Measuring the Effect of Mental Health Chatbot Personality on User Engagement

Joonas Moilanen
joonas.moilanen@oulu.fi
Center for Ubiquitous Computing, University of Oulu
Oulu, Finland

Aku Visuri
aku.visuri@oulu.fi
Center for Ubiquitous Computing, University of Oulu
Oulu, Finland

Sharadhi Alape Suryanarayana
sharadhi.suryanarayana@oulu.fi
Center for Ubiquitous Computing, University of Oulu
Oulu, Finland

Andy Alorwu
andy.alorwu@oulu.fi
Center for Ubiquitous Computing, University of Oulu
Oulu, Finland

Koji Yatani
koji@iis-lab.org
Interactive Intelligent Systems Lab, The University of Tokyo
Tokyo, Japan

Simo Hosio
simo.hosio@oulu.fi
Center for Ubiquitous Computing, University of Oulu
Oulu, Finland

ABSTRACT

Artificial Intelligence is seen as humanity's current best bet to solve the looming crisis in healthcare. Conversational Agents, or chatbots, rely on advances in AI and are increasingly investigated in the context of digital mental health care. Given how they are end-user-facing and interactive communication tools, the user engagement felt when interacting with the bots is a critical consideration. In this work, we examine the effects of chatbot personalities on the experienced user engagement with the bot. We employed personalities that rely on the Big-5 Personality Theory. Among other findings, our quantitative results indicate that a highly conscientious chatbot is likely to foster the highest user engagement. Our qualitative and content analysis also reveals desired and undesired personality features for future mental health chatbots. We discuss our findings in light of digital mental health and propose novel research directions.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Natural language interfaces**; • **Applied computing** → *Health informatics*; *Health care information systems*; *Consumer health.*

KEYWORDS

conversational agent, chatbot, personality, big five, user engagement, mental health, self-care

ACM Reference Format:
Joonas Moilanen, Aku Visuri, Sharadhi Alape Suryanarayana, Andy Alorwu, Koji Yatani, and Simo Hosio. 2022. Measuring the Effect of Mental Health Chatbot Personality on User Engagement. In *21th International Conference on Mobile and Ubiquitous Multimedia (MUM 2022), November 27–30, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3568444.3568464

MUM 2022, November 27–30, 2022, Lisbon, Portugal

ACM ISBN 978-1-4503-9820-6/22/11.
https://doi.org/10.1145/3568444.3568464

1 INTRODUCTION

As Artificial Intelligence evolves, conversational agents, or chatbots, are rapidly developing as well. One of the societally more useful use cases of chatbots is in the context of healthcare, where their potential is evident in easing the looming resource crisis by offloading certain types of work from humans to AI. And while human specialists are still needed across the field, one promising field where chatbots may offer benefits can be found in responding to distinct types of mental disorders [10, 15].

The personality of the chatbot has been shown to be one of the most crucial matters to be considered when designing the chatbot conversation [32] by making the user not only more engaged in the conversation but by also increasing the trust towards the chatbot [38]. This is, in particular, important in a mental health context, and the personality of the chatbot can greatly improve its overall effectiveness [2]. As mental health topics can often be delicate, participants want the chatbot to interact in a way that makes them feel the most comfortable. This is often achieved by having the chatbot adapt its personality depending on the participant [60]. *Self-care* is considered important [34, 56] for improving and helping maintain mental health with methods that do not require the presence of a trained clinician [50]. For these reasons, chatbots have been identified as promising solutions for helping people with their self-care needs, for example, by offering solutions to specific mental health conditions or by reassuring conversations with people [3].

In this article, we investigate how different chatbot personalities impact the interaction, focusing particularly on *user engagement* in the context of mental health self-care. We examined five different personality variants in a crowdsourced within-subjects study setup. Specifically, we contribute:

(1) And empirical investigation of the effects of different chatbot personalities on user engagement,
(2) analysis of the pros and cons of selected mental health chatbot personalities, using the Big Five personality traits, and
(3) a discussion of the implications of our findings for mental health chatbots.

We found a high conscientiousness variant to have the greatest user engagement score, followed by high extraversion and neutral

personalities. While chatbots low in conscientiousness and extraversion got overall poor results for both user engagement and their rankings, we found that some participants preferred these chatbots due to their personality matching with their own. Further, when comparing the likeability of these chatbots in the context of mental health, we found that although some participants enjoyed having a conversation with the low extraversion chatbot, they would not prefer using it for mental health purposes. We also found the opposite to be true for the neutral variation, which was found to have increased potential to be used within mental health care but a decreased general likability. Together, our findings contribute a topical case study to the literature on digital health, specifically in the context of conversational agents in mental health. We believe our recommendations as well as the study in general are helpful for researchers working on the next generation of conversational agents.

2 RELATED WORK

2.1 Chatbot Design

Chatbots have been increasingly used in research in the healthcare industry, particularly in mental health care [5, 6, 26, 33]. While chatbots are generally not sufficient to be used on their own, it has been shown that they can bring various benefits to the user, including information within an interactive environment [6]. Chatbots integrated into various applications can also provide continuous feedback in the form of weekly or daily summaries and help the users with their mental health over a longer period of time [1].

Several factors must be considered in the creation process to design a suitable chatbot. Perhaps the quintessential factor is the conversational design of the chatbot itself [7]. Consider for example a case where the information provided and the general conversation flow feels unnatural. In that case, it can be difficult for the user to trust the chatbot and, consequently, get help for their mental health issues. In addition to the informational contents of the conversation, users tend to appreciate words of empathy given by the chatbot [44] as well as a personalised chatbot [1]. Further, the perceived appearance and security have been shown to make a difference [17].

Other challenges exist in the design of chatbots. As noted by Chaves and Gerosa [9], it is important to keep the user aware of the chatbot's context and to avoid negative stereotypes within the chatbot. Aside from the context, one also needs to balance the personality traits of the chatbot. The personality of a chatbot is a critical consideration in healthcare [2, 32]. From a technical standpoint, chatbot personalities are constructed in two major ways; by having a chatbot self-adapt to the user's personality [19, 60], or by designing a static personality for the chatbot [28].

It has been found that users of chatbots prefer chatbot personalities to be proactive and witty but at the same time caring and encouraging [9, 31], and in our research, we want to find out what features would be preferred by the participants in a self-care focused chatbot and what kind of effect they have on user engagement. Thus, when designing a personality for a chatbot, it is important to balance out personality traits equally. These preferences, however, vary depending on the purpose the chatbot is used for; chatbots in the news or healthcare sites are expected to be professional, while those on shopping websites should be casual and fun [7, 9, 55]. These challenges have been further discussed within the mental health field, specifically by Potts et al. [48].

2.2 Personality Research

Personality on its own is a complex subject and forms a major part of our research. One of the most used personality theories when creating chatbot personalities is the Big Five personality traits theory [27], which divides the personality of an individual into five separate categories; extraversion, conscientiousness, openness to experience, agreeableness and neuroticism. Successful implementations using the theory include, for example, a chatbot designed for peer support [60] and university students [53], but there is an extensive amount of research on chatbot personalities focusing only on a singular few of the personality traits presented in theory, most notably extraversion [54, 55, 58], agreeableness [53, 57] and conscientiousness [9, 37]. We want to broaden the research and study the effects on chatbot personalities between different personality traits, and further by taking into account the low and high variants of said personality traits.

In addition, previous research shows that users prefer a calm and human-like chatbot personality, particularly in mental health [23], but should also follow general chatbot guidelines, such as not being too human-like to suffer from the uncanny valley effect [11]. While we hypothesize this to be somewhat true in our case as well, we are interested in seeing how user engagement changes between these personality types and how critical some of these features are. In addition, we are interested in finding whether in some cases for some participants these features are instead desirable or undesirable to have.

2.3 User Engagement

Finally, user engagement is important. It refers to the quality of the user's investment when interacting with any digital system. Good user engagement leads to sustained attention and positive outcomes in the interaction, in general. As such, it plays a crucial role in mental health chatbots [12]. The most common scale to measure user engagement is the *User Engagement Scale* (UES) [41], which has been further developed into the User Engagement Scale in Short Form [40] (UES-SF).

Measuring user engagement has seen wide use in past chatbot research. For instance, using chatbots has been shown to increase user engagement in mobile apps [46] and in general, has been one of the key features to focus on when designing highly performing chatbots [62]. While we are also comparing user engagement to each other, in our case, we want to find out which personality traits lead to best the best user engagement scores between various chatbot personalities. A similar approach has been presented, for example, in the work by Elsholz et. al. [14], where two different language styles were used for chatbots and the changes in user engagement were reported. It is clear, that the language style and chatbot personality greatly affects user engagement, and thus we want to explore the changes between multiple Big Five personality traits specifically in the mental health context, to find which personality trait could lead to the most engaging chatbot.

3 THE STUDY

We are interested in chatbot personalities, their potential uses, and how we could design better chatbots in the future, e.g. by being able to tailor different personalities for different use cases. In this study, we investigate into how different personalities affect user engagement with a chatbot. As noted earlier, user engagement is one of the key success elements of conversational agents [18, 22, 49].

We first sourced the personalities used by our chatbots from work by other scholars. Then, we implemented a chatbot interface to enable conversations between the participants and the chatbots. Finally, we conducted an online study investigating the connection between chatbot personalities and the resulting felt user engagement.

3.1 The Chatbot Personalities

To design the chatbot personalities, we used findings from research about chatbot personalities covered in the related work section, as well as from our earlier work [36]. We decided to focus on the conscientiousness and extraversion traits, as these two traits have been widely used in past research, and in addition, can be easily used to alter a mental health chatbot personality. For instance, using the neuroticism personality trait would in most cases be undesirable in a mental health context. Extraversion and conscientiousness traits are also easily alterable in text due to a large amount of previous research, and individuals tend to perceive the extraversion [29] and conscientiousness [61] of chatbots differently. For instance, the conscientiousness of a chatbot has been suggested to be a critical factor when dealing with mental health care online [30].

We build on a set of previously developed chatbot scripts of different personalities, developed in [36], in which the personalities were designed by altering the neutral chatbot conversation using various language cues [35]. The scripts follow the Big Five Personality Traits Theory [27]. The design process of the personalities is further covered in Moilanen et. al. [36]. Thus, we investigate five different personalities – neutral (Neutral), low conscientiousness (Low C), high conscientiousness (High C), low extraversion (Low E), and high extraversion (High E), as made available in [36].

3.2 Chatbot Implementation and Operation

We used a programmable third-party chatbot platform, *BotStar*[1]. Botstar offers an easy-to-use conversation builder and typical online input elements, such as dropdowns, buttons, etc. It integrates with Google's Dialogflow for sophisticated language operations. The bots created with Botstar offer a simple user interface to be embedded or used in a full-screen mode online (see Figure 1). Each of our chatbots was given a name: Bot, Clank, Proto, Core and Spark for Neutral, Low C, High C, Low E and High E chatbots, respectively.

The flow of the conversation, presented in Figure 2, was the same for each chatbot interaction, with the only change being their varying personality. After the initial introduction of the purpose of the chat session, the chatbot asked the participant to choose from three mental health conditions; stress, anxiety and low mood. After users selected the preferred condition, the chatbot offered science-backed self-care methods for the conditions - healthy sleep habits [39], physical exercise [47] and spending time with your preferred activity [47] - respectively for previously mentioned conditions. The user could repeat this step until they wished to continue to the next part of the conversation, in which they were asked for open-ended feedback on the methods they were given, in the form of free text. The chatbot then used Google's Dialogflow[2] to detect whether the participant liked or disliked the given methods and provided a response. This feature was implemented by training Dialogflow's natural language understanding algorithm with roughly 100 example sentences and words provided by the paper's authors. For instance, a neutral chatbot would reply 'Good to hear' when receiving positive feedback and 'Sorry to hear that' for negative feedback. In the second part of the conversation, the chatbot gave information on three additional sources for the users to study more about mental health self-care, helping them choose the source most suitable for their needs. These sources consist of literature [21, 51], audio sources [8] and the internet [4, 25]. For example, literature provides information on most mental health topics while helping to build self-confidence and self-understanding of the user in the process [51]. After this, the aforementioned feedback process was repeated, and the chat session ended.

3.3 Study Design

To investigate the connection between the implemented personalities and user engagement, we designed a within-subjects study. As we wanted the participants to rank the chatbots from the worst to best, and wanted to describe what sort of factors made some chatbots better suitable for mental health context than others, we had each participant interact with each of the five chatbots. To mitigate any carryover effects and fatigue from a lengthy study, we used Balanced Latin Square [59] counterbalancing to assign the participants to the chatbots. This resulted in 10 orders of chatbots.

Initially, participants were directed from the crowdsourcing platform to Google Forms. After filling in for consent form in Google Forms, the participants were directed to the study's homepage online and provided with written instructions for the study. This was followed by links on the homepage to each study section, which were to be done in the order presented to the participant. Every participant interacted with each of the five chatbots, and after each chatbot interaction, the participant was directed to fill in the User Engagement Scale survey in Short Form (UES-SF) [40]. This short form consists of 12 survey questions and four different subscales: focused attention (FA), perceived usability (PU), aesthetic appeal (AE) and reward factor (RW). This was done in Google Forms, where we passed the participant ID and the chatbot identifiers as URL parameters. The participants were also asked to name three positive and negative features of the personality of each of the chatbots they had interacted with. Questions **Q1** and **Q2** were asked after every chatbot interaction, thus leading each participant to fill in the survey five times, once for each chatbot.

- **Q1** Please name three good things about the chatbot's personality. [open-ended]
- **Q2** Please name three bad things about the chatbot's personality. [open-ended]

[1] https://botstar.com/

[2] https://cloud.google.com/dialogflow

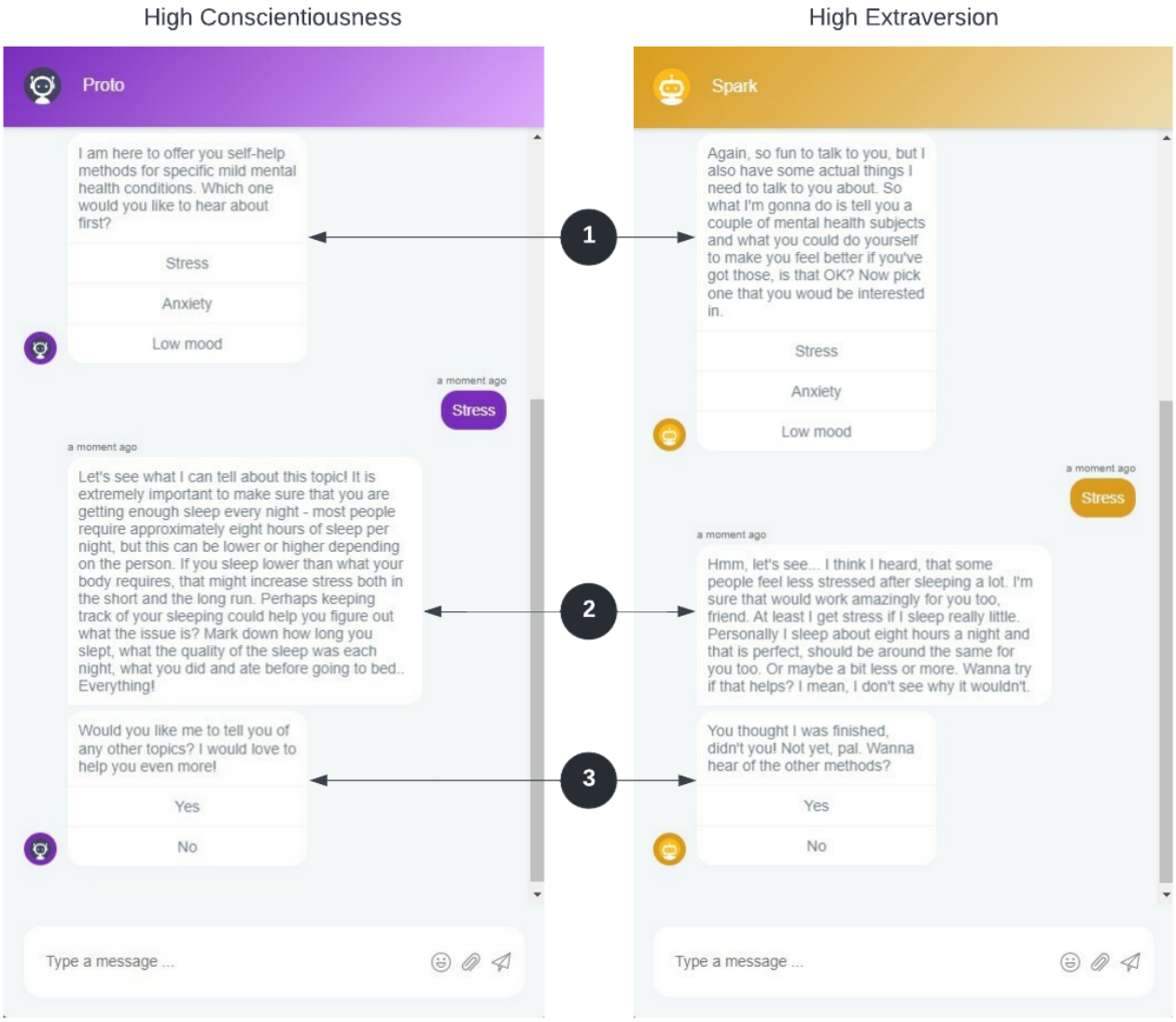

Figure 1: Example parts of the conversation with the High C (left) and High E (right) chatbots. The conversations show (1) the selection of the self-care method the participant wishes to hear about, (2) the self-care method for stress, and finally, (3) the possibility for repetition of the method selection.

Having completed the conversations with all chatbots, the participant was asked to fill in a final online questionnaire in Google Forms. The participant's ID was passed to the questionnaire as a URL parameter. In the questionnaire, we asked the participants to rank the chatbots using two different criteria, and in both of these questions, rank 1 presents the participant's most liked chatbot, while rank 5 presents their least liked chatbot. These ranking questions are presented as **R1** and **R2** given below. Finally, we ask the participant to elaborate on their chosen orders, denoted as **R3** below.

- **R1** Which chatbot did you like talking with the most? [ranking question]
- **R2** Which chatbot would you prefer to use to find self-care solutions for mental health? [ranking question]
- **R3** Please elaborate on the above choices. [open-ended]

3.4 Pilot Study

Before conducting the main study, we recruited three participants to test the implementation and protocol. As with the main study, we used Prolific as our participant source. Prolific is a leading online

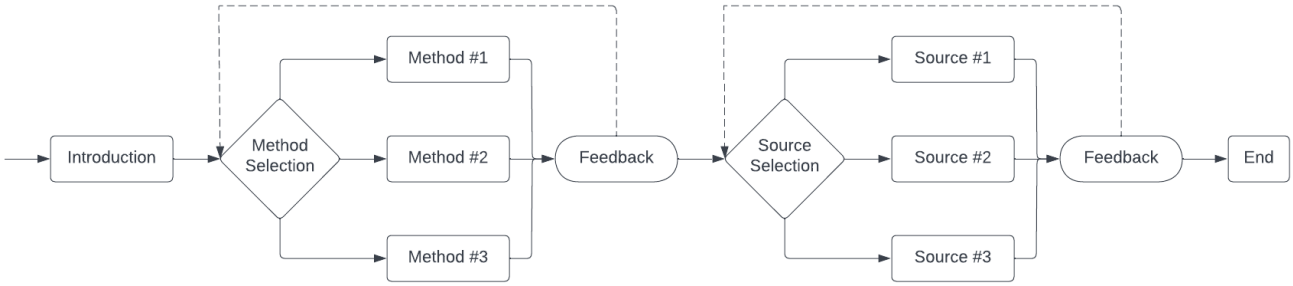

Figure 2: Conversation flow is used for each chatbot. In the method selection, the participants can select from methods for stress, anxiety and low mood offered by the chatbot as methods #1, #2, and #3, respectively, based on the participant's selection. This is repeated similarly for sources #1, #2 and #3, which present the participants with information on additional sources for mental health self-care, namely, literature, audio sources and the internet, respectively. After giving feedback for the given method or source, the user can repeat the selection steps until they wish to continue the conversation.

Table 1: The mean user engagement scores and standard deviations for each chatbot, alongside the four different subscales. FA = Focused Attention, PU = Perceived Usability, AE = Aesthetic Appeal, RW = Reward Factor.

	UES Score	SD	FA	SD	PU	SD	AE	SD	RW	SD
Neutral	3.06	0.62	2.46	0.81	4.11	0.80	2.71	0.96	2.97	1.03
Low C	2.69	0.77	2.28	0.78	3.68	0.99	2.36	1.01	2.46	1.10
High C	3.46	0.72	2.78	0.98	4.35	0.72	3.19	1.02	3.53	1.01
Low E	2.93	0.81	2.53	0.98	3.85	0.91	2.55	1.07	2.79	1.17
High E	3.23	0.75	2.7	0.91	4.11	0.83	2.91	1.10	3.19	1.08

subject pool, widely used in psychological and other online user experiments. It has been shown to provide high-quality data [43]. The participants were compensated with $4.47, and their average study completion time was ~33 minutes. This average completion time of the pilot study was used as an estimate for the main study. In the main study, each participant was paid $6.00 for their participation, with an average completion time of ~40 minutes. We used Prolific's participant prescreening tools to limit the participants to those with at least a 95% approval rate and those with at least 50 previously completed studies. In addition, we excluded the participants from the pilot study.

3.5 Ethics Approval

We followed our local Institutional Review Board protocols, where online studies such as this are not considered interventions or pose a greater risk to participants than online questionnaires. Therefore, our study was considered exempt from a separate ethics approval. Participants were, however, asked for written informed consent before beginning the study, as required by our institution.

4 RESULTS

4.1 Participants

We used the Prolific human subjects pool for the main study as well. Demographic data of the participants was gained directly from Prolific crowdworker platform, including their general demographics as well as information on the chosen prescreeners. We rejected three workers due to not following the instructions of thecrow study to a satisfactory degree and replaced them with new ones.

In the main study, we chose to recruit 100 participants in total, i.e. 10 participants assigned to each order of the chatbots. Of the 100 participants, 39 were female, 60 were male, and one preferred not to state their gender. The mean age of the participants was 27.19 ±1.60 years. Most participants were currently located in Europe (78 participants), and the most presented country was Portugal (21 participants). English was the first language for 1 participant, and the most common first language among participants was Polish with 20 participants.

4.2 User Engagement Scores

We measured the user engagement of all the chatbots using The User Engagement Scale (Short Form), UES-SF [40], a widely used scale in HCI. Table 1 depicts the results. UES-SF scores can range from 1 to 5, and in particular, for the high traits, our scores are similar to those of other chatbot research using UES-SF [16, 18, 49]. We performed a one-way Analysis of Variance (ANOVA) test to examine the user engagement between the employed chatbots and found significant differences ($p<0.001$). Following that, we performed a pairwise Tukey's range test between the chatbots to find differences in the user engagement scores. We found a statistically significant difference between Neutral and High C, Low C and High C, Low C and High E, High C and Low E ($p<0.001$ for all pairs), Neutral and Low C, and Low E and High E ($p<0.05$ for all pairs).

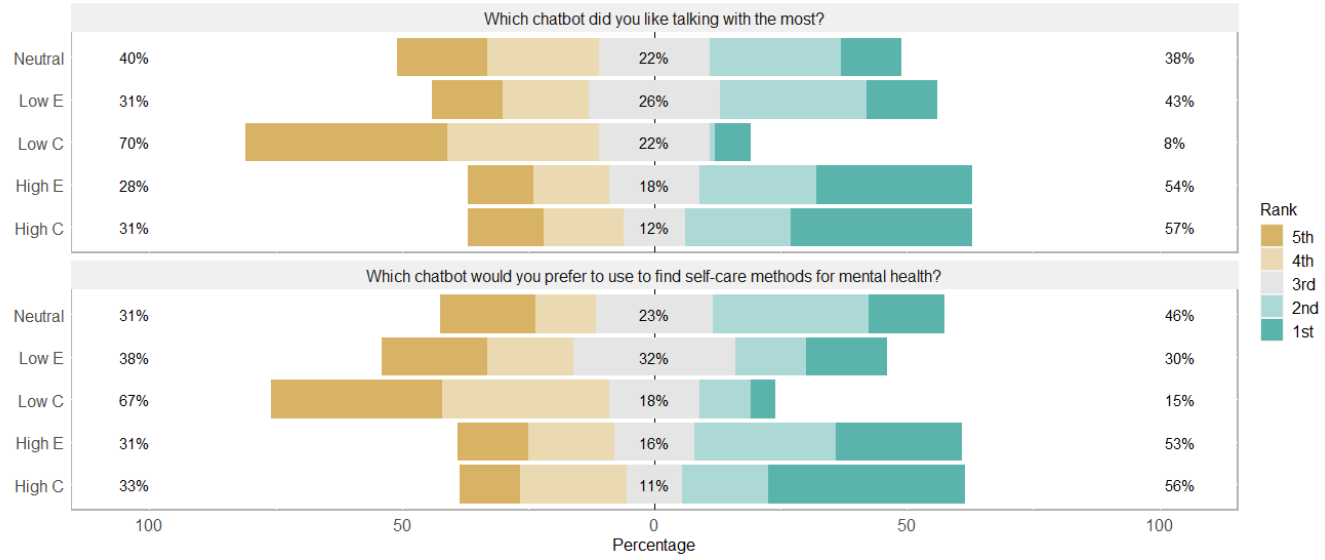

Figure 3: Rankings of the chatbots. The 5th rank on the left presents the participant's least liked chatbot, while the 1st on the right presents their most liked chatbot. The compound percentages of the low ranks (5th and 4th), average rank (3rd) and high ranks (2nd and 1st) are presented on the left, centre and right-hand sides of the plots, respectively.

In addition, we calculated Pearson's correlation coefficients, comparing the correlation between the final user engagement score to that of singular UES-SF subscale scores (FA, PU, AE, RW), and measured the statistical significance of each correlation coefficient ($p < 0.001$). We find that Reward Factor (RW) and Perceived Usability (US) have the highest and lowest scores of the subscales respectively. RW correlates to the user engagement score highest, with $r = 0.91, 0.91, 0.88, 0.90$ and 0.92 for Neutral, Low C, High C, Low E, and High E, respectively. Perceived Usability (US) has the least correlation to the user engagement score, with $r = 0.45, 0.71, 0.57, 0.65$ and 0.57 for Neutral, Low C, High C, Low E, and High E, respectively.

4.3 Ranking of Chatbots

The results of the chatbot-ranking tasks in the Final Questionnaire are depicted in Figure 3.

We performed a one-way ANOVA test to examine the differences between the rankings for different chatbots concerning both ranking-related questions (**R1**: Which chatbot did participants like talking with the most; **R2**: Which chatbot would participants prefer to use to find self-care solutions for mental health). Statistically significant differences in the rankings of the chatbots were found with respect to both **R1** and **R2** ($p<0.001$). This was followed by a pairwise Tukey's range test. For **R1**, we found significant differences between Neutral and Low C, Low C and High C, Low C and Low E, Low C and High E ($p = 0.001$ for aforementioned pairs), Neutral and High C, and Neutral and High E ($p < 0.05$ for aforementioned pairs). For **R2**, we found significant differences between Neutral and Low C, Low C and High C, Low C and High E ($p = 0.001$ for the aforementioned pairs), Low C and Low E, and High C and Low E ($p < 0.05$ for the aforementioned pairs). These results show a significant statistical difference between the rankings for each chatbot.

Then, we tested for statistically significant differences between the responses for ranking questions **R1** and **R2** for each chatbot. This was done by performing t-tests to see find differences for this pair of questions for each chatbot. We found significant differences between these two questions for Neutral ($t = 2.05$, $p = 0.04$) and Low E ($t = -2.61$, $p = 0.01$) chatbots.

4.4 Qualitative Analysis

After ranking the chatbots, the participants were asked to elaborate on their chosen ranks openly. We analysed the results using conventional content analysis [24]. Given how the answers were already primed to be justifications of people's prior answers, we deemed this to be an appropriate solution compared to, e.g. those aimed at deriving theories from large datasets (grounded theory). Two of the article's authors developed a tagging scheme for all the responses and collaboratively derived the following themes.

Professionalism. One of the most important factors that led the participants into liking or dislike a chatbot was their professionalism or lack thereof. Although the chatbot's friendliness mattered to most participants, most of them preferred a professional chatbot which gave them precise information: *"Proto gave me the most useful advice - the one about the importance of sleep. The others were friendly, but they did not quite help me the way Proto did - they just gave me standard methods."* (P90)

In addition to the provided information and professionalism, these participants often appreciated the generally more professional language and the tone of the chatbot; *"I felt Proto was best to talk to because it didn't use weird or awkward language and sounded most professional and I felt taken seriously."* (P13) and *"I like simple*

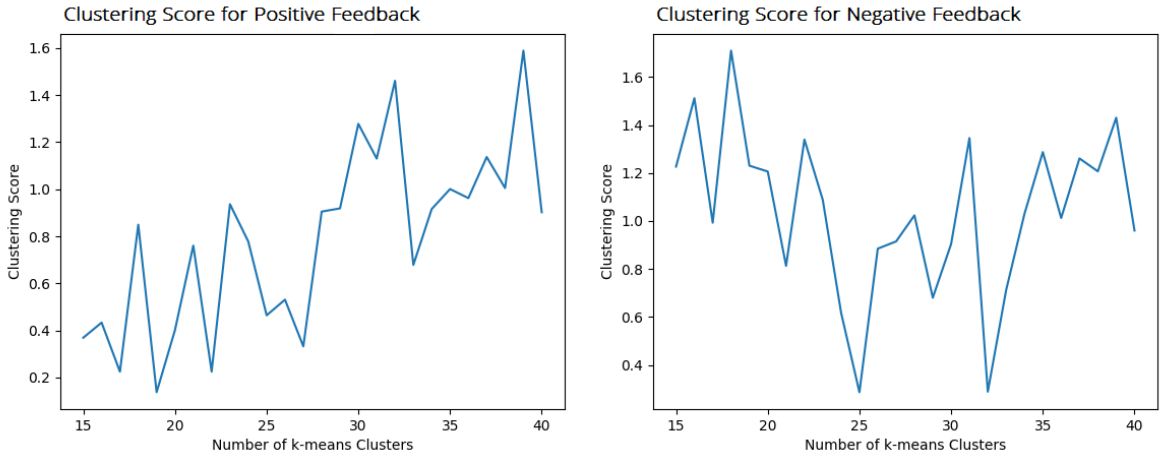

Figure 4: Clustering scores for positive and negative feedback. The clustering score is calculated by adding the normalized Silhouette method and Dunn index scores. Amount of clusters is chosen from the highest values of the scores, presenting us with k=39 clusters for positive feedback and k=18 clusters for negative feedback.

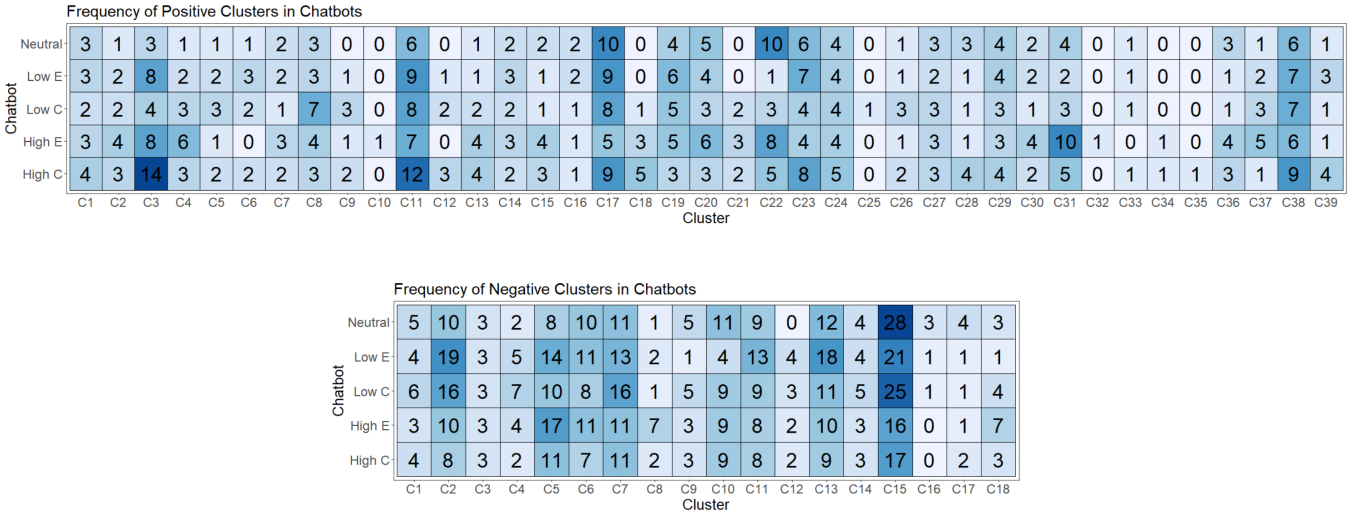

Figure 5: Heatmaps for the frequencies of generated clusters in each chatbot for both the positive and negative feedback.

and short answers for what I look like; I don't want to socialise with Google." (P35)

When designing the personalities, we wanted the chatbots to present the participants with the same methods and information. An exception is the High C chatbot, which offers a few extra methods for each mental health condition, a trait of said personality [35]. This was done to explore whether the amount of information could, in some cases, be overwhelming or just a positive factor. Some of the participants perceived this feature to distinguish it better from the rest of the chatbots professionally; *"Proto gave me the most useful advice - the one about the importance of sleep. The others were friendly, but they did not quite help me the way Proto did - they just gave me standard methods"* (P90), while some found the amount of text to be exhausting; *"I didn't like Proto because it didn't feel like a conversation. It just sent me long pieces of text to read."* (P42)

Friendliness. The second most important factor for the participants was the friendliness of the chatbot, with the High E chatbot being the most mentioned. Participants found the cheerful tone of High E to be reassuring, which could raise their mood just by talking to them: *"Spark gave the impression that it felt interested in what I was thinking about. It was enthusiastic and boosted my mood. On the contrary Clank felt like he was uninterested and in a low mood"* (P15) and *"I feel like if you're feeling down, it would be helpful to chat with a bot that is open, friendly and cheerful and sounds more human, rather than one that sounds fake and more like a robot. Spark was the most fun to "chat" with for me."* (P48)

However, it is also worth noting that several participants commented negatively on too human-like and cheerful chatbots. Having an overly happy and excited chatbot in mental health applications could affect the participants negatively and make them seem less professional. *"Spark's positive communication can help those struggling with such diseases"* (P82) and *"— Spark was a bit too much of friendliness, and the others were just mean and cold and distant."* (P41) and *"Bots like Spark that are too friendly appear less professional and expert. —"* (P54)

The overall personality of the chatbot and how it matches with the participant was also highly presented among the answers: *"I value personality over the information itself. Therefore, "talking" to something closer to how I myself am would make me feel better."* (P18) While most participants seemed to enjoy the talkative High E chatbot, there were also participants who enjoyed talking with the generally less liked Low C and Low E chatbots, as the personality match was much closer to the participant: *"I think Clank suits my personality the best. It was not too careless but not too strict"* (P36) and *"I really felt Core was the most helpful for my personality type. The remaining choices were based on general preference."* (P73)

Middle-ground. Although many participants liked the professionalism and friendliness provided by the High C and High E chatbots, some of the participants found those traits from those chatbots to be undesirable and preferred a chatbot with neutral traits instead: *"When it comes to mental health help, I find the neutral, more professional bots more helpful—"* (P4). Similar to the user engagement score results, the neutral chatbot is often mentioned to fall in between the high and low variants of chatbot personalities due to its personality falling in the middle of both personalities: *"Core and Spark felt the most informative and friendly to me without being too informal and too friendly like Clank for example. The bot was ok, it was neutral, and Proto was straight-up rude."* (P84)

4.5 Clustering the Pros and Cons

Participants were asked to name three positive and negative aspects of each chatbot. To analyze the results, we performed k-means clustering [20] to find which words were most frequently used to describe different personalities either positively or negatively, in a process typically referred to as a bag-of-words categorisation.

We manually processed the feedback, extracting individual descriptive words from each feedback. For example, a positive feedback sentence "This chatbot was nice" would be reduced to a single word "nice". These descriptive words were then categorized into positive and negative groups, with the occurrence of each word associated with a corresponding chatbot personality. For example, the word 'helpful' was used to describe the Low C personality 25 times, and the word 'cold' was used to describe the Neutral personality in 7 items.

To cluster these words from both the positive and negative feedback, we used the numeric tone values of the words. These were gained by using the tone classification model of IBM Watson's Natural Language Understanding (NLU) service[3]. We had the feedback words listed in a CSV file for the positive and negative feedback, and by using a Python script, we separately passed each word to the NLU service. The NLU service returned JSON objects as an output, containing numeric scores for each tone for the given words. Next, these scores were appended to the CSV file with the words, pairing each word with numeric tone scores. These scores were further used to perform the clustering, presented in the following paragraphs. The tones gained from the NLU service consist of *excited, frustrated, impolite, polite, sad, satisfied* and *sympathetic*, and their scores range between 0 and 1. For example, the word *smart* has tone scores of 0.390808, 0.020715, 0.00546, 0.410827, 0.029162, 0.391648, and 0.111325, respectively.

To find the optimal k-value for the k-means clustering algorithm, which defines the number of clusters, we used a balanced combination of the Silhouette method [52] and Dunn index [13] to calculate a Clustering Score for each k-value. Dunn index value determines the within-cluster variance and separation of clusters, and the Silhouette method similarly measures within-cluster cohesion and cluster separation. The minimum k-value was selected as twice the number of features (k=15) to ensure the clusters would not be formed just based on a singular feature. The maximum k-value was set to 40. We repeated this process twice for both the positive and negative words.

The Dunn Index and Silhouette coefficient values were calculated using the Skicit-learn library for Python [45], then normalised both of these values and added them together as shown in Figure 4. The optimal number of clusters can be found from the highest values in the graphs, leading us to generate k=39 clusters for positive feedback and k=18 clusters for negative feedback. The final word clustering was implemented using the Skicit-learn k-means function, and each word was assigned to one of the clusters.

Each cluster contained words that were similarly based on the seven tones provided by the NLU service, e.g., words that commonly ranked high on the 'excited' and 'impolite' tones could be grouped. We then calculated the frequency of each cluster for each personality and ordered the clusters based on popularity. The frequencies of these bag-of-words clusters for each personality type are shown in Figure 5. The heatmaps show that for both positive and negative feedback, a handful of clusters tend to be slightly overrepresented, e.g., C3, C11, C17, and C38 for positive feedback and C15 for negative feedback.

The top five highest ranked clusters for each chatbot are presented in Tables 2 and 3, and present example words chosen from the pool of words in that cluster. The results seem to mirror the traits of each personality type. Neutral is most described as "formal", "honest", or "confident", while Low C is "serious" and "direct", High C is "exhaustive" and "interested". The word "language" in High C relates to using more complicated language. Low E is similar to Low C - "informative" and "serious", while High E is most cited as "interactive" and "likeable".

In addition to clustering the words, we measured the word frequencies for both positive and negative feedback for each chatbot. The most frequent positive words for the various chatbots were helpful (Neutral, Low C, High C), direct (Low E), and friendly (High E), while the most frequent negative words were boring (Neutral, Low C), long (High C), rude (Low E), and talkative (High E).

In conclusion, we managed to find three key themes from the thematic analysis of the participant feedback to what affected their opinion on the chatbots; professionalism, friendliness and the middle-ground. These themes repeat within the clustering as

[3]https://www.ibm.com/cloud/watson-natural-language-understanding

Table 2: The five most common clusters of words for positive feedback (Q1) are given to each chatbot, with a few selected words presented for each cluster.

Rank	Neutral	Low C	High C	Low E	High E
1st	[C17] confident, formal, friend, honest	[C11] attentive, direct, informative, serious, useful	[C3] exhaustive, interested, language	[C11] attentive, direct, informative, serious, useful	[C31] interactive, likeable, words
2nd	[C22] enthusiastic, energetic, modern	[C17] confident, formal, friend, honest	[C11] attentive, direct, informative, serious, useful	[C17] confident, formal, friend, honest	[C3] exhaustive, interested, language
3rd	[C11] attentice, direct, informative, serious, useful	[C8] active, easy-going, kind	[C17] confident, formal, friend, honest	[C3] exhaustive, interested, language	[C11] attentive, direct, informative, serious, useful
4th	[C23] calm, professional, specific	[C38] answers, clear, fast, understandable	[C38] answers, clear, fast, understandable	[C23] calm, professional, specific	[C4] optimistic, relatable, sociable
5th	[C38] answers, clear, fast, understandable	[C19] humane, personality, simple, straightforward	[C23] calm, professional, specific	[C38] answers, clear, fast, understandable	[C20] intelligent, positive, respectful

Table 3: The five most common clusters of words for the negative feedback (Q2) are given to each chatbot, with a few selected words presented for each cluster.

Rank	Neutral	Low C	High C	Low E	High E
1st	[C15] dry, monotone, personality, uninterested	[C15] dry, monotone, personality, uninterested	[C15] dry, monotone, personality, uninterested	[C15] dry, monotone, personality, uninterested	[C5] disconnected, inattentive, uninteractive
2nd	[C13] formal, general, serious, tryhard	[C2] bad, limited, repetitive, simplistic	[C5] disconnected, inattentive, uninteractive	[C2] bad, limited, repetitive, simplistic	[C15] dry, monotone, personality, uninterested
3rd	[C7] automatic, cold, obvious, superficial	[C7] automatic, cold, obvious, superficial	[C7] automatic, cold, obvious, superficial	[C13] formal, general, serious, tryhard	[C6] examples, information, straightforward
4th	[C10] chatty, humane, joking	[C13] formal, general, serious, tryhard	[C10] chatty, humane, joking	[C5] disconnected, inattentive, uninteractive	[C7] automatic, cold, obvious, superficial
5th	[C2] bad, limited, repetitive, simplistic	[C5] disconnected, inattentive, uninteractive	[C13] formal, general, serious, tryhard	[C7] automatic, cold, obvious, superficial	[C2] bad, limited, repetitive, simplistic

well, and in particular, we note helpfulness and friendliness to be one of the most critical factors to increase their opinion, and boringness and rudeness to lower their opinion of the chatbot. We found a few key clusters which were dominant among all of the chatbots, for instance, C11 (attentive, direct, informative, serious, useful) and C17 (confident, formal, friend, honest) for positive feedback, and C15 (dry, monotone, personality, self-centred, uninterested) for negative feedback.

5 DISCUSSION

In this paper, we researched how different chatbot personalities affect user engagement, and further analysed the feedback given by the participants for each chatbot. We found the High C and High E chatbots to have the highest User Engagement score, and these two chatbots were also ranked highest among all of the chatbots. Interestingly, we saw significant differences between the ranking tasks R1 and R2 for Neutral and Low E chatbots. We found three common themes - professionalism, friendliness and middle-ground - for what made the users like the chatbots. From the clustering of the results, we found the most common positive and negative features for each personality, which could be used to further improve the chatbot.

Our research shows the feasibility of using personalities to increase user engagement for mental health chatbots and thus contributes toward creating more impactful chatbots in the field. We can see significant differences in user engagement and overall participant preferences by altering the chatbot's personality, and in addition, recognize new key challenges that must be considered when designing chatbots and their personalities specifically for mental health.

While the chatbots presented in this work are far from complete, we can say using them to offer mental health self-care shows promise. Chatbots offer an easily accessible way to promote mental health, and self-care methods are easily applicable to everyday life. These chatbots do not replace healthcare professionals and are not intended to offer actual clinical intervention. However, we believe that by increasing the number of methods and conditions, regular use of similar chatbots could lead to better mental health for the participants. Using chatbots to offer (mental health) self-care is still a relatively unexplored topic, and in addition to the findings for the chatbot personalities, we contribute towards self-care research as well.

5.1 Using Personalities to Enhance Mental Health Chatbots

As is seen in the results, High C and High E chatbot personalities have the highest user engagement out of the five different chatbots. These results align with previous chatbot research, which shows that participants tend to prefer chatbots with high conscientiousness [9] and high extraversion [58] personalities. While user engagement is but one value to measure the effectiveness of a chatbot, we can tell that similar findings to more general chatbots can be used when designing chatbots for mental health purposes. We also find the differences between the user engagement scores for High C, High E and Neutral chatbots to be relatively small; therefore,

Table 4: Clusters generated for both the positive (Q1) feedback. Few descriptive words of each cluster are selected, with appearances in multiple chatbots.

Cluster	Common Words for Positive Feedback
C1	attentive, juvial, handy
C2	appealing, happy, uplifting
C3	exhaustive, informal, interested, language
C4	optimistic, relatable, sociable
C5	pleasant, smart
C6	approachable, helper, non-robotic
C7	assertive, chatty, responsive
C8	active, easy-going, kind
C9	buttons, dynamic, non-automatic
C10	cheery
C11	attentive, direct, informative, serious, useful
C12	intuitive, trustworthy
C13	enthusiastic, playful, relaxed
C14	concise, efficient, good
C15	cute, entertaining, fun, handsome
C16	motivated, precise, supportive
C17	confident, formal, friend, honest
C18	constructive, feedback, patience
C19	humane, personality, simple, straightforward
C20	intelligent, positive, respectful,
C21	eager, emotional, extraverted
C22	enthusiastic, energetic, modern
C23	calm, professional, specific
C24	affable, easy, helpful,
C25	talkative
C26	cheerful, timely
C27	polite, understanding
C28	accurate, neutral
C29	caring, engaging, resourceful
C30	bright, charming, humorous, nice
C31	interactive, likeable, words
C32	grateful
C33	informative
C34	joyful
C35	empathetic
C36	compassionate, energetic, warm
C37	funny, interesting, upbeat
C38	answers, clear, fast, understandable
C39	friendly, thoughtful

any of the above personalities could have potential use within this context.

When analysing the results further, we saw High C and High E chatbots get similar results for both ranking questions. Neutral chatbot had an increased rating for the question about the preferred chatbot for mental health self-care. This was the opposite for the Low E chatbot. These results further validate the implications based on the chatbots' user engagement scores but raise new aspects to be considered. While Low E was still a more undesired personality for mental health chatbots, it could imply the importance of

Table 5: Clusters generated for both the negative (Q2) feedback. Few descriptive words of each cluster are selected, with appearances in multiple chatbots.

Cluster	Common Words for Negative Feedback
C1	boring, distant, unprofessional, unresponsive
C2	bad, limited, repetitive, simplistic, uninformative
C3	cheesy, energetic, uncanny-valley
C4	annoying, frustrated, inhumane
C5	disconnected, inattentive, uninteractive
C6	examples, information, straightforward
C7	automatic, cold, obvious, superficial
C8	excited, fun, happy, vibrant
C9	cheerful, friendly, wild
C10	chatty, humane, joking, topics, words
C11	arrogant, attitude, harsh, mechanical, robotic
C12	irritating, talkative
C13	formal, general, serious, tryhard
C14	detached, indifferent, useless
C15	dry, monotone, personality, self-centred, uninterested
C16	apathetic, non-empathetic
C17	bot, programmed
C18	abrupt, clumsy, shy

lowering the chatbot's extraversion to accommodate a larger population's preferences. And while the Neutral chatbot was often deemed monotone and robotic, these features could increase the professionalism of the healthcare chatbot, as was found in previous work by Cameron et. al. [7] as well.

However, as opposed to previous research, based on our findings, we suggest a purely neutral personality to not the most desired personality for mental health chatbots. Instead, we believe the general approach suggested by Chaves and Gerosa [9] to be more beneficial, and the chatbot should balance out these personality traits equally. As a Neutral chatbot got a relatively large user engagement score and particularly saw an increase in the ranking for mental health purposes, we believe a chatbot expressing more neutral conscientiousness and extraversion traits to be the most desired for mental health use. In addition, the more neutral tone could help us answer the common uncanny valley challenge that chatbots often face [11], and ensure the chatbot is identifiable as a bot.

Having the chatbot express both personality traits, and avoiding their extreme variants, is further backed by the fact that the participant's personality impacts their opinions. In our work, several participants directly mentioned liking a certain chatbot due to its personality matching their own. When both personality traits are present, we improve the chances of having at least one common trait between the chatbot and its user. Furthermore, by, for example, reducing the extraversion trait within a chatbot, an introverted user would assumably find it to be more enjoyable than one showing much higher extraversion.

5.2 Challenges of Chatbots for Mental Health Self-Care

We see differences in the correlations between the UES-SF subscale scores and the overall user engagement score from those presented

in the original research [42]. In our research, the correlation of the reward factor was the highest, and perceived usability was the lowest, which is the opposite of the original work. This may hint at the importance of the contents of the conversation with the mental health chatbots. As our chatbots were designed to be straightforward and most communication on the participant's end was done via buttons, there were no major issues with the overall usability of the chatbot, apart from the feedback feature. Chatbots using a more non-linear structure, options, and user feedback could further raise the perceived usability subscale's significance. Thus, to enhance the performance of a mental health self-care chatbot, taking into account the three questions from the perceived usability subscale could positively impact the chatbot; make sure that using the chatbot is not frustrating, confusing, or taxing. This can be mainly achieved with a well-designed conversation flow. While giving the participant freedom in the conversation can enhance the overall user experience, it also poses challenges and can lead to lower scores for the perceived usability subscale. Thus, for the best user engagement scores, a more straightforward conversation could often lead to the most reliable results.

Clustering the participant feedback provided us with in-depth knowledge of the desired and undesired features within these chatbots. The generated clusters are not equally impactful. Clusters C11 and C17 appear to be the most common for positive feedback, while most negative clusters are more equally presented. This could be due to the lower amount of clusters generated for the negative feedback but could also simply show that regardless of the personality, chatbots are still perceived as "dry", "monotone", and "uninterested" in conversation. Participants showed much more versatility in their personalities when describing their positive traits.

When designing a chatbot personality for mental health, focusing on distinct key personality features could enhance the participant's opinion of it regardless of the overarching personality. These personality features include informativeness (in C11), confidence (C17), honesty (C7), and formality (C17). Therefore, to make the chatbot more interesting, one could consider increasing traits related to its conscientiousness and traits related to extraversion for interactivity. As the low variants of these personalities were more commonly associated with the negative clusters, they should generally be avoided when creating a static chatbot personality, as they make them seem poorly designed and arrogant. These features include boringness (in C1), laziness (C2), long responses (C15) and unresponsiveness (C1). As some of these traits are also among the language cues used to design personalities, research articles covering the uses of these language cues could provide useful. To combat these negative traits and enhance the chatbot, the personality should be interesting and proactive, the bot should keep the messages informative yet short enough, and it should reply accordingly to the participants' messages. Curiously yet understandably, in our study, many descriptions were used as both positive and negative traits. For instance, as is seen in Tables 2 and 3, the word "formal" was included in the most common cluster for positive feedback, and 2nd most common cluster for negative feedback. Similarly, High C was also described as "serious" as both a positive and a negative trait. The amount of information provided by High E is also reflected on both negative ("information") and positive ("informative") axes.

In addition, we recognise the negative impacts that high amounts of conscientiousness and extraversion traits could have from both the clustering results and the other feedback the participants gave. While the amount of information provided by High C, and the amount of interaction and conversation provided by High E, were mostly seen positively by the participants, in some cases, they lowered the participant's opinion of them. This is in particular notable for the High E chatbot. When providing the participants with mental health help and information, it is also important to consider their moods. While in some cases friendliness of the chatbot was seen to be encouraging and help them feel better, often it had the opposite effect, and the participant would have preferred it to talk in a more neutral tone. This could be done by for example detecting the participant's mood with NLU methods and having the chatbot adapt their personality accordingly, or by simply asking it in the conversation flow. In case this feature is not implemented within a chatbot, we believe more neutral tones would be the safest to be used.

It is clear that based on these results, deciding on the correct chatbot personality to be used within this context is not an easy task. Participants' personality and preferences greatly alter their views on different chatbots. While in an ideal situation, the chatbot would be designed to alter their personality depending on the participant, this is still not an easily accessible approach. Analysing the participant's typing and writing style, matching their mental health problems with personality traits frequently associated with it, or simply asking, could provide key clues on how to adapt the chatbot's personality to the task at hand. As based on our findings, we suggest that a chatbot in this context should express slightly positive extraversion and conscientiousness personality traits.

5.3 Future Work

In future work, we are interested in researching the effects of the other personality traits not presented in this work and how we could use these results for personality-adapting chatbots. In addition, we are looking to include a personality survey in the study to analyse further the impacts of the participant's personality on their preferred chatbot personality. During the data collection, we had the participants fill out a 10-item personality survey (TIPI) during the study session. However, the TIPI is an extremely brief measure, and we opted to not use it in our analysis. Thus, in future studies, we plan on using a different survey for participant personalities. When including other personality traits and variance in how strongly they express the given personality, we hope to find ties between the participant and chatbot personalities. Another future direction is to consider the mental health condition of the participant; as discussed in this work, depressed participants could find certain personalities undesired. We wish to find more detailed knowledge on how these conditions affect their opinions of other personality types.

Apart from the chatbot's personality, we hope to improve the chatbot more broadly. Currently, the chatbot shows a very limited conversation flow, with only one self-care method presented for each given method. We could make the chatbot more engaging to the user by increasing the number of methods given and the mental

health conditions covered. This could further be enhanced by offering the participant more freedom in the conversation, allowing more free-text entries and a non-linear conversation flow.

5.4 Limitations

Our study is not without limitations. First, we focused on two personality traits of the Big Five Personality Traits theory. We opted to include these two, as we could find evidence from related literature for the effects of these two in the context of our study. Now that the study setup is verified to work, other personality traits - openness to experience, agreeableness, and neuroticism - could be included to understand further the impacts of personalities within mental health self-care chatbots. Second, our study focused on single personality traits' high and low variants. While we admit this is a limitation in our work, it also paves the way for future research in this area. For example, the effects of combinations of different personality traits could be included, with more granular variations rather than the extreme low and high traits presented in this work. Last, our participants' cultural and language-based differences could have affected how they perceived the chatbot personalities. However, sourcing representative samples can get extremely costly, and examining the effects of cultures is a future research idea. In addition, we did not ask for demographic information related to the participants' mental health status, which could have affected their responses.

6 CONCLUSION

We investigated five chatbots with different personalities. We found user engagement to be the highest with the chatbot expressing a highly conscientiousness personality. After performing thematic analysis and clustering on the open feedback received from the participants, we found how the general preferences of chatbots and their personalities impact the opinion toward the bots. In addition, we recognised features of both high and low variants of the personalities which affected the user's opinions. Most notably, informativeness and confidence were noted as positive features, while monotonicity was noted as a negative feature for each personality. Based on our results, we believe that a well-performing personality could express conscientiousness and extraversion traits, however, it should be noted that the extreme low and high variants of these traits could often lead to negative reception. Our results shed light on how to design personalities for mental health chatbots.

ACKNOWLEDGMENTS

This research is connected to the GenZ strategic profiling project at the University of Oulu, supported by the Academy of Finland (project number 318930) and CRITICAL (Academy of Finland Strategic Research, 335729). Part of the work was also carried out with the support of Biocenter Oulu, spearhead of project ICON.

REFERENCES

[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research* 23, 1 (2021), e17828.

[2] Rangina Ahmad, Dominik Siemon, Ulrich Gnewuch, and Susanne Robra-Bissantz. 2022. Designing personality-adaptive conversational agents for mental health care. *Information Systems Frontiers* (2022), 1–21.

[3] Arfan Ahmed, Nashva Ali, Sarah Aziz, Alaa A Abd-Alrazaq, Asmaa Hassan, Mohamed Khalifa, Bushra Elhusein, Maram Ahmed, Mohamed Ali Siddig Ahmed, and Mowafa Househ. 2021. A review of mobile chatbot apps for anxiety and depression and their self-care features. *Computer Methods and Programs in Biomedicine Update* 1 (2021), 100012.

[4] Azy Barak and John M Grohol. 2011. Current and future trends in internet-supported mental health interventions. *Journal of Technology in Human Services* 29, 3 (2011), 155–196.

[5] Christopher Burr and Jessica Morley. 2020. Empowerment or engagement? Digital health technologies for mental healthcare. In *The 2019 Yearbook of the Digital Ethics Lab*. Springer, 67–88.

[6] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a chatbot for digital counselling. In *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017) 31*. 1–7.

[7] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2018. Best practices for designing chatbots in mental healthcare–A case study on iHelpr. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*. 1–5.

[8] D Robert Casares Jr. 2022. Embracing the podcast era: trends, opportunities, & implications for counselors. *Journal of Creativity in Mental Health* 17, 1 (2022), 123–138.

[9] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[10] Alton Ming Kai Chew, Ryan Ong, Hsien-Hsien Lei, Mallika Rajendram, Grisan KV, Swapna K Verma, Daniel Shuen Sheng Fung, Joseph Jern-yi Leong, and Dinesh Visva Gunasekeran. 2020. Digital health solutions for mental health disorders during COVID-19. *Frontiers in Psychiatry* (2020), 898.

[11] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (2019), 539–548.

[12] Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps. 2020. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Frontiers in digital health* 2 (2020), 41.

[13] Joseph C Dunn. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. (1973).

[14] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 301–305.

[15] Teresa M Evans, Lindsay Bira, Jazmin Beltran Gastelum, L Todd Weiss, and Nathan L Vanderford. 2018. Evidence for a mental health crisis in graduate education. *Nature biotechnology* 36, 3 (2018), 282–284.

[16] Tamás Fergencs and Florian Meier. 2021. Engagement and Usability of Conversational Search–A Study of a Medical Resource Center Chatbot. In *International Conference on Information*. Springer, 328–345.

[17] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? An exploratory interview study. In *International conference on internet science*. Springer, 194–208.

[18] Silvia Gabrielli, Silvia Rizzi, Giulia Bassi, Sara Carbone, Rosa Maimone, Michele Marchesoni, Stefano Forti, et al. 2021. Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. *JMIR mHealth and uHealth* 9, 5 (2021), e27965.

[19] Boris Galitsky. 2021. Adjusting chatbot conversation to user personality and mood. In *Artificial Intelligence for Customer Relationship Management*. Springer, 93–127.

[20] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[21] T Mark Harwood and Luciano L'Abate. 2009. Self-help in mental health: A critical review. (2009).

[22] Linwei He, Erkan Basar, Reinout W Wiers, Marjolijn L Antheunis, and Emiel Krahmer. 2022. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health* 22, 1 (2022), 1–14.

[23] Camilla Gudmundsen Høiland, Asbjørn Følstad, and Amela Karahasanovic. 2020. Hi, can I help? Exploring how to design a mental health chatbot for youths. *Human Technology* 16, 2 (2020), 139.

[24] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[25] Robert C Hsiung. 2000. The best of both worlds: An online self-help group hosted by a mental health professional. *CyberPsychology & Behavior* 3, 6 (2000), 935–950.

[26] David-Zacharie Issom, Marie-Dominique Hardy-Dessources, Marc Romana, Gunnar Hartvigsen, and Christian Lovis. 2021. Toward a Conversational Agent to Support the Self-Management of Adults and Young Adults With Sickle Cell Disease: Usability and Usefulness Study. *Frontiers in Digital Health* 3 (2021), 1.

[27] Oliver P John, Sanjay Srivastava, et al. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. (1999).

[28] Minjeong Kang. 2018. A Study of Chatbot Personality based on the Purposes of Chatbot. *The Journal of the Korea Contents Association* 18, 5 (2018), 319–329.

[29] Jeffrey A Kelly, Jeffrey M Kern, Betty G Kirkley, Jana N Patterson, and Terence M Keane. 1980. Reactions to assertive versus unassertive behavior: Differential effects for males and females and implications for assertiveness training. *Behavior therapy* 11, 5 (1980), 670–682.

[30] Daniel N Klein, Roman Kotov, and Sara J Bufferd. 2011. Personality and depression: explanatory models and review of the evidence. *Annual review of clinical psychology* 7 (2011), 269.

[31] Matthias Kraus, Philip Seldschopf, and Wolfgang Minker. 2021. Towards the development of a trustworthy chatbot for mental health applications. In *International Conference on Multimedia Modeling*. Springer, 354–366.

[32] Karolina Kuligowska. 2015. Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research* 2 (2015).

[33] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[34] Mike Lucock, Steve Gillard, Katie Adams, Lucy Simons, Rachel White, and Christine Edwards. 2011. Self-care in mental health services: a narrative review. *Health & social care in the community* 19, 6 (2011), 602–616.

[35] François Mairesse and Marilyn A Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction* 20, 3 (2010), 227–278.

[36] Joonas Moilanen, Aku Visuri, Elina Kuosmanen, Andy Alorwu, and Simo Hosio. 2022. Designing Personalities for Mental Health Conversational Agents. *Joint Proceedings of the ACM IUI Workshops* (2022).

[37] Kellie Morrissey and Jurek Kirakowski. 2013. 'Realness' in chatbots: establishing quantifiable criteria. In *International conference on human-computer interaction*. Springer, 87–96.

[38] Lea Müller, Jens Mattke, Christian Maier, Tim Weitzel, and Heinrich Graser. 2019. Chatbot acceptance: A latent profile analysis on individuals' trust in conversational agents. In *Proceedings of the 2019 on Computers and People Research Conference*. 35–42.

[39] Shannon B Myers, Alison C Sweeney, Victoria Popick, Kimberly Wesley, Amanda Bordfeld, and Randy Fingerhut. 2012. Self-care practices and perceived stress levels among psychology graduate students. *Training and Education in Professional Psychology* 6, 1 (2012), 55.

[40] Heather O'Brien, Paul Cairns, and Mark Hall. 2018. 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 (04 2018). https://doi.org/10.1016/j.ijhcs.2018.01.004

[41] Heather L O'Brien and Elaine G Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69.

[42] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. 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 (2018), 28–39.

[43] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[44] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, Bongwon Suh, et al. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of medical Internet research* 21, 4 (2019), e12231.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[46] Olga Perski, David Crane, Emma Beard, and Jamie Brown. 2019. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digital health* 5 (2019), 2055207619880676.

[47] Karen Pilkington and Lisa Susan Wieland. 2020. Self-care for anxiety and depression: a comparison of evidence from Cochrane reviews and practice to inform decision-making and priority-setting. *BMC complementary medicine and therapies* 20, 1 (2020), 1–15.

[48] Courtney Potts, Edel Ennis, RB Bond, MD Mulvenna, MF McTear, Kyle Boyd, Thomas Broderick, Martin Malcolm, Lauri Kuosmanen, Heidi Nieminen, et al. 2021. Chatbots to Support Mental Wellbeing of People Living in Rural Areas: Can User Groups Contribute to Co-design? *Journal of Technology in Behavioral Science* 6, 4 (2021), 652–665.

[49] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[50] Kelly Richards, C Campenni, and Janet Muse-Burke. 2010. Self-care and well-being in mental health professionals: The mediating effects of self-awareness and mindfulness. *Journal of Mental Health Counseling* 32, 3 (2010), 247–264.

[51] Rachel Richardson, David A Richards, and Michael Barkham. 2008. Self-help books for people with depression: a scoping review. *Journal of mental health* 17, 5 (2008), 543–552.

[52] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[53] Elayne Ruane, Sinead Farrell, and Anthony Ventresque. 2020. User perception of text-based chatbot personality. In *International Workshop on Chatbot Research and Design*. Springer, 32–47.

[54] Michael Shumanov and Lester Johnson. 2021. Making conversations with chatbots more personalized. *Computers in Human Behavior* 117 (2021), 106627.

[55] Tuva Lunde Smestad and Frode Volden. 2018. Chatbot personalities matters. In *International Conference on Internet Science*. Springer, 170–181.

[56] Kim Storrie, Kathy Ahern, and Anthony Tuckett. 2010. A systematic review: students with mental health problems—a growing problem. *International journal of nursing practice* 16, 1 (2010), 1–6.

[57] Sarah Theres Völkel and Lale Kaya. 2021. Examining User Preference for Agreeableness in Chatbots. In *CUI 2021-3rd Conference on Conversational User Interfaces*. 1–6.

[58] Sarah Theres Völkel, Ramona Schoedel, Lale Kaya, and Sven Mayer. 2022. User Perceptions of Extraversion in Chatbots after Repeated Use. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[59] Evan James Williams. 1949. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry* 2, 2 (1949), 149–168.

[60] Akihiro Yorita, Simon Egerton, Jodi Oakman, Carina Chan, and Naoyuki Kubota. 2019. Self-adapting chatbot personalities for better peer support. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 4094–4100.

[61] Debra M Zeifman. 2003. Predicting adult responses to infant distress: Adult characteristics associated with perceptions, emotional reactions, and timing of intervention. *Infant Mental Health Journal: Official Publication of The World Association for Infant Mental Health* 24, 6 (2003), 597–612.

[62] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.