Can a Chatbot Comfort Humans? Studying the Impact of a Supportive Chatbot on Users’ Self-Perceived Stress

Lenin Medeiros, Tibor Bosse, and Charlotte Gerritsen

Abstract—This article is part of a project that explores the potential of chatbots for providing online emotional support to humans tailored to stressors. Based on a number of empirical studies, we have developed a socially interactive agent able to have simple dialogues with stressed humans seeking for emotional support. In the current article, we address the question to what extent this chatbot is effective in helping users cope with stressful situations. To this end, we present a study in which participants were asked to interact with our proposed chatbot for three days. Participants are distributed over the following three conditions—namely: 1) receiving support from the chatbot, knowing the support is computer-generated; 2) receiving support from the chatbot, while believing the support is human-generated; 3) not receiving any support. During the three days, participants’ self-reported stress levels are measured on a daily basis before and after each interaction. Results indicate that the best results are obtained in the ‘human’ condition, while the worst results are obtained in the ‘computer’ condition. These findings lead us to conclude that the presumed sender of a stress support message (i.e., a human or a computer) might be more important than the content of the message.

Index Terms—Chatbots, computer-generated emotional support, emotion regulation, emotionally supportive agents, online experiment, stress coping strategies.

I. INTRODUCTION

EVEry human being is likely to face undesired situations in life that can significantly affect one’s mood, even though the individual does not necessarily suffer from a mental disorder such as depression. Examples of such situations, to which we refer in this article as everyday stressors, are employees facing problems at work, students dealing with final exams, families concerned about relatives with medical conditions and many other challenges in life. All these situations have the potential to make humans experience stress.

Fortunately, people may apply a variety of coping mechanisms to reduce the stress they experience. Since the advent of social media, one of the most frequently used coping mechanisms is to share personal problems via online social networks; by posting their problems online and receiving supportive replies from peers, people may indeed feel a bit better [1]. Obviously, this phenomenon is not new; in an offline context, it has been studied for decades, and is defined by researchers as social support [2]. Moreover, substantial evidence has been found that this form of support might indeed have a positive impact on experienced stress [3]. Hence, it is worth exploring to what extent social support can be effective in an online situation.

However, the concept of online social support has a number of limitations. First of all, immediate support is not always available. For example, when someone is looking for support in the middle of the night, most peers are likely to be offline. Moreover, some people simply do not have many friends and relatives who are able to support them [4]. Second, people may be reluctant to share their personal problems with peers. Several studies have shown that the threshold of sharing (negative) information about one’s private life can be very high, even to the extent that people prefer sharing their problems with a computer than with another human [5]. Third, peers are not always willing or able to provide social support, in particular because it may cost them too many emotional resources. Indeed, it is known from the literature that therapists could absorb negative emotions after having contact with many patients, which brings them at risk to develop secondary trauma themselves [6].

To tackle the limitations described above, we proposed in previous work the idea of an emotionally supportive agent in the form of an online chatbot. The core of the developed chatbot is an algorithm that is able to: 1) classify incoming messages from users into various “stress categories” (e.g., health issues, relationship problems); and 2) generate supportive messages that are tailored to the particular stressor [7]. Theoretically, the system is based on an empirical study on which types of daily

1It is important to state that this solution neither replaces a real therapist nor is an attempt to simulate one. Instead, this chatbot is an effort to simulate an online social network peer providing social support in the form of text messages to stressed friends. Hence, if incorporated in a real world application, its target users would be healthy individuals, not people with mental health conditions.
life problems people share online [8], as well as on the emotion regulation theory by Gross [9].

As a next step in the project, this article aims to investigate to what extent the developed chatbot is able to effectively reduce stress in its users. Additionally, we aim to understand whether it makes a difference if users know the supportive messages are computer-generated or not. To this end, a between-subjects experiment is conducted where groups of participants are requested to interact with different variants of the system for three days in a row, while reporting their subjective levels of stress before and after each interaction.

The rest of this article is organized as follows. Section II presents an overview of related studies that provide a theoretical basis for the current research. Section III provides a technical description of the stress-support chatbot developed previously that is used as stimulus material in our study. Section IV explains the method used to investigate the effectiveness of this piece of software in helping users cope with stress. Section V describes the results of the experiment. Finally, Section VI concludes this article with a summary and discussion about the main findings, pointing out weaknesses and strengths of this study as well as ideas for future work.

II. BACKGROUND

Social interactions between humans and computers have been studied since a number of decades. In their seminal work, Reeves and Nass [10] were the first to demonstrate that individuals tend to treat computers the same way as they treat human beings. In this work, participants turned out to be as polite to computers as they would be to real humans after completing the following task: first they were tutored by a computer about a given topic, and after that they had to make a test about what they had learned. After the test, the participants received feedback from the computer regarding its own performance when helping them with the task. All the participants were told by the computer that it had successfully accomplished the mission of tutoring them. Next, the participants had to evaluate this computer. They were separated into two groups: the first group evaluated the computer using the same machine that had just tutored them, whilst the second group did so on a different machine. It turned out that, on average, the first group came up with a better evaluation of the computer, for instance using words such as friendly, helpful, and competent.

Other studies have reached similar conclusions. One recent example is described by Carolus et al. [11]: they performed a similar experiment, using smartphones instead of desktop computers. Indeed, their results were along the same lines as the study by Reeves and Nass, i.e., these devices elicit social norms of politeness among their users. Furthermore, there is evidence that humans sometimes treat computers that are acting as teammates in the same manner as they would treat real colleagues, as a number of papers suggest (e.g., [12] and [13]). In addition to individuals dealing with computers in work-related environments, it has also been demonstrated that humans may engage in other types of social interactions with machines. Kanda et al. [14], for instance, conclude that interactive robots could establish social relationships with humans, based on an experiment in which they demonstrated that children developed friendly relationships with robots in elementary school. Finally, another example is a study conducted by Schrammel et al. [15] in which participants watched virtual characters displaying some social cues typically present in humans, such as facial expressions and interactive gaze. Facial muscle activity of the participants was measured during these interactions, which led the researchers to conclude that the emotional response humans show toward interactive machines is potentially the same as when they are interacting with real individuals.

If humans can have social responses toward computers that are similar to the ones they have toward real individuals and social support, as mentioned earlier, provided by humans to other humans can make people cope better with challenges in life, then one can argue that the scientific community could investigate the use of computers to provide social and emotional support (e.g., by “listening” to the problems faced by a user and coming up with supportive words to make him or her feel better). This would be a particularly appealing alternative in cases when people do not have access to effective social support from peers online or when they just do not want to use this coping mechanism, as indicated in Section I.

Indeed, the concept of computer-generated emotional support (CGES) has recently appeared in the literature, with promising results. For instance, Morris et al. [16] state that, even though people prefer emotional support provided by real peers, in general they tend to accept CGES. Following this idea, Kindness et al. [17], presented a study in which CGES is used to support medical emergency personnel that experiences stress on the job. They developed an algorithm to map stressors onto categories of supportive messages humans would follow to help their peers cope with stress in these situations. Moreover, in the mental health domain, Woebot and Tess are examples of real applications currently available on the market aiming to provide emotional support to users that experience self-identified symptoms of stress, anxiety, and depression. Results of two empirical studies, respectively, demonstrate that these two examples of emotion-aware conversational agents seem to be effective in reducing such symptoms among participants [18], [19]. Note that these systems are based on cognitive behavioral therapy (CBT) [20], a well-established psycho-social intervention. This is different from the approach proposed in this article, which aims to develop an “artificial online friend” that provides emotional support in daily-life situations, rather than a virtual therapist providing support to mental health patients. Finally, more examples of virtual agents providing emotional support are discussed in [17], leading to the conclusion that this type of support can indeed have a positive effect on a user’s mood.

To explore to what extent artificial agents can effectively provide CGES to peers in an online social network, we proposed in early stages of this project the idea of an “artificial friend” [4], [7]. As a first step to develop such a system, online social network data were analyzed to identify the most common stressors shared by the respective users. This resulted in categories such as “health issues” and “relationship
problems.” Next, similar data were analyzed to identify the typical supportive strategies used by peers when helping stressed friends via social media [8]. These strategies were then categorized in accordance with Gross’s theory on emotion self-regulation² [9]. This theory describes how humans cope with their own emotions via a number of mechanisms—namely, 1) situation selection—avoiding potentially stressful situations; 2) situation modification—modifying a given stressful situation; 3) attentional deployment—stop focusing on a given stressful event; 4) cognitive change—reinterpreting the emotion-eliciting meaning of a given stressful situation; 5) response modulation—controlling the emotional response to a given stressor [9]. Originally, these mechanisms were identified as strategies that people use to regulate their own emotions; however, they can also be used in an interpersonal way, i.e., to regulate someone else’s emotions (as discussed, for instance, in [23]). In addition, a sixth strategy was, to which we refer as general emotional support—this strategy aims to support peers by expressing general words showing care and empathy [24].

Hence, at this point we had two categorizations, namely a list of stressor categories people share online and a list of supportive strategies people use online. In the next step, the data were analyzed in order to map these lists to each other. As a result, a contingency table was established, indicating which emotion regulation strategy is most appropriate for which type of stressor (e.g., in case the stressor is the end of a relationship, then the most suitable strategy is attentional deployment). Based on this table, an algorithm was implemented to generate appropriate supportive responses tailored to text messages representing stressors. This algorithm is the core of a chatbot that, by means of text mining and natural language processing, is able to carry out small dialogues with social media users. During such dialogues, the chatbot aims to understand when users share stressors and send back emotionally supportive messages to the respective senders [7]. More details about the system are presented next in Section III-A.

As a next step in the project, we aim to investigate to what extent the developed chatbot, which provides emotional support tailored to everyday stressors, is effective in reducing users’ experienced stress. Second, and inspired by the finding that people typically prefer human-generated emotional support over CGES [16], we aim to investigate whether or not it matters if users are aware that the support is generated by a computer. This is an interesting question since it may well be the case that, even though the support generated by the system is similar to the support that human beings provide each other online, people do not appreciate it simply because they know it comes from a machine. By systematically manipulating people’s expectations about the source of the support, we will attempt to answer this question.

²We selected this theory as the theoretical framework behind our work because of the following. 1) It is a well-established theory, which has been used before in the context of computer-mediated stress regulation (see [21] for a scoping review). 2) The authors have developed a computational model of the theory in the past [22], which makes it very suitable to incorporate within a chatbot. A more extensive discussion on the considerations underlying this choice can be found in [4].

In summary, the research has two main aims, namely to study 1) the effect of interacting with our chatbot on (self-reported) stress; and 2) whether it matters if users are aware that the support is computer-generated. In line with [25], we define stress as a state of high arousal and low (negative) valence. To answer our research questions, a between subjects experiment has been designed, in which participants are asked to share their everyday stressful experiences online, while being allocated randomly to one of three conditions: Group B (bot), Group H (human), and a Control Group. In both Groups B and H, participants receive support on their stressful experiences from the developed chatbot. However, in Group B participants are informed that the support is computer-generated, whereas participants in Group H are told that the support is provided by another participant. In the Control Group, participants do not receive any form of support. More information about the design is provided in Section IV-A.

Following results reported in similar studies [17]–[19], we hypothesize that after interacting with the chatbot, people will have a lower state of stress than before. This hypothesis is independent of people’s belief about the nature of the communication partner; hence, we expect it to hold both for Group B and Group H. Based on this line of reasoning, the following hypotheses have been formulated.

1) Hypothesis 1 (H1): For participants in Groups B and H, the self-reported levels of arousal after interacting with the chatbot are lower than before interacting with the chatbot.

2) Hypothesis 2 (H2): For participants in Group B and H, the self-reported levels of valence after interacting with the chatbot are higher than before interacting with the chatbot.

Furthermore, as people typically prefer human-generated support over computer-generated support [16], we expect that the reduction in stress postulated above will be larger for people who think they are talking to another human (i.e., Group H) than for people who know they are talking to a computer (i.e., Group B). Nevertheless, we expect that this effect will still be larger in Group B than for the people who do not receive any support at all (i.e., the Control Group). This leads to the following additional hypotheses.

1) Hypothesis 3 (H3): The decrease of arousal postulated in H1 is larger in Group H than in Group B and it is larger in Group B than in the Control Group.

2) Hypothesis 4 (H4): The increase of valence postulated in H2 is larger in Group H than in Group B and it is larger in Group B than in the Control Group.

III. TECHNICAL SPECIFICATIONS

As explained earlier, the stimulus material used in our experiment consists of an emotionally supportive chatbot that uses the strategies from Gross’ emotion regulation theory to provide supportive text messages tailored to the stressful situations shared by participants. In this section, the technical implementation of the chatbot is described. First, the algorithm used by the chatbot to determine which text message to send based on the inputs
provided by users is explained. After that, a description of the software architecture is provided.

A. Algorithm

The algorithm which is the core of this application consists of a mapping of stressors to support strategies. Based on a previous empirical study [8], a number of categories of stressors were identified that are shared most frequently on social media. In particular, there are seven clusters of stressors:

1) financial—issues with money, bills, debts, etc., (as in “Hello - i’m a bit worried about how I will pay next month’s rent because I received an unforeseen expense today”);
2) health—health-related situations (as in “I fell on some ice, and I really badly hurt my back and leg”);
3) grief—situations in which a loved-one passed away (as in “My dog passed away last night, he was 10yo... it is so hard to not having him with me anymore”);
4) relationships—problems involving interpersonal relationships (as in “I had a bad argument with my girlfriend today. She kept insisting she wanted to buy soda when I prefer drinking water…”);
5) school—matters related to school, college, university, etc. (as in “thinking about the start of my master thesis stresses me out. testing will start tomorrow”);
6) work—stressful situations about work environments (as in “I’ve been really frustrated with work today. Had to rely on other people and they have let me down”);
7) other—when the stressor does not fit into any of the previous categories (as in “everyday I go to the mail to see if I finally got the package I ordered, and it still didn’t arrive.”)

Note that the messages used as examples were taken literally from this reported experiment, i.e., they are messages sent by the participants to the proposed chatbot in the experiment defined in Section IV.

In addition to the stressor categories, support strategies were also collected in order to match them to specific stressors. As explained earlier, these strategies are mostly based on Gross’s [9] study about emotion self-regulation. One of them can be mapped into a concept of emotional support inspired by Heaney and Israel [4], [24]. In total, based on a previous investigation [4], four different coping strategies are used in our algorithm:

1) attentional deployment (AD)—suggesting to move one’s focus of attention away from the stressor (as in “forget this and think about something else”);
2) cognitive change (CC)—suggesting to re-interpret the stressful situation from a different perspective (as in “look at the bright side: this is not too bad”);
3) general emotional support (GES)—simply expressing care, trust, love or empathy (as in “I am so sorry to hear this from you, I am here for you whenever you need me, take care”);
4) situation modification (SM)—suggesting to take some practical action to change a given stressful situation (as in “do something about it, I am pretty sure you can change this”).

Based on a previous study on which strategy is most appropriate in which situation [8], these support strategies are related to stressors as depicted in Table I.

The code of the chatbot which can be accessed via a GitHub repository\(^3\) contains a set of pre-established supportive messages written by the researchers that are in accordance with the emotion regulation strategies defined earlier. The messages are divided into groups (25 messages per group), where each group is related to one support strategy. There is a pointer to store what was the last message sent to the user to avoid repetitions. Details about the structure of the proposed software which uses this algorithm to support stressed users appear in the following.

B. Architecture

Participants are able to interact with the chatbot via Facebook. For the experiment, two versions of the same chatbot have been developed; the only difference between them was the time needed to send back supportive messages to users. A delay was included in one version making the users wait for at least 15 seconds more than the other version to see the bot’s reply. This delay was incorporated to make participants believe the chatbot was in fact another participant in the experiment. More details about the experiment are presented in Section IV. A visual representation of the architecture of the chatbot is depicted in Fig. 1 and the respective explanation is presented below.

The architecture consists of the following five modules:

1) a user interface module (represented by the Facebook box);
2) a sentiment analysis module to extract sentiment from messages sent by users (represented by the IBM W NLU box);
3) a dialogue flow management module to determine how the chatbot should respond based on the messages sent by users (represented by the IBM W Assistant box);
4) a database module to store information (represented by the mLBox box);
5) a central module to connect all the modules which is intended to stay available in the cloud waiting for messages sent by the users (represented by the Webhook box). The dashed box entitled Heroku represents the cloud platform service where the central module is hosted.

As a specification of the Facebook API, a webhook\(^4\) written in JavaScript was developed to respond to events that occur


\(^4\) In web development, webhooks are intermediary layers between clients and servers that respond to registered events and make the appropriate communication between both sides.
in the Facebook module (i.e., messages sent by users and transmitted via HTTP POST requests). The communication between modules, represented by the arrows in Fig. 1, works as follows. First, a given user sends a message to the chatbot using their Facebook account (arrow 1). Then this message (which should represent a stressful situation experienced by the user) is transferred to an IBM service (IBM Watson Natural Language Understanding) that recognizes human sentiments present in text messages (arrow 2). As a result, this service returns one out of three different values: positive, negative or neutral (arrow 3). If the sentiment is not negative, a warning message is returned to the user (arrow 8) stating that they should use the chatbot only to share undesired situations, which means any other type of conversation is not allowed. If the sentiment is negative, the webhook will continue to make the appropriate calls to analyze the message.

IBM Watson Assistant is a service that allows developers to create dialogue schemes in order to implement chatbots. When creating a project in this platform, one should provide examples of text that is expected as input. Based on these examples, the system can learn to classify new input texts. In our case, examples were provided about stressful situations shared by social media users. Hence, using machine learning, this service is able to decide whether any incoming message represents a user seeking for social support tailored to a particular stressor. Consequently, via text mining, it is possible to determine how to respond to the identified stressor: keywords representing stressors were provided in advance, so the platform could generate very specific answers in accordance with the algorithm explained in Section III-A, because the stressors could be extracted from the messages sent by users. In arrow 4, the webhook sends the message received from a user to IBM Watson Assistant, and in arrow 5 a classification of the message into one of the predefined categories of stressful situations is received. It is also possible that the message is not a description of a stressful situation, even though it contains a negative sentiment. In this case, the webhook should return a warning message to the user (arrow 8). If there is indeed a stressor in the message, the webhook receives one out of four possible support strategies that can be used as a response message (see Table I).

Finally, after confirming that an incoming message contains a description of a stressful situation, the webhook will convert the selected support strategy into a comforting message that should be sent back to such an user. In arrow 6, two pieces of information are requested by the webhook to a cloud MongoDB database service called mLab: 1) whether the user who sent the message is already registered in the system; and 2) a pointer to the last text message that was sent. Based on the pointer returned by the database (arrow 7), the webhook selects the appropriate text message to send to the user (arrow 8). In case the user is a new user, the webhook will perform an extra call to the database (not shown in Fig. 1) to store their information.

IV. METHOD

A. Design

The purpose of the reported experiment was to study the effect of interactions with our emotionally supportive chatbot on the stress level experienced by its users. A second goal is to investigate whether it matters if users are aware that the support is computer-generated. To this end, an experiment was conducted in which participants were invited to share their personal stressful experiences with (different versions of) the chatbot via Facebook. A one-factor between subjects experimental design was adopted, with the description of the conversation partner as independent variable and the experienced stress as dependent variable. This approach is common in social psychology [26], since it allows participants to report their experience in a specific condition while minimizing the potential effect of other factors [27], [28]. Participants were allocated randomly to one of three conditions: Group B (bot), Group H (human), and a Control Group (see Section II). In Groups B and H, participants received support from the proposed chatbot. However, in Group B participants were informed that the support was computer-generated, whereas participants in Group H were told that the support was provided by another, randomly selected participant. An additional difference between the two conditions was that in Group B, the chatbot provided instantaneous responses to user input, whereas in Group H a delay of 15 seconds was incorporated. In the Control Group, participants did not receive
any form of support at all. Hence, they were also asked to type in their stressful experiences via the Facebook interface, but they did not receive any response. Moreover, they were informed that nobody would read their messages. More details about the instructions that were provided to the participants are presented in Section IV-D.

B. Stimulus Material

The participants of the experiment were asked to share their stressful experiences via a dedicated Facebook page. Three different Facebook pages were created, each of them representing one of the three conditions introduced earlier (see Fig. 2). Upon entering one of the pages, participants could type in their stressful experiences and click on the Send Message button in order to “share” them. Note that these messages were actually kept private, which means that, beside the participants themselves, only the respective researchers could read them. The pages were available to the participants at any time while the experiment was being conducted.

Participants allocated to the Control Group were directed to the page entitled Share your Stressful Events. As mentioned above, this page was nonresponsive, i.e., participants could type in whatever they wanted but did not receive any message back. On the other hand, the pages called Stress Supportive Bot (for Group B) and Stress Supportive Friend (for Group H) were connected to the chatbot described in Section III. Therefore, participants in these conditions received answers to the messages they sent. Since both pages were connected to the same piece of software, their behaviors were identical: any given message would result in the same type of answer for both pages. An example of a conversation between a user and the chatbot is shown in Fig. 3. The participants could interact with the pages via any device connected to the Internet as long as it provides access to Facebook Messenger.

C. Participants

This study received ethical approval from the ethics Committee of the Social Sciences faculty at Radboud Universiteit (project number ECSW-2018-10). A total of 210 participants (70 per condition) were recruited via the online crowdsourcing platform Prolific. Only adult English speakers with daily access to the Internet and Facebook Messenger and who never received any diagnosis of mental disorder were eligible to participate. Participants were requested to sign a consent form online. Each participant received £1 per day as a reward for their contribution (the experiment took no more than 10 minutes per day). As mentioned earlier, participants were randomly allocated to one of three groups: Group B, Group H, and a Control Group. After data collection, the data were cleaned according to the following conditions: all participants that contributed for only one day (B: 3, H: 9, Control: 6) or two days (B: 3, H: 3, Control: 3), did not share any valid stressors (i.e., random or empty messages) (B: 32, H: 21, Control: 27) and that did not write their messages in English (Control: 4) were excluded from the analysis. Additionally, 13 individuals from Group H who did not fully believe they received human-generated support were removed. As a result, Groups B, H, and Control ended up with 32, 24, and 30 participants, respectively.

The participants’ age varied from 20 to 64 years. No significant differences were found between the means of the respective groups: Group B ($M = 31; SD = 10.38$), Group H ($M = 35; SD = 13.16$) and the Control Group ($M = 32; SD = 13.01$);
The participants were asked to read a text with instructions and sign a consent form.

2) For three days in a row they were asked to do the following. 
   a) Report their current emotional state by filling in the affect grid (see next section for details).
   b) Open the respective Facebook page and enter at least two messages describing (two different) stressful events they recently experienced.
   c) Fill in the affect grid again.

3) After completing the steps above (i.e., after sending messages for three days in a row), participants were debriefed. In particular, participants in Group H were informed that they had been interacting with a chatbot instead of a real person. Additionally, just before this debriefing, participants in Group H were asked to share their thoughts about the profile of their conversation partner. This was done to verify whether they actually believed that they were interacting with a human being.7

E. Measures

Self-reported stress levels of participants were measured by means of the Affect Grid, a single-item measure introduced by Russell et al. [29]. Using this approach, participants can report on their current affective state by indicating their own levels of arousal and valence in a 9 × 9 grid (see Fig. 4), where the horizontal axis represents valence and the vertical axis represents arousal. Here, valence refers to the level of positivity of the person’s experienced emotional state (ranging from negative to positive), whereas arousal refers to its intensity (ranging from sleepy to aroused). The cells highlighted in red indicate “feelings of stress and tension,” whereas the green cells indicate “feelings

7The following open-ended question was used for this: “You MUST use the space below to describe your thoughts about the profile of the participant who was replying to your messages during the experiment. You might believe, for instance, it was a male/female, a young adult, an elderly, a non English native speaker, etc.”
of calmness and relaxation.” The orange cell in the center of the grid can be interpreted as a “neutral, average feeling.” Hence, for our study, stress is assumed to correspond to high arousal and low (negative) valence, cf. [25]. As a result, we split our hypotheses about the effect on experienced stress into two separate subhypotheses (one for valence and one for arousal).

The affect grid was presented to the participants via Qualtrics. As mentioned above, they were asked to fill it in six times: before and after their interactions with the Facebook page, for three consecutive days. Hence, in total six values for arousal and valence were recorded per participant who completed the whole task.

A randomization check was performed to check whether there were no significant differences in the affective state of the participants at the start of the experiment. To this end, the initial values for arousal and valence (i.e., the values reported in the first measurement on the first day) of the three groups were compared via one-way ANOVA’s. These tests indicated that the initial values for arousal among the conditions were somewhat different, with Group B ($M = 5.44; SD = 1.85$) having higher mean than Group H ($M = 4.87; SD = 2.42$) and the Control Group ($M = 5.60; SD = 2.09$) being in between the other conditions. However, there were no significant differences between these means: $F(2, 83) = 0.85, p = .43$.

For valence, the mean value in Group B ($M = 4.34; SD = 2.04$) was also slightly higher than in Group H ($M = 3.29; SD = 1.49$) and the Control Group ($M = 4.00; SD = 2.16$). But again, these differences were not significant: $F(2, 83) = 2.02, p = .14$. Consequently, we can conclude that the participants were equally distributed over conditions with respect to their initial affective state.

F. Analysis

Hypothesis H1 and H2 were tested via paired t-tests in which the mean level of arousal (or valence) of the participants from Groups B and H at the start of day 1 was compared to the mean level at the end of day 3. Hypothesis H3 and H4 were tested via two one-way ANOVAs. For each participant, the mean “change” in arousal (or valence) was calculated by taking the difference between the reported values before and after the interactions, averaged over the three days. Next, a one-way ANOVA was applied to test whether there was a significant different in these mean changes among the three groups. In addition, Welch’s t-tests were conducted to make pairwise comparisons between Groups H and B, and between B and the Control Group.

V. RESULTS

In this section, the results of the statistical analysis are presented. Hypothesis H1 and H2, which have been formulated to test the effect of interacting with the chatbot on self-reported stress, are addressed in Section V-A. Next, Section V-B addresses H3 and H4, which postulate the effect of participants’ expectations about the nature of the interaction partner (human or bot). The entire analysis was performed via an R script. Both the script and the data can be accessed via a GitHub repository.

A. Effect of the Chatbot (H1 and H2)

The results regarding H1 and H2 are shown in Fig. 5. With respect to H1, contrary to our expectations, the participants’ self-reported arousal after interacting with the chatbot ($M = 5.19, SD = 1.49$) was almost the same as it was before interacting with the chatbot ($M = 5.27, SD = 1.58$). A paired t-test confirmed that this difference in arousal was not significant; $t(55) = 0.48, p = .32$. Hence, H1 should be rejected.

In contrast, with respect to valence (H2), a clear effect can be observed: the participants’ self-reported valence after interacting with the chatbot ($M = 4.55, SD = 1.42$) was higher (i.e., more positive) than it was before interacting with the chatbot ($M = 3.97, SD = 1.44$). Indeed, this difference is significant; $t(55) = −3.24, p < .001$. Hence, H2 can be accepted.

B. Effect of Expectations (H3 and H4)

Next, to see if the way the interaction partner was presented makes a difference, we look more closely at the average change

---

in self-reported stress by comparing the three conditions. These results are depicted in Fig. 6. For arousal (H3), the differences between the three groups turn out to be rather small. For all three conditions, the change in arousal is minimal (Bot: \( M = +0.02, SD = 1.08 \); Human: \( M = -0.21, SD = 1.38 \); Control: \( M = -0.03, SD = 1.33 \)). Indeed, no significant effect was found for arousal: \( F(2, 83) = 0.24, p = .79 \), which means that H3 should be rejected.

For valence (H4), the picture is more interesting. As expected, the are indeed significant differences between the three groups; \( F(2, 83) = 3.78, p = .03 \), with the participants in the Human condition reporting a large increase in valence (\( M = +1.10, SD = 1.63 \)), and the participants in the Bot condition reporting a much smaller increase in valence (\( M = +0.18, SD = 0.85 \)). Finally, the participants in the Control condition also reported an increase in valence (\( M = +0.52, SD = 1.24 \)).

To further analyze these results, additional Welch’s t-tests were performed to make pairwise comparisons between the groups. These tests pointed out that there was indeed a significant difference between Group H and Group B; \( t(49.48) = 2.50, p = .008 \). However, no significant differences were found when comparing Group H to the Control Group, \( t(41.92) = 1.43, p = .08 \), nor when comparing Group B to the Control Group, \( t(51.04) = 1.27, p = .10 \). Consequently, hypothesis H4 can be partially confirmed: the increase in valence for participants who believed to be interacting with a human being was indeed larger than for those who knew they were interacting with a bot. However, the increase in valence for participants who knew they were interacting with a bot was not larger than for the participants in the Control group, who simply shared their stressful experiences without receiving any feedback.

VI. DISCUSSION

A. Conclusion

The present research aimed to investigate the effect of an emotionally supportive chatbot on its users’ self-reported stress. This was done through an experiment in which the changes in participants’ levels of arousal and valence were measured after interacting with the chatbot. The chatbot used supportive messages tailored to different types of daily-life stressors, inspired by the emotion regulation strategies proposed by Gross [30]. A secondary aim was to investigate to what extent the presumed source of the emotional support (a bot or a human) had an effect on the experienced stress. This was done by asking participants to share their stressful experience via a Facebook interface and comparing three conditions: participants in Group B received support from the bot and were fully aware of this. Participants in Group H received support from the bot but were told the support was generated by another human. And participants in the Control Group did not receive any form of support at all.

The results led us to conclude that on average, participants’ self-reported levels of valence increased after interacting with the chatbot. Moreover, as expected, this increase in valence was larger for participants who believed the support came from another human compared to those who knew it came from a computer. However, contrary to our expectations, for the participants in the Control Group, the increase in valence was not significantly different from the two other groups. Moreover, the self-reported levels of arousal did not seem to be affected for any of the groups.

To a certain extent, these findings are consistent with other research from the literature. First of all, the fact that the chatbot was able to increase users’ levels of valence (2) corresponds to existing claims that computer-generated emotional support may be effective in reducing stress (see Section II). Also, the fact that this increase was larger for individuals who believed the support came from another human (2) is in accordance with well established scientific findings: virtual agents are still not so competent in simulating empathy as well as humans [16] and this might affect users’ impressions making them prefer human-generated support. In fact, it is reasonable to expect that most people are still more open to social support provided by humans due to the fact that emotionally aware computer systems are not very common yet [17]. This is also consistent with the finding that social support given via one-click reactions on social media is less effective when it is perceived as an automatic, mindless reaction [31]. It looks like people want to be “taken seriously” in their search for support, and they probably consider human-generated support more serious than automated support.

However, an interesting finding was that the emotional support from the bot was hardly more effective than simply letting people share their problems without providing support, which also seems to have resulted in a (small) reduction in stress level. In hindsight, this may be explained by referring to research by Pennebaker [32], who found that writing about stress could have a therapeutic function. Indeed, it seems that the mere act of writing down their problems already had some positive effect on the participants’ mood. What is nevertheless remarkable is that in our study, this effect was already reached (to some extent) by writing as little as two sentences, whereas in Pennebaker’s paradigm participants were asked to write about their emotions for 15–30 minutes per day.

Another unexpected finding is the fact that, in contrast to valence, none of the interventions seemed to have any effect on self-reported arousal. A posthoc explanation for this might be that participants were so engaged with the experiment that this already increased their arousal, which may have nullified any
potential effect of the intervention. After all, it is well known from the literature that cognitive workload has an effect on (physiological) arousal [33]. The fact that reported arousal levels were relatively high overall supports this explanation.

Finally, it should be noted that the current form of support was still relatively simple (i.e., consisting of one single supportive message). A more extensive form of support (e.g., in the form of a longer and more personalized conversation) might be more effective. Nevertheless, it is important to state that our chatbot at least did not have a negative impact on participants’ stress level (neither in terms of valence nor arousal). This is an indication that the supportive strategies followed by our agent seem to be appropriate, as improper emotional support is likely to have negative effects on one’s mood [34].

B. Strengths

The strengths of the reported research are mainly related to the design of the experiment. By manipulating participants’ expectations on the source of the emotional support, we could investigate the impact of knowing the support is computer-generated versus believing it comes from a human. Perhaps the idea that someone cares about you is sometimes more important than the content of the support. This is an important lesson for future developers of emotionally supportive chatbots, as they need to find ways to make people have a more open mind to computer-generated emotional support. As technology becomes more sophisticated and at the same time people’s attitude toward technology is changing, we expect that there is still an interesting market for future versions of supportive chatbots.

Another strength of the current article is that it described an experiment where participants were asked to report their arousal and valence levels for three consecutive days. To the best of our knowledge, most existing research about the effect of CGES algorithms on users’ well-being focuses on “one-shot” experiments. Moreover, in our study the dependent variables were measured by means of the affect grid, an intuitive and reliable technique for collecting information on people’s mood.

Finally, it is worth mentioning that we instructed participants to come up with stressful events they actually experienced in their daily life. This makes the experiment more ecologically valid compared to studies in which people are asked, for instance, to imagine such problems.

C. Limitations and Future Research

Despite the above, the presented study also suffered from a number of limitations, which should be further addressed in future research.

First of all, the current study focusing on a specific type of stress—namely, psychological stress—that is caused by everyday events. However, stress can also be related to physical events such as pain after a physical injury. Future research is needed to understand to what extent our approach also applies to these other forms of (physical) stress.

A second limitation concerns the setup of our experiment: since we asked participants to report any stressful events that they could think of, we cannot guarantee that they actually experienced these events recently, nor that they felt a lot of stress about them. Indeed, as the reported levels of arousal and valence were relatively close to neutral, one might argue that the stress participants experienced during the experiment was limited. Future research should therefore look for more substantial stress induction mechanisms.

Moreover, one might argue that a scenario in which a user writes one message per stressor is not realistic. When human peers share their problems with each other, their messages might also contain a combination of stressors. Moreover, typically not just one message, but an entire dialogue may play a role when it comes to providing emotional support. This means there should also be room for small talk that is not directly related to a stressful situation. In addition, human peers have a memory and are able to ask stressed friends about problems from the past. And finally, in a human-like conversation, multiple interaction modalities may be present, such as voice messages and pictures. Therefore, we argue that future versions of computer-generated emotional support should include mechanisms to cover these aspects, in order to create a more realistic simulation of a supportive human peer.

Regarding the algorithm, ideally the chatbot should be able to dive deeper into the content of the stressors. In some cases, knowing the “category” of a given stressful situation is not enough to provide optimal support. For instance, knowing whether a user is facing a stressor herself or whether she is stressed because someone else cares about her is important. And to a certain extent, it is possible to realize this via AI techniques. Other potentially relevant features to extract might be the moment a stressor happened, the amount of money someone needs to pay a given bill, when is the deadline that is bothering a stressed worker, and so on. In some cases, it could already be sufficient to simply ask questions to the user to infer these features.

Another potential improvement involves the duration of the study. Although the study went beyond the traditional one-shot experiment, the number of days (3 in our case) could still be increased. However, in this case, an open question is what would be the ideal duration. On the one hand, individuals do not face stressful situations every day that make them so stressed that they need to seek for help. On the other hand, people often talk about their problems for a longer period. Therefore, future research might address questions like “Is there an added value of having multiple chatbot interactions per day?” and “Is there an ideal duration of a dialogue?” Related to this, another limitation to take into account is that having to interact with the chatbot for 3 days in a row already resulted in a considerable number of dropouts.

Additionally, it is worth mentioning that we did not do an in-depth analysis of the interaction between the participants and the chatbots. That is, we did not investigate measures like the number of messages each participant sent to our chatbot or the time they spent on the interaction. Similarly, we have not analyzed the content of the stressors and the respective supportive messages that were exchanged during the study. These metrics might be used in future research (e.g., using a content analysis) to gain more insights into the types of problems people share and the effectiveness of various forms of social support.
Regarding the dependent variables, we used the affect grid to measure self-reported stress. However, as various other techniques for measuring affective states are available in the literature (e.g., skin conductance, facial expression, body language, tone of voice, etc.), future studies can provide more insight by considering these additional modalities.

D. Implications

Overall, the current study has provided some useful pointers for follow-up research on the use of chatbots for providing emotional support via social media. Although many questions are still open, including technical, social, and ethical challenges, the presented results shed some interesting light onto the possibilities and impossibilities of computer-generated emotional support.

From a practical perspective, we conclude that there is a potential market for conversational agents that provide emotional support in everyday stressful situations, but that these applications are most effective when users believe that the supporting messages are produced by human beings. This poses a serious challenge to the developers of such systems, because transparency is often seen as an important feature: many scholars consider it unethical to develop social machines that falsely pretend to be humans [35]. The challenge is therefore to develop these agents in such a way that they stay effective without having to “fool” the user about their nature. A potential solution, which deserves more attention in the future, is to incorporate techniques for building “rapport” with the user [36]. This way, the interaction with the agent could feel natural while users are rationally aware that they are talking to a bot.

REFERENCES


