastly, three questions are asked about the perceived empathy to test if the chatbot successfully mimicked empathy. All questions are answered on a 5-point Likert scale. After question creation, the survey is checked for potential cognitive biases using the Cognitive-Biases-in-Crowdsourcing Checklist [24]. Furthermore, the entire survey can be found in appendix C

Participants

A total of 30 participants (53% female) took part in the experiment. The only inclusion criteria they have to meet is being a student at TU Delft and being able to speak English. Exclusion criteria are inabilities to read or type.

4 Results

Each participant got assigned a chatbot at random. The amount of participants chatting with a certain chatbot is displayed in table 2

Gender	Conversational Style	Participants
None	Non-empathic	5
None	Empathic	6
Female	Non-empathic	5
Female	Empathic	5
Male	Non-Empathic	4
Male	Empathic	5

Table 2: Division of the participants amongst the chatbots

The answers to the multiple-choice questions are converted into numerical values. The most positive option, depending on the phrasing of the question either strongly disagree or strongly agree, was mapped to 5. The most negative option was mapped to 1. As a result, values below 3 are on the negative end, and values above 3 are on the positive end.

The data is analyzed with a two-way ANOVA test with significance level $\alpha = 0.05$. Acording to this test, all results are not statistically significant (p>$0,05$), apart from the effect of gender on QX ($p = 0,005$). The mean and standard deviation of all questions in a certain category, for every experi-

ment condition, are shown in table 3. Overall the results are mainly positive, with two negative outliers, the NG E chatbot on enjoyment and the empathy of the M NE chatbot.

	QX*		Enjoyment		Empathy	
Condition**	**M**	**SD**	**M**	**SD**	**M**	**SD**
NG NE	4,16	0,34	3,27	0,65	3,4	0,68
NG E	4,07	0,41	2,89	0,97	3,61	0,52
F NE	3,84	0,34	3,33	0,47	3,6	0,39
F E	3,96	0,53	3,33	0,7	3,73	0,77
M NE	4,65	0,26	3,25	0,36	2,58	0,6
M E	4,64	0,45	3,87	0,54	3,67	0,7

*Questionnaire Experience.
**Gender: No gender (NG), Female (F) and Male (M). Conversational Style: Non-empathic (NE) and Empathic (E)

Table 3: Mean and Standard Deviation for all dependent variables and each chatbot condition

4.1 Questionnaire Experience

The hypotheses regarding the effects on QX concern the presence of an avatar and conversational style (H1 and H2). Chatbots with an avatar present are all female and male-gendered chatbots (F NE, F E, M NE, M E). In table 4 the means and standard deviations for the chatbots with an avatar present are shown. Overall the chatbots with avatars seem to have a slight preference over the chatbots without avatars. However, the means for QX vary very little, with (0,04) for a non-empathic conversational style, and a bit more for an empathic conversational style (0,23). Comparing the QX for empathic and non-empathic conversational styles yields hardly any difference for the male chatbot (0,01) and a small difference for the female (0,12) and no-gendered (0,08) chatbot. The female chatbot is the only version in which the empathic conversational style is preferred over the non-empathic conversational style.

	QX*		Enjoyment		Empathy	
Condition**	**M**	**SD**	**M**	**SD**	**M**	**SD**
NG-NE	4,16	0,34	3,27	0,65	3,4	0,68
NG-E	4,07	0,41	2,89	0,97	3,61	0,52
A NE	4,2	0,51	3,3	0,43	3,15	0,7
A E	4,3	0,59	3,6	0,68	3,7	0,74

*Questionnaire Experience.
**No gender and absence of avatar (NG), presence of Avatar (A). Conversational Style: Non-empathic (NE) and Empathic (E)

Table 4: Means and standard deviation for all dependent variables and the chatbots without and with avatar

4.2 Enjoyment

For enjoyment, the hypotheses concern gender and conversational style (H2 and H3). Comparing the female and male chatbots shows the following. The female non-empathic chatbot is slightly preferred over the male, non-empathic chatbot (0,08). However, the male empathic chatbot gets preferred over the female, empathic chatbot (0,54). Comparing conversational styles shows us a preference for the non-empathic conversational style for the non-gendered chatbots (0,38), no preference for the female chatbots, and a preference for the empathic conversational style for the male chatbots (0,62).

4.3 Empathy

The experiment shows the following effects on empathy for changing the gender and conversational style (H2 and H4). The female chatbots are preferred over the male chatbots, for both non-empathic (1,03) and empathic (0,06) conversational styles. At all times the empathic conversational style is rated as more empathic than the non-empathic chatbots. This difference is biggest for the male chatbots (1,09) and slightly smaller for the non-gendered (0,21) and female (0,13) chatbots.

5 Responsible Research

As research integrity is valued highly, this section will reflect on the ethical aspects of the research and the research process. First, the ethical implications of this research will be assessed, to make the reader aware of the negative impacts of not using the outcomes ethically. Secondly, the research process is assessed to view that it is in line with the Netherlands Code of Conduct for Research Integrity [25] and the TU Delft vision on integrity [26].

5.1 Ethical Implications

Finding ways to properly assess wellbeing digitally possibly includes recognizing mood or feelings. If these are recognized correctly, this information must be only used for health-related purposes. The users are potentially very vulnerable and their mental state or their current emotion should never be used for commercial purposes.

Furthermore, as discussed in section 2, female chatbots are usually preferred over male chatbots, likely because of social biases. Feeding these gender cues is debatable. However, literature proposes that "it is ethically permissible to insert gender cues into ECA design as long as those cues do not spread a discriminatory vision of gender dynamics" [27, p. 2].

5.2 Research Process

Reproducibility

"Generating verifiable knowledge has long been scientific discovery's central goal" [28, p. 8]. To verify results it is important that this research can be reproduced and will yield the same answers. A few important factors that contribute to irreproducible research are selective reporting, unavailable methods, poor experimental design, and fraud. Therefore the chatbot scripts and the survey questions are included for transparency, also linking to the Mertonian norm (CUDO-norms) of Communism.

Interviews and Surveying

A part of the research is done by conducting interviews and surveys. As this involves human Research Subjects it is important to keep human research ethics in mind. Therefore the TU Delft Risk Planning Tool [29] is consulted and it is ensured that the participants do not face any risks while partaking in this experiment. There is no gathering of personal data, so there is no traceability to any of the participants.

Next to that, the questions are not posed with a desired outcome in mind. The participants are chosen with the principle of diversity in mind. The researcher has no personal interest in any outcome of the research, linking back to the norm of Universalism.

6 Discussion and Limitations

This study aims at finding adaptations to make Conversational User Interfaces used for wellbeing assessment more engaging and create better experiences. The conducted experiment shows no statistically significant effects on QX for the presence of an avatar (H1) and no overall improvement for empathic conversational styles (H2). Lastly, there is no statistical ground to claim that female chatbots are perceived as more empathic than their male counterparts (H4), same goes for the effect on enjoyment (H3). While no hypothesis could be confirmed, there is a statistically significant preference for male chatbots over female chatbots when considering QX. A probable cause of some results being statistically insignificant could be the small sample size of the experiment (N=30). Redoing the experiment with either a bigger sample size or fewer different attributes might contribute to more significant results.

Disregarding the significance of the data, the results do not always seem to be in line with the literature. While rapidly emerging advances in AI or Natural Language Processing provide great bases for even more effective CUIs, the expectations of conversing with a chatbot might be higher than it was a decade, or even a few years ago. It is hard to argue whether an article is outdated, or maybe still suitable for certain demographics, e.g. older people being less experienced with using technologies and likely having lower expectations.

Furthermore, [9] argues that there is little standardization in reviewing CUIs. Therefore it can be hard to make comparisons between certain parts of the literature and form hypotheses fitting to this specific demographic and goal; TU Delft students and wellbeing assessment. Next to that CUIs are presented in a lot of different forms and on a lot of different platforms. While for instance CUIs with spoken in- and output systems are considered in the literature review, these systems might not be that comparable to systems with written in- and output, like the chatbots used in this experiment.

6.1 Chatbot gender

In marketing-related contexts, female chatbots are oftentimes preferred over male chatbots [18–20]. This is likely because these bots are perceived as more warm and human, which are traits that are stereotypically associated with the female gender. Performing a two-way ANOVA showed us that there was no ground to confirm any of the hypotheses. However, there are still some remarkable findings concerning gender that are statistically sound. The male chatbots are preferred over the female chatbots on QX ($p = 0,005$) and comparing F NE and M NE using a one-way ANOVA yields a preference for F NE ($p = 0,022$). As the effect on QX was only tested for avatar presence, the samples with the male and female chatbots are aggregated into one category. This aggregation disregarded the differences between the female and male versions, even though there was a significant difference between them. Regarding empathy, the experiment presents some evidence that the female chatbots are perceived as more empathic than their male counterparts, whenever they do not use an empathic conversational style. It seems that the male chatbot needs to be equipped with an empathic conversational style before it is actually labeled as empathic, while for the female counterpart the perception of talking to a female is enough.

6.2 Avatar presence

Literature shows us that chatbots with an avatar are more effective than the ones without [12], however, these visual design elements seem to affect younger people less [11]. Whether there was an avatar present had little influence on the QX, but there were some differences in enjoyment and empathy. While using empathic conversational styles did not differ that much for the non-gendered chatbot, it had a way more significant impact on the male chatbot. Likely the presence of an avatar creates some sort of expectation for the bot to be more sympathetic and empathic rather than just chatting to a robot.

6.3 Empathic Conversational style

The hypotheses stating that empathic conversational styles would overall score higher than non-empathic conversational styles could not be confirmed. While generally validating responses bring out positive reactions [15], and empathy also positively affects user satisfaction while using CUIs [12, 13], these effects did not come forward from the experiment. Likely the way empathy was attempted to mimic did not particularly suit the purpose of this CUI. While a short validating response might work for a customer-service chatbot, a person sharing how they feel might expect a bit more. The chatbots used in the experiment were not able to ask further or more in-depth questions. Furthermore, the way certain emotions are identified was not always successful. For instance, a participant rating their mood of the day fairly high, but mentioning they are tired received an enthusiastic reaction, which was not the best fit.

6.4 Design Takeaways

The iterative co-creation process for the chatbot and the qualitative data gathered in the surveying process form a source of takeaways for further development of CUIs for wellbeing assessment.

Telegram is an easy platform for chatbot creation, but is not that frequently used among the participants. One participant even mentioned that they would not be willing to download telegram for this chatbot. Messages sent with delays, about one or two seconds, seem to be preferred over instant messages. Creating shortcut buttons gives the idea that those are

the only options, which can be confusing. Some participants seem to be confused about what the goal of the chatbot is, sometimes expecting more like a redirection to relevant tools or receiving advice.

6.5 Future Research

As the outcomes still were predominantly positive, using a chatbot for wellbeing assessment seems worthwhile. However, finding an optimally engaging chatbot remains a challenge. There seems to be a desire for more interaction and more in-depth questions or better recognition of certain emotions.

Overall this experiment mostly looked into adapting according to the written input of the user, not necessarily considering other user attributes like age or gender. For instance by testing whether certain genders have a preference for perceived chatbot gender. Next to that, literature shows the possible benefits of making a CUI culturally fitting [16, 17]. While this is something not assessed in this experiment, due to among others time constraints, adapting a chatbot to one's culture might positively influence the user experience. Furthermore, certain effects of the influence of chatbot gender are found in the data. However, this data was only used in an aggregated form in which these differences were no longer present. Looking further into these effects, for instance by using a non-aggregated form is recommended. Lastly, in this experiment, the non-gendered chatbot is depicted as having no avatar or name. However, opting for a non-binary avatar and name is something that still can be explored.

7 Conclusion

The aim of this research was to find the effects of adaptive CUIs on engagement and enjoyment while assessing wellbeing. A chatbot was created to evaluate Questionnaire Experience, Enjoyment, and Empathy. Changes in visual design and conversational style were made to measure the effects of gender, avatar presence, and empathy.

After conducting the experiment with 30 participants, the results showed no statistically significant effects of avatar presence or conversational style on QX. The same goes for the effects of conversational style or gender on empathy and enjoyment. However, chatbot gender does seem to affect QX. Some areas worthwhile investigating are highlighted, like cultural affiliation, non-binary chatbots or further examining gender effects. Moreover, some takeaways for further development of CUIs for wellbeing assessment are presented.

References

[1] W. van der Maden, D. Lomas, S. Fonda, and P. Hekkert, "Designing a feedback loop for community wellbeing," 2022.

[2] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, p. 36–45, jan 1966.

[3] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association : JAMIA*, vol. 25, p. 1248—1258, September 2018.

[4] J. Masche and N.-T. Le, "A review of technologies for conversational systems," in *International conference on computer science, applied mathematics and applications*, pp. 212–225, Springer, 2017.

[5] W. van der Maden, D. Lomas, U. Gadiraju, and S. Qui, "Using a conversational user interface to assess wellbeing.".

[6] R. Maharjan, D. A. Rohani, P. Bækgaard, J. Bardram, and K. Doherty, "Can we talk? design implications for the questionnaire-driven self-report of health and wellbeing via conversational agent," in *CUI 2021-3rd Conference on Conversational User Interfaces*, pp. 1–11, 2021.

[7] D. Siemon, R. Ahmad, H. Harms, and T. de Vreede, "Requirements and solution approaches to personality-adaptive conversational agents in mental health care," *Sustainability*, vol. 14, no. 7, p. 3832, 2022.

[8] S. Nelekar, A. Abdulrahman, M. Gupta, and D. Richards, "Effectiveness of embodied conversational agents for managing academic stress at an indian university (aru) during covid-19," *British Journal of Educational Technology*, vol. 53, no. 3, pp. 491–511, 2022.

[9] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental health: a review of the psychiatric landscape," *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456–464, 2019.

[10] R. G. Curtis, B. Bartel, T. Ferguson, H. T. Blake, C. Northcott, R. Virgara, and C. A. Maher, "Improving user experience of virtual health assistants: Scoping review," *Journal of Medical Internet Research*, vol. 23, no. 12, 2021.

[11] C. Straßmann, N. C. Krämer, H. Buschmeier, and S. Kopp, "Age-related differences in the evaluation of a virtual health agent's appearance and embodiment in a health-related interaction: Experimental lab study," *J Med Internet Res*, vol. 22, p. e13726, Apr 2020.

[12] C. Lisetti, R. Amini, U. Yasavur, and N. Rishe, "I can help you change! an empathic virtual agent delivers behavior change health interventions," *ACM Trans. Manage. Inf. Syst.*, vol. 4, dec 2013.

[13] H. Nguyen and J. Masthoff, "Designing empathic computers: The effect of multimodal empathic feedback using animated agent," Persuasive '09, (New York, NY, USA), Association for Computing Machinery, 2009.

[14] Y.-C. Lee, N. Yamashita, Y. Huang, and W. Fu, *"I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot*, p. 1–12. New York, NY, USA: Association for Computing Machinery, 2020.

[15] C. E. Shenk and A. E. Fruzzetti, "The impact of validating and invalidating responses on emotional reactivity," *Journal of Social and Clinical Psychology*, vol. 30, no. 2, pp. 163–183, 2011.

[16] L. Yin, T. Bickmore, and D. E. Cortés, "The impact of linguistic and cultural congruity on persuasion by conversational agents," in *International Conference on Intelligent Virtual Agents*, pp. 343–349, Springer, 2010.

[17] S. Zhou, T. Bickmore, M. Paasche-Orlow, and B. Jack, "Agent-user concordance and satisfaction with a virtual hospital discharge nurse," in *International conference on intelligent virtual agents*, pp. 528–541, Springer, 2014.

[18] C. L. Esmark Jones, T. Hancock, B. Kazandjian, and C. M. Voorhees, "Engaging the avatar: The effects of authenticity signals during chat-based service recoveries," *Journal of Business Research*, vol. 144, pp. 703–716, 2022.

[19] S. Borau, T. Otterbring, S. Laporte, and S. Fosso Wamba, "The most human bot: Female gendering increases humanness perceptions of bots and acceptance of ai," *Psychology & Marketing*, vol. 38, no. 7, pp. 1052–1068, 2021.

[20] D.-C. Toader, G. Boca, R. Toader, M. Măcelaru, C. Toader, D. Ighian, and A. T. Rădulescu, "The effect of social presence and chatbot errors on trust," *Sustainability*, vol. 12, p. 256, Dec 2019.

[21] J. Murray, J. Elms, and M. Curran, "Examining empathy and responsiveness in a high-service context," *International Journal of Retail & Distribution Management*, 2019.

[22] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload* (P. A. Hancock and N. Meshkati, eds.), vol. 52 of *Advances in Psychology*, pp. 139–183, North-Holland, 1988.

[23] J. Baumgartner, N. Ruettgers, A. Hasler, A. Sonderegger, and J. Sauer, "Questionnaire experience and the hybrid system usability scale: Using a novel concept to evaluate a new instrument," *International Journal of Human-Computer Studies*, vol. 147, p. 102575, 2021.

[24] T. Draws, A. Rieger, O. Inel, U. Gadiraju, and N. Tintarev, "A checklist to combat cognitive biases in crowdsourcing," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 9, pp. 48–59, 2021.

[25] K. Algra, L. M. Bouter, A. Hol, and J. van Kreveld, "Nederlandse gedragscode wetenschappelijke integriteit," 2018.

[26] Committee Reassessment Integrity Policy TU Delft, "TU Delft Vision on Integrity 2018-2024," 2018.

[27] F. Fossa and I. Sucameli, "Gender bias and conversational agents: an ethical perspective on social robotics," *Science and Engineering Ethics*, vol. 28, no. 3, pp. 1–23, 2022.

[28] V. C. Stodden, "Reproducible research: Addressing the need for data and code sharing in computational science," 2010.

[29] TU Delft Integrity Office, "Risk-Planning Tool. Managing Risk in Human Research." Accessed: 2022-30-04.

A Chatbot scripts

Message Sent	User Input	Quick Reply Buttons
Welcome to the 'My Wellness Check' bot. I will ask you some questions about your wellbeing. Let's start. At which faculty are you following your study programme?		
	string	"Aerospace Eng. (AE)", "Applied Science (AS)", "Architecture", "Civil Eng. (CiTG)", "EEMCS", "Industrial Design", "3mE", "TPM"
Bachelors of Masters?		
	string	
Rated from 0 (terrible) to 10 (excellent), how are you feeling today?	int	
Do you have to share anything about your mood today?		
	string	"No"
Taking all things together, how satisfied or dissatisfied are you with your life as a whole these days? Rated from 0 to 10		
	int	
Can you elaborate on that?		
	string	"No"
How would you rate your physical health, from 0 to 10?		
	int	
Thank you for answering these questions. Could you now please go to this link and answer some questions about this chatting experience?		

Table 5: Script for the baseline chatbot (NG-NE)

B Empathic reactions

Identified input	Reaction Style	Measured by
Positive	Enthusiastic	Rating >7
Neutral	Neutral	Rating >5 && <7
Negative	Sympathetic	Rating <5
Open about positive mood	Enthusiastic	Gives answer to open-text question && high rating
Open about neutral/negative mood	Appreciative and sympathetic	Gives answer to open-text question && neutral or low rating
Closed off	Accepting	Does not answer open-text question

Table 6: Reaction styles of the empathic chatbot for certain inputs.

C Survey Questions

Welcome

Question	**Question form**
As what gender do you identify	Multiple choice with open option

Questionnaire Experience (QX)

Question	**Question form**
Chatting with the chatbot was mentally demanding/complex	5-point Likert scale
Chatting with the chatbot was taking too much time	5-point Likert scale
Chatting with the chatbot was a good way to tell how I feel	5-point Likert scale
Chatting with the chatbot was frustrating or stressfull	5-point Likert scale
Chatting with the chatbot costs me little effort	5-point Likert scale
Comments	Open-question

Enjoyment

Question	**Question form**
I think that I would like to use this system frequently	5-point Likert scale
I enjoyed using this system	5-point Likert scale
I liked the look and feel of the system	5-point Likert scale
Comments	Open-question

Empathy

Question	**Question form**
I felt understood by the chatbot	5-point Likert scale
I felt sympathy from the chatbot	5-point Likert scale
I did not feel judged by the chatbot	5-point Likert scale
Comments	Open-question

Table 7: Survey questions per subsection and corresponding answer form.