2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)

Towards Understanding Emotional Intelligence for Behavior Change Chatbots

Asma Ghandeharioun
MIT Media Lab
Cambridge, MA, US
asma_gh@mit.edu

Daniel McDuff
Microsoft Research
Redmond, WA, US
damcduff@microsoft.com

Mary Czerwinski
Microsoft Research
Redmond, WA, US
marycz@microsoft.com

Kael Rowan
Microsoft Research
Redmond, WA, US
kael.rowan@microsoft.com

***Abstract*—A natural conversational interface that allows longitudinal symptom tracking would be extremely valuable in health/wellness applications. However, the task of designing emotionally-aware agents for behavior change is still poorly understood. In this paper, we present the design and evaluation of an emotion-aware chatbot that conducts experience sampling in an empathetic manner. We evaluate it through a human-subject experiment with N=39 participants over the course of a week. Our results show that extraverts preferred the emotion-aware chatbot significantly more than introverts. Also, participants reported a higher percentage of positive mood reports when interacting with the empathetic bot. Finally, we provide guidelines for the design of emotion-aware chatbots for potential use in mHealth contexts.**

***Index Terms*—Mobile applications, affective computing, experience sampling, agent, emotional intelligence, mental health.**

I. Introduction

The Experience Sampling Method (ESM) is a technique in which feelings or activities are recorded at the moment, either at randomly selected or predefined times [1]. ESM is less influenced by memory-bias, compared to retrospective self-reports. It has shown promise in understanding affect in context, linking it to events, and unraveling its temporal patterns [2]. ESM has proved to be valid and reliable [3], and particularly useful in symptom tracking in psychosomatic medicine [4]. There have been multiple efforts for streamlining ESM (e.g. [5]) and improving user engagement (e.g. [6]). However, other aspects of ESM design, such as delivery via an agent, have not been fully studied.

ESM provides a means for self-reflection and both unmediated and technology-mediated self-reflection have been shown to improve wellbeing [7]. In this paper, we aim to explore if we can further improve the positive effects of self-reflection on behavior change and one's sense of wellbeing by delivering ESM through an emotionally sentient agent. More specifically, we study the influence of interaction with an emotion-aware chatbot on one's mood.

Nowadays, virtual agents (VA) have been successfully deployed in multiple settings, ranging from education [8] to healthcare [9], [10], particularly in symptom tracking. While there is a strong focus on building applications to assess health, there is scientific evidence that making such applications empathetic plays a significant role in their acceptance and success and improves user experience [11]. An agent that is adaptive to the user's emotional state is perceived as more trustworthy, valuable, and intelligent [12]–[14].

However, when the context is more personal and nuanced, such as when the agent asks about the user's mood or mental wellbeing, there are further intricate details that need to be studied. What does the concept of emotional intelligence for an automated personal assistant mean in such contexts? Should the technology be designed for replicating human experience and emotional expressiveness? Or do users feel more comfortable opening up to a neutral and objective assistant? Does personality type influence one's preference for affective personal assistants? Does interaction with an emotionally-aware agent influence users' behavior in and of itself?

To address these questions, we designed and implemented an emotion-aware chatbot for the general population. This chatbot conducts experience sampling in an empathetic manner; this means that it is not only the instrument to capture self-reports, but also responds with emotionally appropriate conversations. It acknowledges the user's emotional state, similar to what an empathetic companion would do. We evaluated different aspects of this emotion-aware chatbot through a week-long randomized controlled trial with N=39 participants where we compared an emotionally appropriate condition (Emotion-Aware) to a neutral condition (Control). Our results showed that participants recorded a higher percentage of positive emotions using the empathetic bot compared to the neutral bot. Also, extraverts preferred the emotionally expressive bot significantly more than introverts. We present future directions of how such a chatbot can be used for longitudinal symptom tracking and for delivering mHealth interventions.

II. Related Work

The experience sampling method (ESM) [1] was born in 1970s with the advent of pagers. It is a validated technique used for capturing frequency, intensity, and overall patterns of one's emotional experience [3]. ESM alleviates people's inability to provide accurate retrospective information on their daily behavior and experience by capturing such information in the moment [15].

Despite the strengths of ESM, it also has limitations. This method puts heavy demand on research participants, which makes it a better fit for conscientious individuals rather than the general population [2]. There have been multiple efforts

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

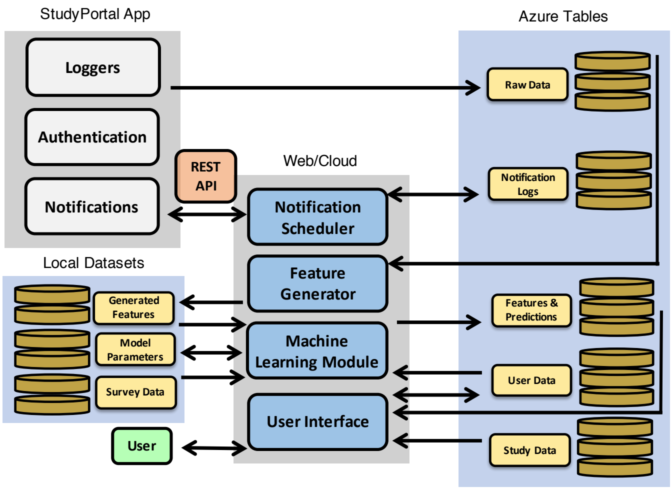
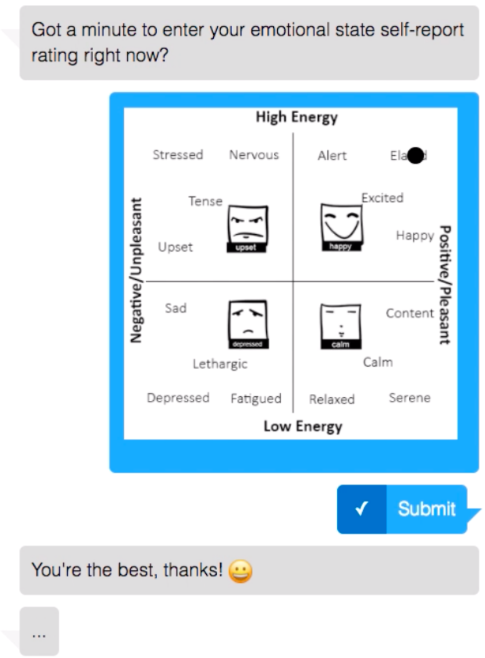


Fig. 1. System design

in improving the average user's engagement and compliance. Some researchers have tried to reduce the burden on the user by incorporating ESM more seamlessly into their daily pipelines, for example, by placing it in the phone unlock screen [5]. Other researchers have designed engaging games, making short questionnaires part of the game flow, and have validated ESM responses captured in the game in comparison to the traditional setting [6].

While these efforts have tried to address the issue of user engagement with ESM, the tone of delivery of ESM and its implications are not well studied. Specifically, ESM goes beyond being purely a method for capturing data. It provides a means for reflecting on one's past more objectively, which has the potential to improve psychological wellbeing and personal growth [16]. ESM is widely used in applications for improving mental health and wellbeing. Recently, personal assistants, chatbots, and virtual assistants (VAs) are are being used more often for conducting experience sampling. Therefore, it becomes imperative to study the characteristics of such an agent delivering ESM.

In health and mental health contexts, patients are sometimes reluctant to respond honestly. Extensive research on VAs shows that people are more comfortable disclosing health symptoms to VAs when they know they are automated and there is no human behind the scenes [14]. Interestingly, when VAs exhibit human qualities, they could further improve the relationship with the user. VAs are better at creating rapport when contingent on the human's responses [13]. Also, they are more successful in establishing trust when using relational conversational strategies [12]. It remains an open question how emotional-awareness and empathy in an agent which conducts mood experience sampling is perceived by users, if it is mediated by the user's personality, and if it affects the user's mood.



Got a minute to enter your emotional state self-report rating right now?

You're the best, thanks! 😀

...

Fig. 2. The visual design of the user interface. The bubbles are color coded to show if they are coming from the agent (gray) or have already been answered (blue).

III. System Design

We designed a conversational bot interface to conduct experience sampling. It specifically samples user's mood and is crafted to respond to it appropriately. For example, it responds positively to an expression of excitement, while responding sympathetically to an expression of stress. In this section we describe the system design in detail.

A. *Mobile Application*

The mobile application administers experience sampling and uses affective accompanying text. The app adjusts its behavior based on the group condition. Fig. 1 depicts the system design.

The mobile app consists of a web-based user interface (UI) (Fig. 2). The UI visualizes the conversations between the agent and the user. The content appears within bubbles that are left- or right-aligned based on the speaker (human or agent). Also, the bubbles are color coded to show if they are coming from the agent (gray), are prompts for the user to respond to (green), or have already been answered and are no longer editable (blue). To make the experience more realistic, the agent starts typing for one second before the text appears (see Fig. 2). The content is selected from the pool of scripted texts by a rule-based decision tree according to the group condition and the user's most recently recognized affective state.

The web-based UI is built upon the StudyPortal platform which is designed to handle different OS types [17]. In our case, StudyPortal is in charge of delivering notifications to the participants' phones and considering their history of previous responses.

TABLE I
AN EXAMPLE OF AFFECTIVE VS. NEUTRAL TEXTUAL CONTENT AFTER RECEIVING A MOOD REPORT. TL, TR, BL AND BR REFER TO THE SPATIAL LOCATIONS ON THE 2X2 RUSSELL CIRCUMPLEX MODEL OF EMOTION, E.G. TL REFERS TO TOP LEFT QUADRANT.

Condition	**Content**
Neutral	Thanks for the rating.
TL	Oh I see. I'm sorry to hear that. 😔
TR	I'm so proud of you 👏, thanks! 😊
BL	So sorry 😞, Hang in there and thanks.
BR	Nice. Thanks! 🙂

B. Experience Sampling

To make the chatbot design emotionally intelligent, it needs to reason about the user's current affective state [18]. To capture ground-truth emotion labels, we administered experience sampling five times a day and explicitly asked the participants to rate their mood. We adopted Russel's two-dimensional model of emotion [19] as our primary "gold-standard" mood measurement technique. This is one of the most prevalent and highly cited models of emotion and considers two dominant dimensions for mood: valence (pleasure - displeasure) and arousal (high energy - low energy). Horizontal and vertical axes correspond to valence and arousal respectively. To make it easier for users to self-report their mood, we included sample icons (visual cues) and emotional states (textual cues) that fall under the corresponding quadrants (See the experience sampling grid in Fig. 2). This visual grid captures continuous values between 0 and 1 for both valence and arousal.

C. Agent Dialog/Communications

For smooth communications between the agent and the user, we scripted dialog that was emotionally expressive and added emojis (from the set depicted in Fig. 3) when appropriate to better communicate emotions. In the emotional condition, each textual interaction had an average of 1.3 emojis, where there was an emoji per 6.5 words. In order to keep the content more realistic and engaging, we have scripted 6 different phrasings for each dialog interaction and randomly selected one when starting a conversation. For the Control condition, we scripted similar texts, but so as to be completely neutral without any expression of affect or use of emojis. Table I provides an example of the affective vs. neutral text when the user has just responded to an ESM prompt.

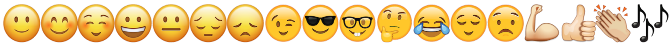


Fig. 3. Set of emojis used.

IV. HUMAN SUBJECTS

The study protocol was approved by the institutional review board at [anonymous institution]. Participants signed-up for the study online and were randomly assigned to the Emotion-Aware condition or the Control group. Forty one participants were recruited. One participant dropped out early due the app's phone battery usage. Another participant had previous knowledge about the hypotheses of the study and thus was excluded. Among N=39 participants that completed the protocol successfully, 7 were females and 32 were males. This included 17 full-time employees (FTE), 17 interns, and 5 external members or contractors. The participants' ages ranged between 16 and 49 (M=29.4, SD=7.9). Gift-card raffles were held at this week-long experiment for 50 dollars. Among active users, three were randomly selected as winners of the raffle [1].

To better understand our population in terms of their mental health and wellbeing status, we administered the Depression Anxiety Stress Scales (DASS) [20]. Overall, the participant population was generally healthy in terms of their mental health scores, as measured by DASS. This questionnaire includes a set of self-report scales designed to measure negative emotional states of depression, anxiety, and stress. We utilized the short version of DASS, which includes 21 items, 7 per scale. Each item is rated on a Likert scale, ranging between 0 (never) to 3 (almost always). Total DASS has possible scores of 0-63, and depression, anxiety, and stress sub-scales have possible scores of 0-21. See Table II for baseline values and their standard deviation among participants. Values under 4.5 for depression scale, under 3.5 for anxiety scale, and under 7 for stress scale are considered in the normal range.

Group		**DASS**	
	Depression	Anxiety	Stress
Emotion-Aware	3.6±4.1	2.3±3.3	4.7±3.6
Control	3.4±4.0	2.1±2.0	4.4±3.1

TABLE II
PARTICIPANTS' MENTAL HEALTH STATE AS MEASURED BY DASS. MEAN AND STANDARD DEVIATION OF BASELINE VALUES FOR DEPRESSION, ANXIETY, AND STRESS SUB-SCALES ARE INCLUDED.

V. EXPERIMENT: THE INFLUENCE OF INTERACTING WITH THE EMOTION-AWARE CHATBOT

Our research question is regarding the influence of interacting with an emotionally expressive bot compared to a neutral agent. Previous research has shown that interacting with a textual agent that shows minimal support of affect already helps to relieve strong negative affect. Also, when combined with a system that is designed to be frustrating, i.e., a game with unexpectedly long delays, participants prefer to continue to use such a system for longer if they are interacting with the emotional bot [21]. Subtle emotional expressiveness in agents has also been associated with higher trust and likability [22]. Other researchers have looked into the role of personality (introversion/extraversion dimension) in interacting with virtual agents [23]–[25]. Building upon previous research, we would like to explore the following questions: Does interacting with the Emotion-Aware chatbot improve users' self-reported mood? Do extraverts benefit more from adding emotional expressiveness to bots compared to introverts?

[1] This study was part of a larger experiment that lasted several weeks. Participants received $200 gift-cards for successfully completing all the experiments.

To answer these questions, we designed a one-week, longitudinal experiment. We randomized participants into two groups: Emotion-Aware and Control. The Emotion-Aware group had access to the mobile app that administered experience sampling. The app would generate 5 probes at random times throughout the day, between 9AM and 9PM, approximately every 2.5 hours, and we made sure that the probes were at least 30 minutes apart. Each experience sampling prompt started with a phone notification from the app, saying "Hi! Have a minute?". The participants could then click on the notification, or start the app by clicking on the application icon on the home screen. After the app opened, the chatbot would randomly select from a set of initial prompts that asked the participant to report his/her emotional state. Then, it would provide the experience sampling visual grid (See Fig. 2). After the participant responded to the prompt by dragging the indicator to express his/her emotional state, the chatbot would detect the selected quadrant, and randomly draw from a set of emotionally relevant phrases scripted for the respective quadrant. Note that the Control group had access to a similar interface, with the same methodology in triggering experience sampling probes. However, the responses to the experience sampling would always be selected from a pool of plain neutral texts without any expressive emotions. In summary, the difference between Emotion-Aware and Control users was in the responses that the participants received after reporting their mood. In the Control group, the app was only an instrument to capture data. Regardless of the user's selection, it would thank the user politely afterwards with a neutral tone; but in the Emotion-Aware group, the app would acknowledge the user's current status, respond appropriately, and resemble an empathetic companion.

A. Measures

To test our hypotheses regarding the interplay between personality and agent likability, we captured personality traits in the pre-study survey. We used well-validated measures of affect in the pre- and post-study surveys to capture affect and further validate the experience sampling data. We introduced satisfaction measures to study agent likability and user experience. Also, we analyzed the momentary mood sampled by the bot.

1) Big Five Personality Traits: The Big Five Personality Trait scale is a model based on common descriptors of personality that includes five factors: openness to experience, conscientiousness, extraversion, agreeableness, and Neuroticism [26]. The scale is composed of 44 items, where each item is rated on a Likert scale, ranging between 1 (strongly disagree) to 5 (strongly agree). Each personality factor is associated with 8-10 questions, thus possible scores are between 8-50.

2) Positive and Negative Affect Schedule: The Positive and Negative Affect Schedule (PANAS) consists of 20 words that describe different emotions [27]. Half of the items indicate positive affect (PA) and half indicate negative affect (NA). Items are rated on a Likert scale, ranging from 1 (very slightly or not at all) to 5 (extremely). PA and NA are calculated separately and each range between 10-50. the PA/NA ratio is another commonly used measure derived from PANAS. PANAS has been used to capture affect in different time scale ranges. These include momentary, daily, over the past few days, weekly, for the past few weeks, yearly, and general affect. In our study, we have used PANAS to capture affect over the past week.

3) User Preference: We assessed satisfaction and efficacy of the system through different questions using a Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). These questions asked about the agent's likability, intelligence, and the appropriateness of its "tone". Questions were also asked about user preference for continuing to interact with the agent, and his/her improvement in awareness of daily emotions. They also asked if the notifications from the app where too frequent. Also, we included an open-ended question at the end of the week for general comments.

4) Experience Sampling: Using the visual experience sampling grid, we capture valence (v) and arousal (a) on a continuous scale, $v, a \in [0.0, 1.0]$. To match the discrete categories of the pool of scripted text explained in Section III-C, we discretize v to have positive and negative valence:

$$\hat{v} = \begin{cases} Negative & \text{for} \quad v < 0.5 \\ Positive & \text{otherwise} \end{cases} \tag{1}$$

We also discretize a to have high and low arousal.

$$\hat{a} = \begin{cases} Low & \text{for} \quad a < 0.5 \\ High & \text{otherwise} \end{cases} \tag{2}$$

The 4 possible combinations of $\hat{v}$ and $\hat{a}$ are mapped to the 4 quadrants on the visual grid: Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR).

B. Results

1) User Perception of the Emotion-Aware Chatbot: Given that the Emotion-Aware chatbot is emotionally expressive, we questioned whether different personality types would prefer the agent more or less. Specifically, do extraverts prefer the Emotion-Aware chatbot more than intraverts? To answer this question, we discretized the Big Five extraversion scores into binary values: extravert (above median) vs. introvert (below median). Focusing only on the Emotion-Aware group, we compared the overall likability of the agent as averaged across all likability questions. An independent-samples t-test showed a significant difference in the overall likability scores for extraverts (M=5.17, SD=.91) and introverts (M=4.43, SD=.55); t(17)=2.08, p=.05 (Fig. 4)[2].

2) Influence of the Emotion-Aware Chatbot on Mood Reports: To answer this question, we compared the daily percentage of positive and negative ESM mood reports across groups. Granular daily self-reported emotion samples revealed significant differences across Emotion-Aware and Control groups. Fig. 5 shows the average percentage of the positive and

[2] The Pearson correlation coefficient and p-value between extraversion score and agent likability are the following: r=0.432, p=0.065.

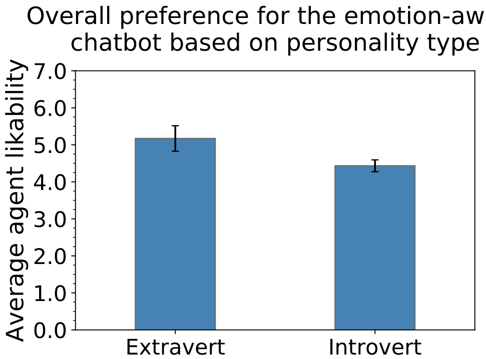


Fig. 4. Overall agent preference based on personality type with one standard error bars. Extraverts preferred the Emotion-Aware chatbot significantly more.

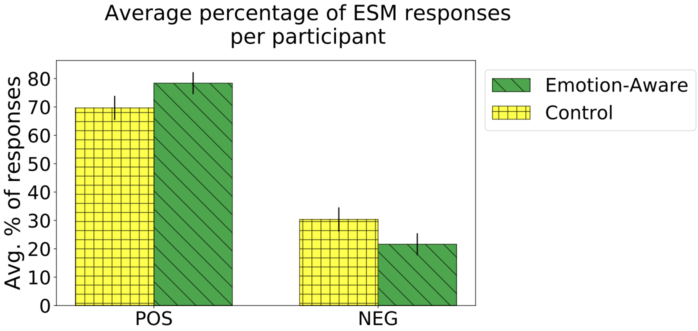


Fig. 5. Average percentage of ESM responses per participant with one standard error bars. POS: positive valence quadrants on the 2x2 Russell circumplex model of emotion (top right and bottom right), NEG: negative valence quadrants (top left and bottom left). Participants in the Emotion-Aware condition reported higher percentage of positive emotions.

negative ESM self-reports per participant. An independent-samples t-test was conducted to compare percentage of positive emotions reported daily between the Emotion-Aware and the Control conditions. There was a significant difference in the percentage of positive emotions for Emotion-Aware (M=80.55, SD=3.65) and Control (M=69.08, SD=4.16) conditions; t(37)=2.74, p = .009 [3]. The Emotion-Aware group reported a higher percentage of positive emotions (Top Right and Bottom Right affective quadrants in Russell's 2x2 model) and a lower percentage of negative emotions (Top Left and Bottom Left quadrants) compared to the Control group. Note that the weekly PANAS survey and the daily ESM are capturing instantaneous vs. weekly mood which are different by definition, but our analysis showed that the PA score derived from PANAS and the total number of positive self-reports over the course of the week were correlated (Pearson r=0.217, p=0.020). However, looking more closely at the influence of the Emotion-Aware chatbot on PA from weekly PANAS scores, a 2 (group) x 2 (pre-post PA) RM-ANOVA did not show a significant group x pre-post interaction [4].

3) User Feedback: Several participants reported interacting with the app as "an interesting experience" (pa070), "pretty quick" (pa081) and "fun" (pa045).

Some mentioned that the experience sampling made them more self-aware, or amplified their emotional state; pa050: "notifications from the agent amplified how I was feeling."; pa064: "It is a good exercise to periodically reflect on my emotions. I really like that aspect."; pa063 mentioned surveys acted as a feedback loop, too: "answering this survey forces me to define an emotional profile, to which I somehow become committed or identify with, which in turn influences my daily ratings.".

Originally, we did not fully absorb the extensive role of the bot on self-awareness, but the overwhelming feedback from participants recognizing how it influenced their behavior highlighted that any behavior change application needs to support self-reflection. This result is in line with previous research findings, suggesting that self-reflection is an important part of behavior change and has the potential to improve wellbeing and mood [7], [28]. However, it is worth mentioning that encouraging users to self-reflect should be done in moderation. There are downsides with interrupting users too frequently to self-report. First, the possible consequences should be considered. Some participants mentioned that an extremely high frequency of self-reflection could be harmful in certain circumstances; pa050: "When I was stressed/worried at work and saw that I had to report on my feelings then those feelings felt more intense."; pa088: "I'm not sure if thinking about my feeling so many times in a day is a good thing. I realized that I've been picking happy only infrequently, which made me a little sad.". Second, there can simply be high missing data rate. pa011: "I frequently miss notifications.". Therefore, if the application is solely relying on users' self-reported data, it will significantly hurt performance. Partial or full automation could help address these caveats.

Also, the open responses shed light on what could be improved. Some participants mentioned the chatbot's responses in the Emotion-Aware condition were exaggerated and unable to capture subtle or nuanced emotional states; pa073: "The agent's responses seemed very *narrow* responding with just a few generic phrases to my self-assessments[...]. Although the emotion quadrant consists of four squares, the actual coordinates within each square have a wide range of meanings [...]. However, the agent did not appear to respond particularly differently [to the intensity of the reported emotion]."; pa028: "The reactions to input could be better. They seem to come from the four basic zones (+- x/y), and the feedback from the bot doesn't indicate that the input I send is any more granular than that."; pa067: "The agent seems to respond as if your emotional state is either great or terrible[...]. It would be nice if it could adopt a more neutral tone in some circumstances. It's just kind of weird when it says something like *Bummer* when I report that I'm feeling [almost] neutral."; pa031: "It would be nice if the agent had better design and some kind of persona. The way it is now seems simplistic (though, still useful in the sense of reminding)".

[3] Since the percentage of negative emotions is 100 minus the percentage of positive emotions, similarly there was a significant difference in the percentage of negative emotions for Emotion-Aware (M=19.45, SD=3.65) and Control (M=30.92, SD=4.16) conditions; t(37)=-2.74, p = .009.

[4] F=2.430, p=.098

We need to emphasize that there was overwhelming feedback from participants highlighting that they wanted to be able to enter more nuanced emotional state self-report data, but that the bot's responses were coarse and rough and did not account for the subtleties in their reports. In other words, users wanted to select very particular feelings during self-report, and they wanted an agent that reflected that level of precision. This lack of granularity bothered our users, even though they still liked the reminding facility.

Some of the responses mentioned the difficulty users had in expressing precise emotional reports. This could be due to the UI; pa078: "It's difficult to be precise in positioning the dot on the axes.". It could also be due to difficulty identifying emotions and mapping them to the quadrants; pa061: "Sometimes I found it hard to describe my feelings", pa052: "[the] subtle changes in my emotions are not being captured by my current way of recording it." Providing better user interface support for users to reflect upon and enter their emotions remains an important issue for future work.

VI. Discussion

A. Empathetic Experience Sampling and Mood

Our results showed that providing an emotionally appropriate response when conducting experience sampling, similar to what happens in a successful human-human interaction, resulted in a higher percentage of positive responses being recorded. However, interaction with the agent did not significantly influence positive and negative affect, as captured by the weekly PANAS surveys. We have three possible interpretations. First, the influence of the agent may be subtle and, since it only appeared in granular experience sampling about five times a day, was possibly not enough to show its influence over one week. Second, the emotional expressiveness of the bot may have resulted in self-report bias. One participant said: "Sometimes, the responses when the mood is marked as negative seem somewhat validating or disheartening, subconsciously making me reluctant to mark my mood as such." This suggests that the affirmative response from the agent might have affected the ratio of missing self-reports asymmetrically for negative vs. positive samples. Third, our population was overall quite healthy and happy–improved positive affect in a clinical sense would probably be unlikely. Further studies are needed to get a deeper understanding about empathetic experience sampling to tease these issues apart.

B. Personality and Preference for an Affective Agent

Our findings revealed that there is value in adding emotional understanding and expression to conversational agents. The emotionally expressive bot was generally liked (on average more than 4 from on a 7-scale Likert scale). However, extraversion was an important personality factor influencing the likability of the agent: extraverts' average likability measures were significantly higher than introverts'. This suggests that certain personality types may benefit more from adding emotional intelligence or expressiveness to conversational agents.

C. Design Guidelines

A synthesis of users' feedback shed light on guidelines that could prove useful for designing affective conversational bots.

Emotional intelligence is sometimes a neutral response. Feedback from participants revealed that providing emotionally expressive responses to subtle emotions decreased the perception of emotional intelligence of the bot. For example, expressing sympathy in response to minor expressions of sadness was received as unnecessary exaggeration. Instead, a neutral or nuanced response was preferred. We learned that low intensity emotions should be responded to with more subtle and neutral interactions.

Behavior change applications benefit from supporting self-reflection. The overwhelming feedback from our participants shed light on the influence of self-reflection on behavior change. We suggest that any behavior change application should consider supporting self-reflection to improve the efficacy of the system. We need to highlight that supporting self-reflection does not necessarily require sole reliance on the user to provide data frequently. It rather means intelligent support systems could provide opportunities for the user to self-reflect at the right pace and frequency, while still being able to function without needing high rates of user data.

D. Limitations

We manually scripted all the textual interactions. Though we created multiple phrases with similar, but slightly different messages, their occurrence soon became "expected" over the course of the study. In the future, we would like to use machine learning to automate the intervention text generation and make it emotionally expressive by adding emojis or sentiment that works for an individual according to personal preference and context [29]–[31]. This work lies on the boundary between a data-gathering and a behavior change tool. It is worth mentioning that the emotional expressiveness of the Emotion-Aware bot may have confounded the experience being sampled. Further studies are required to fully investigate validity of ESM responses, with and without these modifications. Further replication of this work using larger populations is encouraged.

VII. Conclusions and Future work

In this paper, we quantitatively and qualitatively evaluated an emotion-aware chatbot for conducting experience sampling, over the course of a week, with N=39 participants. Our results show that an emotionally expressive agent is likable, particularly to extraverts. Furthermore, an emotionally sentient agent such as we deployed has the potential to improve positive affect and reduce negative affect. We identified several design guidelines for future work. Specifically, we found that an emotionally appropriate response is sometimes neutral and that support of self-reflection is a crucial part of behavior-change applications. In the future, we would like to extend our system to detect a user's mood from passive smartphone sensor data and use automatically predicted emotional states to drive emotional dialog and relevant micro-interventions just-in-time.

References

[1] S. Prescott and M. Csikszentmihalyi, "Environmental effects on cognitive and affective states: The experiential time sampling approach," *Social Behavior and Personality: an international journal*, vol. 9, no. 1, pp. 23–32, 1981.

[2] C. N. Scollon, C.-K. Prieto, and E. Diener, "Experience sampling: promises and pitfalls, strength and weaknesses," in *Assessing well-being*. Springer, 2009, pp. 157–180.

[3] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," in *Flow and the foundations of positive psychology*. Springer, 2014, pp. 35–54.

[4] T. S. Conner and L. F. Barrett, "Trends in ambulatory self-report: the role of momentary experience in psychosomatic medicine," *Psychosomatic medicine*, vol. 74, no. 4, p. 327, 2012.

[5] A. Ghandeharioun, A. Azaria, S. Taylor, and R. W. Picard, ""kind and grateful": a context-sensitive smartphone app utilizing inspirational content to promote gratitude," *Psychology of well-being*, vol. 6, no. 1, pp. 1–21, 2016.

[6] S. Taylor, C. Ferguson, F. Peng, M. Schoeneich, and R. W. Picard, "Use of in-game rewards to motivate daily self-report compliance: Randomized controlled trial," *Journal of medical Internet research*, vol. 21, no. 1, p. e11683, 2019.

[7] E. Isaacs, A. Konrad, A. Walendowski, T. Lennig, V. Hollis, and S. Whittaker, "Echoes from the past: how technology mediated reflection improves well-being," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 1071–1080.

[8] S. D'Mello, R. W. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," *IEEE Intelligent Systems*, vol. 22, no. 4, 2007.

[9] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[10] L. Ring, T. Bickmore, and P. Pedrelli, "An affectively aware virtual therapist for depression counseling," in *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) workshop on Computing and Mental Health*, 2016.

[11] K. Liu and R. W. Picard, "Embedded empathy in continuous, interactive health assessment," in *CHI Workshop on HCI Challenges in Health Assessment*, vol. 1, no. 2. Citeseer, 2005, p. 3.

[12] T. Bickmore and J. Cassell, "Relational agents: a model and implementation of building user trust," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2001, pp. 396–403.

[13] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *International Workshop on Intelligent Virtual Agents*. Springer, 2007, pp. 125–138.

[14] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "Its only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, 2014.

[15] H. R. Bernard, P. Killworth, D. Kronenfeld, and L. Sailer, "The problem of informant accuracy: The validity of retrospective data," *Annual review of anthropology*, vol. 13, no. 1, pp. 495–517, 1984.

[16] F. B. Bryant, C. M. Smart, and S. P. King, "Using the past to enhance the present: Boosting happiness through positive reminiscence," *Journal of Happiness Studies*, vol. 6, no. 3, pp. 227–260, 2005.

[17] K. Rowan, "Studyportal api," http://studyservice.cloudapp.net/docs/, 2013, online, Retrieved August 14, 2017.

[18] S. Jeong and C. L. Breazeal, "Improving smartphone users' affect and wellbeing with personalized positive psychology interventions," in *Proceedings of the Fourth International Conference on Human Agent Interaction*. ACM, 2016, pp. 131–137.

[19] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[20] P. F. Lovibond and S. H. Lovibond, "The structure of negative emotional states: Comparison of the depression anxiety stress scales (dass) with the beck depression and anxiety inventories," *Behaviour research and therapy*, vol. 33, no. 3, pp. 335–343, 1995.

[21] J. Klein, Y. Moon, and R. W. Picard, "This computer responds to user frustration: Theory, design, and results," *Interacting with computers*, vol. 14, no. 2, pp. 119–140, 2002.

[22] T. Bickmore and R. Picard, "Subtle expressivity by relational agents," in *Proceedings of the CHI 2003 Workshop on Subtle Expressivity for Characters and Robots*, 2003.

[23] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996.

[24] C. Nass and K. M. Lee, "Does computer-generated speech manifest personality? an experimental test of similarity-attraction," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2000, pp. 329–336.

[25] S. Buisine and J.-C. Martin, "The influence of users personality and gender on the processing of virtual agents multimodal behavior," *Advances in Psychology Research*, vol. 65, pp. 1–14, 2010.

[26] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[27] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[28] L. Taber and S. Whittaker, "Personality depends on the medium: differences in self-perception on snapchat, facebook and offline," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 607.

[29] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," *arXiv preprint arXiv:1708.00524*, 2017.

[30] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1587–1596.

[31] A. Ghandeharioun, J. Shen, N. Jaques, C. Ferguson, N. Jones, A. Lapedriza, and R. Picard, "Approximating interactive human evaluation with self-play for open-domain dialog systems," *arXiv preprint arXiv:*, 2019.