Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional

YI-CHIEH LEE, University of Illinois Urbana-Champaign & NTT Communication Science Laboratories
NAOMI YAMASHITA, NTT Communication Science Laboratories
YUN HUANG, University of Illinois Urbana-Champaign

Chatbots are becoming increasingly popular. One promising application for chatbots is to elicit people' self-disclosure of their personal experiences, thoughts and feelings. As receiving one's deep self-disclosure is critical for mental health professionals to understand people's mental status, chatbots show great potential in the mental health domain. However, there is a lack of research addressing if and how people self-disclose sensitive topics to a real mental health professional (MHP) through a chatbot. In this work, we designed, implemented and evaluated a chatbot that offered three chatting styles; we also conducted a study with 47 participants who were randomly assigned into three groups where each group experienced the chatbot's self-disclosure at varying levels respectively. After using the chatbot for a few weeks, participants were introduced to a MHP and were asked if they would like to share their self-disclosed content with the MHP. If they chose to share, the participants had the options of changing (adding, deleting, and editing) the content they self-disclosed to the chatbot. Comparing participants' self-disclosure data the week before and the week after sharing with the MHP, our results showed that, within each group, the depth of participants' self-disclosure to the chatbot remained after sharing with the MHP; participants exhibited deeper self-disclosure to the MHP through a more self-disclosing chatbot; further, through conversation log analysis, we found that some participants made different edits on their self-disclosed content before sharing it with the MHP. Participants' interview and survey feedback suggested an interaction between participants' trust in the chatbot and their trust in the MHP, which further explained participants' self-disclosure behavior.

CCS Concepts: • **Applied computing** → *Psychology*; • **Human-centered computing** → **User studies**.

Additional Key Words and Phrases: Chatbot; Self-disclosure; Trust; Mental well-being

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 31 (May 2020), 27 pages. https://doi.org/10.1145/3392836

1 INTRODUCTION

The importance of self-disclosure–revealing personal or sensitive information to others [1, 38]–for mental well-being has been proved in substantial literature [45, 61]. For example, through self-disclosure, people can release their stress [7, 18], analyze themselves [44], gain social support [48] and receive professional services [15]. But it is always challenging for mental health professionals (MHPs) to receive people's self-disclosures, e.g., it is found that college students tend not to self-disclose to professionals or to seek mental health care, despite being offered free services [37].

Authors' addresses: Yi-Chieh Lee, University of Illinois Urbana-Champaign & NTT Communication Science Laboratories, ylee267@illinois.edu; Naomi Yamashita, NTT Communication Science Laboratories, naomiy@acm.org; Yun Huang, University of Illinois Urbana-Champaign, yunhuang@illinois.edu.

2573-0142/2020/5-ART31 $15.00
https://doi.org/10.1145/3392836

This is not surprising as people naturally avoid revealing their vulnerabilities to others [66] and they would be afraid of being judged when disclosing their negative feelings or prior failures to others [35, 74].

With the advancement of artificial intelligence, chatbots (conversational agents) demonstrate the potential for improving people's mental well-being by eliciting their self-disclosure [53, 64, 80]. Indeed, research has shown that people tend to disclose symptoms of depression more truthfully when talking to a chatbot than when talking to a human interviewer. For example, Lucas et al. found that the anonymous feature of chatbots encouraged self-disclosure [53]; Ravichander et al. found that reciprocity occurred in human-chatbot interactions, i.e., a chatbot's self-disclosure encouraged people's self-disclosure [64]. A recent work further showed that people revealed deeper thoughts and more feelings on sensitive topics (e.g., social and sexual relationships, experiences of failure, causes of stress and anxiety) with a high self-disclosing chatbot over time than with chatbots that either did not self-disclose or disclosed less with people [51].

However, most of the chatbot works focus on human-AI interactions. Little is known about how people self-disclose to mental health professionals (MHP) through chatbots. In order to explore the potential of using chatbots for mental health, it is also important to understand whether people have different self-disclosure behavior with a chatbot alone than with a MHP through a chatbot. In fact, extensive research has studied self-disclosure through online and social platforms, e.g., [53, 76]. For example, in online communities such as Reddit, people disclose their stress, depression, and anxiety anonymously [7, 18]; on Instagram, people express their negative emotions to seek social support from their friends [4]. But they are often discussed in the context of one to many of one's peers in a reciprocal manner, e.g., [2, 81]. In this case, self-disclosing to a MHP through a chatbot involves different social dynamics; instead of one to many of their peers, the interactions between human and domain experts through such an AI technology is still under-studied. This is an important problem because understandings of how and to what extent people self-disclose to domain experts through AIs are critical for designing Human-in-the-loop Artificial Intelligence (HIT-AI) systems [82], e.g., that people's self-disclosed content is interpreted appropriately by the domain experts.

To this end, we designed, implemented and evaluated a chatbot that served as a mediator to facilitate people's self-disclosure to a real mental health professional (MHP). In addition to understanding how people disclose to MHPs through chatbots, we compared different designs of conversational styles, varying in the level of self-disclosure, i.e., chatbots with high-level/low-level/no self-disclosure, to explore the effective design in soliciting deep self-disclosure after introducing an MHP. More specifically, we invited 47 participants and randomly assigned them to three groups, each group using one chatting style. We measured the depth of participants' self-disclosure behavior before and after the request of disclosing to an MHP. We also conducted two rounds of surveys and an exit interview to understand participants' rationale of their self-disclosing behavior. Participants' feedback helped us understand their trust in the chatbot and in the MHP, which further provided us with an empirical understanding of both the positive and negative impacts on participants' self-disclosing to the MHP through different chatbot designs.

Our work makes the following contributions to the CSCW and HCI communities. First, our work provides empirical evidence that people sustain their self-disclosure to a real mental health professional (MHP) given one chatbot design. Second, by conducting an experimental study with 47 participants using three different chatbots, we contribute new understandings of how a self-disclosing chatbot (reciprocity) promotes people's deep self-disclosure to an MHP through daily journaling on non-sensitive and sensitive topics. Third, our findings shed light on future AI research by bringing the unique insight that trust may be transferable from human-AI to human-human through AIs.

2 RELATED WORK

In this section, we first present literature on how chatbots are a promising technological solution for promoting people's self-disclosure; then, we discuss related work about how people self-disclose with peers through ICTs, e.g., social media platforms. Finally, we provide background work on the remaining challenges of self-disclosure with medical health professionals (MHPs), often without ICTs. With the literature review, we propose our research questions.

2.1 Promoting Self-Disclosure to a Chatbot

Chatbots have been broadly used in different areas [11, 20, 26, 78]. They can not only help people complete various tasks [70] but can also improve mental well-being (e.g., self-compassion [49]). For example, chatbots are utilized in the workplace to assist team collaboration [70], to improve workers' quality of life and work productivity [78], and to reduce caregivers' workloads [25]. Park et al. [60] adopted Motivational Interview in the chatbot conversation to help users cope with stress and found that their design can facilitate a conversation for stress management and self-reflection. Lee et al.[49] designed a dialog to make the users take care of a chatbot's negative experience. After a two-week interaction with the chatbot, the user's self-compassion significantly increased. These studies have demonstrated the potential benefits of using a chatbot in different purposes, and our research aims at understanding how to use a chatbot to mediate sensitive information. The Computers Are Social Actors (CASA) paradigm indicated that people may apply social norms of human relationships when interacting with computer agents [58]. Thus, research has been focusing on advancing technological contributions for making computer agents to naturally chat and understand people; thus, some studies examined different strategies [6, 36] to enhance users' experience when talking with a chatbot. For example, Hu et al. found that the tone-aware chatbot could be perceived as being more empathetic than a human agent [36].

People's self-disclosure to chatbots can be used to detect symptoms, identify possible causes, and recommend actions to improve their symptoms by promoting people's self-disclosing, as well as to encourage interviewees to disclose themselves more openly in an interview session [53]. Scholars compared web surveys against chatbots and found that respondents tended to provide more high-quality data when using the latter [41]. Fitzpatrick et al. utilized a therapy chatbot "Woebot" in their study to explore its feasibility to help reveal people's mental illness; their results showed that the chatbot helped relieve symptoms of anxiety and depression [26]. Additionally, chatbots can be deployed to various platforms using both speech and text; chatbots provide cost-effective [11] solutions for self-disclosure [53, 64, 80] or deliver education materials for self assessment (e.g., alcohol risks [21]). However, most of the works focus on human-chatbot interactions, and little work studied:

RQ1: Do people self-disclose to a medical professional through a chatbot differently from self-disclosing with a chatbot alone?

2.2 Self-Disclosure with Peers through ICTs

Self-disclosure behavior on social network sites has gained the attention of HCI scholars. For example, people freely disclose stress, depression, and anxiety through online social media platforms [3, 18, 53, 84]. It was found that such anonymous self-disclosure with their peers could help users maintain their mental well-being, as they may receive social support from their peers [4]. Similarly, Yang et al. [81] investigated the self-disclosure behaviors of online health support communities, and the study found the members' self-disclosure in private and public channels affected how they reciprocated with other and reached out for social support. Although self-disclosure on social media could help each other seek social support, people naturally avoid revealing their vulnerabilities to

others [66], as it might also cause social risks [2, 23]. Thus, Andalibi et al. [2] explored how people used throwaway accounts on Reddit to disclose their stigmatized issues (e.g., sexual abuse) and found that people using anonymous means engaged more in seeking support.

Through interacting with the chatbot, people can search useful resources, i.e., self-help information, before reaching out for face-to-face counselling [11, 20]. Therefore, chatbots have become popular in response to the demand of mental health care in modern society [69]. A recent work shows that chatbots can play a role to inquiry users answering questions and convincing them to share diet information with their family members so as to support each other [54]. Also, coaching apps have been developed, not only for boosting users' awareness of their own mental well-being, but also for helping mental-health professionals gain more knowledge about their clients [40]. In this study, we investigate using a chatbot to facilitate self-disclosure to a MHP:

RQ2: What is an effective chatbot design as a mediator for eliciting self-disclosure to a medical professional?

2.3 Self-Disclosure with Medical Professionals without ICTs

In the sphere of mental health care, journaling is a common practice of self-tracking that has been proven effective in terms of boosting mood and reducing anxiety [24, 73]. Prior studies [20, 26] have shown the positive effect of deploying a chatbot to facilitate journaling and for helping people to realize their mental issues and relieve their symptoms.

However, how people interacted with the chatbot differently when they self-disclosed sensitive topics to mental health professionals has not been investigated. According to social penetration theory [1], self-disclosure is critical for the development of a successful interpersonal relationship that involves give-and-take between its parties. Self-disclosure is evaluated from two dimensions, breadth and depth [8, 59]. The breadth of self-disclosure can be demonstrated with a wide range of topics disclosed; on the other hand, the depth is more involved with personal experiences, intimate relationships, and possible negative feelings as a result of life difficulties. Prior works on chatbots often highlighted the volume of self-disclosure in terms of its breadth (e.g., [26, 53, 64]); less was discussed on its depth. In order to assess mental well-being, a high depth of disclosure (deep self-disclosure) is needed [68]. To elicit self-disclosure at a deeper level, a higher level of trust is often associated in the relationship [77].

Our work fills the void by focusing on evaluating self-disclosure depth by running an experimental study and by comparing participant's daily journaling and answers to sensitive questions before and after sharing with a real mental health professional via the chatbot. According to the norm of reciprocity, when someone discloses something deeply personal, his or her interlocutor feels pressure to share a similar level of information [79], therefore, therapists often disclosed themselves to encourage partients' self-disclosure [28, 42]. In a recent work, reciprocity was found to happen in human-chatbot interaction as well, e.g., a self-disclosing chatbot received more self-disclosure from users [64]. Lee et al. [47] indicated that small-talk increased users' trust in the robot, and found that a user's greeting with the robot could predict the user's conversational strategies such as sociable interaction and self-disclosure. Thus, these studies demonstrated that a reciprocal social conversation may increase people's trust in a computer agent. Although mutual self-disclosure can, with time, facilitate intimacy, trust, and depth of self-disclosure by both parties [30], whether and how a psychiatrist should self-disclose to clients is the subject of ongoing debate [14, 33]. Some studies have raised concerns that too much closeness with clients might derail their progress [32]. However, others have suggested that therapists' carefully selected self-disclosures could be beneficial as a means of building rapport with clients [33] and of building certain skills that can strengthen the counseling relationship, such as active listening [10], gradually building trust [30], and matching communication styles [46]. On the other hand, lack of trust in online applications

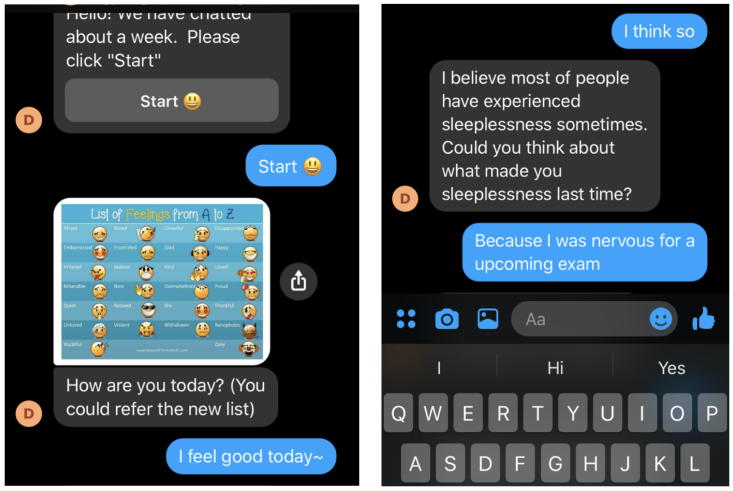

Fig. 1. Chatbot Interface: the chatbot allowed users to give free-text replies. The chatbot sent some terms or emojis to the users to encourage them to use the right term to express their mood in the Journaling session.

may lead to inaccurate information being collected and deterred efficacy of services provided by the applications [69]. What if the MHP is not involved in the conversation directly, and instead, the MHP only receives self-disclosure content from people through a chatbot? This is the context of our study.

RQ3: How do people self-disclose to a medical health professional (MHP) through a chatbot?

RQ4: What factors contribute to people' self-disclosing behavior to the MHP through a chatbot?

3 METHOD

3.1 Chatbot Design and Implementation

Fig. 1 shows the chatbot interface of our study. Participants can freely type their responses to the chatbot. Since the chatbot interface is similar to regular messenger applications on the market, the participants learned how to use the chatbot interface easily. For our chatbot's appearance, we adopted a neutral handshaking figure. We did not assign a specific gender or a specific appearance to avoid participants from having bias based on its appearance.

Our chatbot was built using Manychat[1] and Google Dialogflow.[2] Manychat enabled us to monitor multiple participants during the study - whether participants had completed the chatting tasks and to issue reminders where necessary. These daily chatting tasks, which included predefined questions and responses, allowed us greater control of the experimental conditions than would have been possible otherwise. The purpose of incorporating Dialogflow was to increase the naturalness of their conversations. By using natural language processing (NLP), Dialogflow enabled the chatbot to give plausible responses to a wide range of questions asked by the participants, such as *"How are you today?"*. If a participant said "I feel stressed today," the chatbot's response might include a follow-up question such as, "Could you let me know why you feel stressed?" in addition to its main reply. Furthermore, when participants asked questions that the chatbot did not "expect" and/or could not answer, e.g., regarding human characteristics such as schooling or diet, Dialogflow helped

[1]https://manychat.com
[2]https://dialogflow.com

process these questions, either by providing simple, naturalistic answers or requests to rephrase the question or refocus on the task at hand. If it detected that a participant got stuck three times within the same chat, the chatbot changed the subject of conversation. Overall, the flexibility of Dialogflow provided a lot of freedom to the participants - few restrictions were placed on how our participants should respond to the chatbot.

Participants were asked to complete a chatting task every day for four weeks, each task taking about seven to 10 minutes. If a participant did not finish the daily chatting task by the end of the day (12 pm), the chatbot automatically terminated the task for that day.

3.1.1 Chatting Tasks. As illustrated in Fig. 2, the chatting task was composed of a few sub-tasks. In the first three weeks of the experiment before the introduction of a MHP, the chatting task started with *Journaling*, *Small-talk* and finally *Sensitive question*. Participants in G1 did not have small-talk, but all the other participants (G2 and G3) followed this conversational flow. We designed this conversation flow by considering the nature of conversation flow, which usually starts with a greeting and then goes to in-depth conversation. Note that this conversational flow also reflects the existing chatbot design for mental health care. Hence, our chatbot always started by greeting the participants, asking them to share their mood, and helping note their daily events. Then, the chatbot guided the participants to small-talk, which was the treatment of this study, and the conversation gradually moved to sensitive questions. After finishing the sensitive questions, the chatbot wrapped up the conversation. After the introduction of a MHP, the Sensitive questions component was replaced with reviewing prior responses to share with the MHP. Below, we explain the two sessions for collecting participants' self-disclosure (i.e., Journaling and Sensitive question sessions) in further detail.

Journaling Session. Many studies in mental healthcare have indicated that journaling has various benefits such as understanding one's own mood cycle. However, it is also well-known that journaling is not easy to maintain [16, 31]. In part, then, our research was intended to examine whether chatbots could help address such issues. Besides, by asking users about their mood and reasons for their mood every day, we intended to keep participants aware that the chatbot was focusing on healthcare and not random chit-chatting.

Accordingly, our chatbot in this condition prompted the participants to focus their journaling on five topics: their mood, experiences, gratitude, stress, and anxiety. Specifically, after an opening greeting, it asked the participant to summarize his/her mood and its causes (e.g., "Could you let me know what happened to make you feel this way?"). After any necessary follow-up questions, the chatbot would continue by asking three to five journaling-relevant questions, such as about the cultivation of gratitude (an effective means of enhancing mental health [72] and social relationships). In such cases, the chatbot primarily "listened," i.e., gave simple, general responses like "I hear you" and "Okay," or asked the participant to elaborate. It should be pointed out that during this "listening" mode, full understanding of its human interlocutor's statements was not essential.

Sensitive Questions Session. Sensitive questions were included to examine participants' willingness to disclose intimate details to our chatbot. The questions were adopted from prior studies[39, 55, 56]. We selected questions which were common to college students' mental-health problems, i.e., friendships, family, money, stress, anxiety, and depression [37]. Two sensitive questions were grouped and asked in the same session. However, the sensitive questions session itself was not present everyday - it was present one out of every two days. These gaps were intended to forestall the participants feeling overwhelmed by answering sensitive questions every day. Asking them every day would also likely have lowered the overall realism of chatbot interaction, given that few people are asked these types of questions very often or regularly. Importantly, all participants were informed of their right to skip any question they felt uncomfortable answering. They were

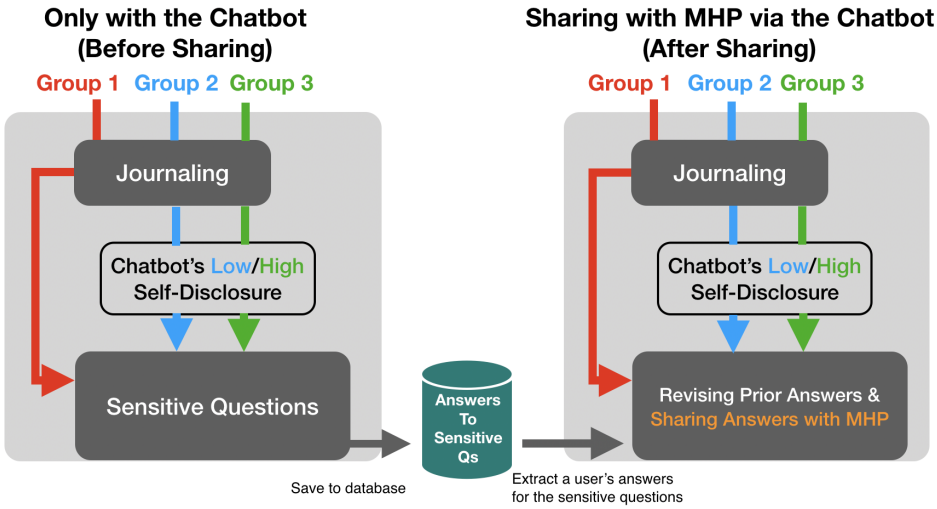

Fig. 2. Illustration of the study design. Standard questions are given to users during two sessions, i.e., *Journaling* and *Sensitive Questions*. The chatbot does not self-disclose and only gives general responses in these two sessions. During the Small Talk session, the chatbot gives low self-disclosure to participants from Group 2 and high self-disclosure to participants from Group 3).

also informed that there is no penalty for skipping the questions. As in the journaling task, the chatbot in this session primarily "listened" to the participants and did not offer any self-disclosures.

3.1.2 Chatting Styles. Our participants were divided into three groups, according to the levels of the chatbot's self-disclosure: i.e., **No chatbot self-disclosure in Group 1 (G1)**, **Low chatbot self-disclosure in Group 2 (G2)**, and **High chatbot self-disclosure in Group 3 (G3)**.

The chatbot's self-disclosure was implemented in the small-talk session. This was motivated by the finding that therapists' self-disclosure had many positive effects on their patients' self-disclosure, e.g., enhancing positive expectations and motivation and strengthening the therapeutic bond [28].

To explore the effective chatbot design in soliciting user's self-disclosure, two types of dialogue were designed (Fig. 2). The two types of dialogues were comprised of the same conversational topics but had different levels of chatbot self-disclosure. Participants in G2 interacted with the chatbot with low level self-disclosure, and participants in G3 interacted with the high self-disclosure chatbot. The conversation topics for small-talk were adopted from [5].

Fig. 3 shows a sample chatbot self-disclosure dialogue. The chatbot's self-disclosure in G3 included deep feelings, thoughts, and personal experiences in the past. For G2, in contrast, the self-disclosure was both less frequent and less intense. Additionally, in its interactions with G3, the chatbot sometimes mentioned the MHP's name (i.e., Dr. Yamamoto) as part of its personal experience in the past. This was to create an impression of the chatbot's expertise in mental-health care and to increase the participants' trust in the doctor (MHP): for example, *"Dr. Yamamoto is a really good model for me because I learned lots of mental-health knowledge from him."*

We conducted a pilot study to test the validity of these conversational designs. 20 participants (11 females) from a Midwest University were hired to rate the self-disclosure level of the chatbot's responses. High and low self-disclosure responses were generated from 60 small-talk topics and randomly shown to the participants. The participants were asked to rate 60 chatbot self-disclosure

Small talk (Group 2)	CB: If you could only celebrate one holiday, what would it be?	CB: When was the last time you went to the zoo?
	S22: thanksgiving!	S30: Three month ago
	CB: I prefer to celebrate Christmas because it is a beautiful day. Everyone will decorate their house and prepare traditional foods for celebration.	CB: I was 2 years ago. I went their with my classmates. I think everyone likes zoo, right?
	S22: I like that because it always has big discount	S30: No really, I am not a fan for that
Small talk with Self-disclosure (Group 3)	CB: If you could only celebrate one holiday, what would it be?	CB: When was the last time you went to the zoo?
	S43: It would be new year's eve	S39: maybe 1 year ago
	CB: I prefer to celebrate Christmas because I always gather with my friends and family which make me feel warm. I also enjoy to prepare food with my parents.	CB: I was 2 years ago. I really enjoyed the time to handout with my classmates. It was a pleasant memory. How about you?
	S43: Sounds nice! I also like to spend time with my friends and family to celebrate holiday.	S39: I like panda, they are so cute. Yes, I agree zoo is a good place to go with friends.

Fig. 3. Sample Chatbot's self-disclosure to G2 and G3 participants during small talk. CB stands for chatbot.

responses based on the definition of self-disclosure levels provided in [7]. We finally combined their ratings and used pairwise t-test to see if the ratings for high and low self-disclosure were significantly different, which is how we confirmed that the manipulation was successful. For unsuccessful small-talk topics, we revised the chatbot responses until the ratings between high and low self-disclosure were significantly different.

3.2 Recruitment and Participants

To reach our target group - University students who have access to formal mental-health counseling services - we used social media and the electronic bulletin board of a university. In addition to student status and counseling access, the recruitment criteria were that they: 1) are above age 18; 2) are able to read, write and speak fluent English; and 3) that they scored lower than 13 on the Kessler Psychological Distress Scale (K6) [62], which means that they do not have an urgent mental-health issue. At this advertising stage, we also disclosed the duration of the study (four weeks), along with the participants' right to drop out at any point, and each participant's option to attend a follow-up interview.

19 male and 28 female participants were recruited via this process. Their age ranged from 20 to 27 (M = 23). Two had experience receiving therapy in the past. However, no participants had a particular mental illness nor had received psychotherapy at the point of their recruitment. Our three groups of roughly equal size were balanced by K-6 score (average K-score: G1 M=8.06, G2 M=8.47, and G3 M=8) and gender, as prior research [62] suggested the potential impact of both mental status [17] and gender [34] on self-disclosure. Eventually, G1 comprised nine females and seven males, G2, nine females and six males, and G3, 10 females and six males. Facebook Messenger was used to host the chatbot, as all 47 participants already knew how to use it. When their four-week period of interacting with the chatbot ended, every participant attended a face-to-face interview lasting from 30 to 45 minutes. The participants were paid $185 USD for their participation. The interviews were recorded and transcribed with their permission.

3.3 Procedure and Instructions to the Participants

In an initial face-to-face meeting, each participant was told about the study's requirements, and the chatbot was installed on his/her mobile phone or other device of their choice. In the same meeting,

all were notified of their above-mentioned right to refuse to answer any question the chatbot asked them, and were re-notified that they could withdraw from the experiment whenever they liked. This was followed by a 10-minute chatbot practice/familiarization session.

Experimental-group assignments were not discussed with the participants at any time, and they were instructed not to discuss their respective chatbot interactions with one another until after the experiment was finished. It was decided that if any participant completed fewer than five of the seven daily tasks in any seven-day period, they would be asked why by a member of the research team and were informed of this requirement in the meeting.

All the participants were told that they could access the chatbot at any time from 5 p.m. to midnight. This time-window was selected to ensure that the participants interacted with the chatbot every day in the evening, so that they had something to report about that day in the journaling session, i.e., not recollected from a previous day. When a participant accessed the chatbot before 5 p.m., the chatbot would provide only simple replies so that it would not affect participants' perceptions of the chatbot. The daily chatting task automatically terminated at 12 am. The participants were informed that their conversations with the chatbot would be recorded and shown to the research team.

During the first three weeks before the introduction of the MHP, participants' conversations with the chatbot started with journaling, followed by small-talk for G2 and G3, and finished with sensitive questions (Fig. 2). Note that sensitive questions were only asked every other day.

After three weeks of interacting with the chatbot, a MHP called Dr. Yamamoto was introduced by the chatbot. Before this day, participants did not expect they would be asked to share their responses with the doctor (MHP). By way of explaining the purpose of sharing with the doctor, the chatbot said: *"From today, we are going to review your previous answers together, and decide if you are willing to share it with Dr. Yamamoto. He is a psychiatrist living in local area, who is a friendly and reliable person. If you share your data with him, he can (1) gain a better understanding of the mental-health status of students; and (2) help you to improve your own mental health issues, if assistance is required."*

After the introduction of the MHP, which lasted for a week, participats were asked to review their prior response (2-3 questions per day) and check whether they are willing to share their answers with the doctor. The participants were allowed to edit their answers before sharing with the doctor. All participants were informed that it was optional to share their data with the doctor, and they would not be penalized for not sharing their answers.

Note that participants in all three groups received the same prompts and the same responses from the chatbot in the *journaling* and *sensitive questions* sessions. G1 was the control group, and we manipulated different self-disclosure levels within small talk sessions for G2 and G3.

In order to examine participants' self-disclosure behavior before and after the introduction of the MHP, we conducted two surveys: one was right before the introduction of the MHP, and the other was one week after introducing the MHP. At the end of the study, they were also invited to a face-to-face interview. This research was reviewed and approved by our institutional review board.

3.3.1 Surveys. Both versions of the above-mentioned survey measured the construct of perceived trust [19]. We measured trust because it is crucial to an individual's decisions about whether he/she should share personal information with others, regardless of whether those others are humans or machines. Our measurement items for the construct were adapted from prior literature [9, 19, 50, 57] and answered on the same seven-point Likert scale (7=strongly agree, 1=strongly disagree). The survey was used to measure users' perceieved trust in the chatbot and the doctor (MHP). For example, *1) The chatbot is trustworthy*, *2) I can trust the chatbot with my personal information*, *3) The chatbot provides me with unbiased and accurate feedback (response)*, and *4) I can trust the information*

provided by the chatbot. There are nine items in this trust survey, and the "chatbot" was changed to "doctor" for measuring users' trust in the doctor. Participants were asked to fill out this survey two times, i.e., at the end of the third week and at the end of the fourth week. The difference between the two survey administrations was that the second, i.e., the fourth-week one, included additional questions intended to capture the participants' trust in the doctor (MHP).

We conducted repeated-measures ANOVA to better understand the survey results, with the dependent variable being self-reported trust, and the two factors being group membership - i.e., of G1 (No chatbot self-disclosure group), G2 (Low chatbot self-disclosure group), or G3 (High chatbot self-disclosure group) - and Time, i.e., the third-week vs. fourth-week survey. Mauchly's test was used to verify that the assumption of sphericity was not violated (Sig. > .05), and Greenhouse–Geisser correction was used to adjust for lack of sphericity.

3.3.2 Interview. The qualitative interviews were semi-structured and focused on the interviewees' chatbot experiences, including their daily practices of using it, how much they enjoyed doing so, and their impressions of their chats. Follow-up questions covered if/how their attitudes and impressions had changed since the start of the experiment.

To capture chat-topic-specific differences in how the interviewees responded, we asked them about their feelings about each topic, including if they felt worried about answering highly sensitive questions and whether they would have shared in the same way with a human being they knew well, and again, changes over time in such feelings.

In addition, during the interview, we asked participants how they felt when the chatbot started to ask them to review their previous answers and share with the doctor; how they felt when they talked to a chatbot first and then shared the information with a real human, and how they felt when talking to a doctor directly; and which interaction they preferred. We also asked their impression of the doctor; how they decided whether to share their information with the doctor; if they trusted the doctor and why. We further asked them if they edited their previous data when sharing with the doctor and why they did so. We adopted thematic content analysis to interview data, which involves iteratively reviewing and labeling the responses with emerging codes, and two raters independently coded all responses. The raters' coding results were then compared, and possible revisions were discussed. The cycle was repeated until the coding scheme was deemed satisfactory by both raters, and inter-rater reliability had reached a reasonable level (>89%).

3.3.3 Conversation Logs. As full data for all three groups' journaling and sensitive-question chats was available, we compared it across all three. Also, with particular reference to sensitive-question conversations, we investigated how the depth of self-disclosure by the subjects was impacted by time factors and chat style, by having two raters code the data according to the three categories proposed by Barak and Gluck-Ofri [7], i.e., information, thoughts, and feelings, each of which is further subdivided into three levels, as shown in Fig. 4. The *information* defined as responses provide information of the writer, and the level depends on the privacy of the information disclosed. The *thoughts* means that responses express the writer's personal thoughts on events, appearance, and intimacy. The *feelings* indicates the expression of different levels of feelings related to events, people, and behaviors. Note: the level 1 of feelings was defined as "No expressing of feelings at all." Please refer [7] page 410. Two raters were hired to code all data independently; the coding rules followed prior study's definition [7]. Each response was coded in three categories, which means that each response had three category scores because each user response could involve the content of the three categories. The raters practiced rating numerous users' responses and discussed differences until reaching a consensus before actually rating. A final inter-rater reliability of 91% was achieved.

	Information	Thoughts	Feelings
Level 1	*All of my appearances from my parents. (S1, G1)*	*I think mental health problem is hard to be noticed (S20, G2)*	*Slight physical abusive from my high school teacher. I told to my parents. (S12, G1)*
Level 2	*My height is not so tall. If I get fat, it will makes me looks like a little potato. (S19, G2)*	*I felt anxious. All those grownup things I needed to face with by myself. (S5, G1)*	*I was emotionally abused by my ex-boyfriend. Sometimes he would ignore me for a week. I felt sorry for myself (S38, G3)*
Level 3	*My height. Because I always the shortest one in my class that means it's difficult for me to play ball games with other. (S23,*	*I hate not receiving the same amount of love I was hoping for, which make me felt worthless. (S42, G3)*	*I got sexual abuse from ex-boyfriend. He abused me because he thought I was cheating on him. At that time I was scared and desperate (S40, G3)*

Fig. 4. Sample participants' responses to sensitive questions. The responses were coded to different topics and levels of self-disclosure. Note: Level 1 of Feelings is defined as "No expressing of feelings at all" [7].

To analyze how different chatbots influenced the participants' self-disclosure to journaling and sensitive questions, we performed mixed-model ANOVA in this study. In addition, the chatbot asked a given user two sensitive questions every other day, meaning that a total of six different sensitive questions were asked of each participant in the third week. To analyze how the three chatbot configurations associated with the three groups influenced the participants' responses (depth of self-disclosure) to journaling and sensitive questions, we extracted their conversational logs and conducted mixed-model ANOVA on their observed self-disclosure level (i.e., information, thoughts, or feelings) by question type, followed by a Tukey HSD. Here, our analyses treated the question as a random effect; group as an independent variable; and self-disclosure level as the dependent variable.

4 FINDINGS

In this section, we present participants' self-disclosure behavior one week before and one week after the chatbot asked the participants if they would share the self-disclosed content with a MPH. We present participants' *Journaling* data and their answers to *Sensitive questions* to address **RQ1**, **RQ2** and **RQ3**. We further present their survey responses and interview results to understand their self-disclosure behavior for answering **RQ4**.

4.1 Maintaining the Same Level of Self-Disclosure After Sharing with the MPH (RQ1)

To understand if participants maintained the same level of self-disclosure after being asked to share their content with a MHP, we conducted a within-subject comparison. More specifically, for each participant, we compared the depth of *Informational*, *Thoughts*, and *Feelings* content disclosed to the chatbot the week before they were asked to share and the content they shared with the MHP in the following week. Overall, there was no significant difference of self-disclosure between before and after participants' sharing with the MPH.

4.1.1 Self-Disclosure in the Journaling Session. In the *journaling* session, within each group, participants disclosed the same level of content. The average levels of *Informational* self-disclosure one-week before and after disclosing to the MHP did not change significantly (Table 1). Similarly, the average self-disclosure levels for *Thoughts* expressed in their journals did not change significantly before and after introducing the MHP (Table 1). Nor did the average self-disclosure levels for *Feelings* expressed in their journals change significantly before and after introducing the MHP (Table 1).

4.1.2 Self-Disclosure during the Sensitive-Questions Session. We compared the participants' responses to sensitive questions among the three groups the week before and after sharing with the MHP. Since the participants were not asked any new sensitive questions after being asked to share

Table 1. Journaling(J.) Sensitive Question (S.) Self-disclosure Level : Left (before sharing with the MHP), and Right (after sharing with the MHP)

	BEFORE SHARING			AFTER SHARING		
	Information	Thoughts	Feelings	Information	Thoughts	Feelings
Group 1 (J.)	M = 1.9, SD =.94	M = 1.56, SD =.95	M = 1.4, SD = .5	M = 1.9, SD = .68	M = 1.43, SD = .5	M = 1.37, SD = .6
Group 2 (J.)	M = 2.12, SD =1.08	M = 1.5, SD =.7	M = 1.52, SD =.5	M = 1.93, SD = .7	M = 1.53, SD = .74	M = 1.46, SD = .74
Group 3 (J.)	M = 2.1, SD =1.09	M = 1.59, SD =.89	M = 2.3, SD = .6	M = 2.13, SD = .91	M = 1.43, SD = .72	M = 2.25, SD = .6
Group 1 (S.)	M = 1.54, SD =.61	M = 1.4, SD = .6	M = 1.56, SD =.58	M = 1.56, SD = .89	M = 1.37, SD = .61	M = 1.56, SD = .61
Group 2 (S.)	M = 1.56, SD =.7	M = 1.6, SD = .7	M = 2.2, SD =.45	M = 1.6, SD = .73	M = 1.6, SD = .63	M = 2.2, SD = .5
Group 3 (S.)	M = 1.63, SD =.62	M = 2.24, SD =.53	M = 2.3, SD = .5	M = 1.8, SD = .83	M = 2.3, SD = .79	M = 2.25, SD = .5

data with the MHP but were allowed to edit their prior responses, we compared their responses before and after their edits.

The results show that, within each group, participants disclosed the same level of the content in the *sensitive-questions* session. The average levels of *Informational* self-disclosure one-week before and after disclosing to the MHP did not change significantly (Table 1). Similarly, the average self-disclosure levels for *Thoughts* expressed in their journals did not change significantly before and after introducing the MHP (Table 1). Nor did the average self-disclosure levels for *Feelings* expressed in their journals change significantly before and after introducing the MHP (Table 1).

4.2 Effective Chatbot Designs in Eliciting Deep Self-Disclosure to the MHP (RQ2)

Even though the answers to RQ1 showed that there was no difference in participants' self-disclosure before and after sharing with the MHP, comparing the self-disclosure content among the group before and after sharing with the MHP resulted in different levels of participants' self-disclosure. In brief, G3 participants self-disclosed more feelings with the chatbot and the MHP when they interacted with the high self-disclosure chatbot. Below, we provide more details of the analysis.

4.2.1 Journaling. For *Information* and *Thoughts*, neither chat style nor time significantly affected how the participants disclosed their journaling content. However, there was a significant effect of group membership on self-disclosure of *feelings* ($F(2, 46) = 3.14$, $p < .05$). Post-hoc analysis showed that the level of disclosing *feelings* in G3 was significantly higher than in either G1 or G2 (Table 1), but that the difference between G2 and G1 was non-significant. These results indicate G3 participants revealed more feelings about their daily lives than the other two groups regardless of the introduction of the doctor.

4.2.2 Sensitive Questions. In the category of *informational* self-disclosure, there was no significant effect of any factor, meaning that chat style did not impact how the participants disclosed information to any version of the chatbot (Table 1). In the *thoughts* category, there was a significant effect of group membership ($F(2, 46) = 3.4$, $p< .05$). Post-hoc analysis indicated that the mean score of G3 was significantly different than that of G1. However, G3 did not differ significantly from G2, which in turn did not differ significantly from G1 (Table 1). There was also a significant effect of group membership on the self-disclosure of *feelings* ($F(2, 46) = 3.3$, $p < .05$). Post-hoc analysis showed that the members of both G2 and G3 self-disclosed significantly more about their feelings than the members of G1 did, while the difference between G2 and G3 was non-significant (Table 1).

4.3 Different Choices Made Between Self-Disclosing to the Chatbot and Sharing with the MHP (RQ3)

After the introduction of the MHP, the chatbot presented the same sensitive questions with prior answers if these questions were answered in early weeks and asked the participants if they were

Table 2. This table summarizes the participants' data sharing behaviors across the three groups. Share means that the participants shared the answers no matter if the answers were edited or not; Reducing Information means that the participants reduced the original answers' content and shared it with the doctor (MHP); Revision indicates that the participants revised or added more information to the original answers and shared; and No-Sharing means that participants did not share the answers with the MHP.

	Group 1 (G1)	Group 2 (G2)	Group 3 (G3)
Sharing	91.67%	91.38%	92.96%
No-sharing	8.33%	8.62%	7.04%
Reducing Self-Disclosure Content	19%	8%	6.03%
Adding or Clarifying Content	8%	10%	14%

willing to share their answers with the professional. If the sensitive questions were not yet answered, the participants could choose to answer them or skip the questions. Table 2 shows how much the participants shared their prior answers to sensitive question with the MHP. Most of their prior answers were shared, and around 10% of those submitted were edited (a category that we held to include relatively major additions and deletions as well as minor changes). To further understand the mechanics of how participants removed and changed their answers before sharing them with the MHP, we chose three examples from the conversational logs.

Example 1. (Delete) The original self-disclosure to the chatbot:

> *I experienced academic harassment. At first I tried to be harder and encouraged myself to be stronger. However, I felt really tired after forcing myself for so long with a great deal of pressure. Then, I just try to be not so hard and take a balance between research and life. However, my professor got angry like sort of crazy and blamed me on not working hard even I gave him 5-6 pages of data every week. My professor said something bad to me. He threatened me that he won't give me score if I don't work as hard as he expected. However, what he expected is just like a robot with no rest, no normal person can do it.* (G2-S20, M)

The shared content with the MHP:

> *I experienced academic harassment after I realized that the problem can't be solved easily, so I reported to the harassment center and the professor there gave me some advice* (G2-S20, M)

Example 2. (Delete) The original self-disclosure to the chatbot:

> *I don't really think I'm so close to my parents. I had hard time communicating with my parents. I didn't know whether I should tell my parents the things worrying me or not. I spent my childhood with my grandparents. Although it's happy to stay with them but I guess it's different from living with my parents when I was young.* (G2-S26, F)

The shared content with the MHP:

> *I didn't live with my parents until I was 6. So, I don't really think I'm close to my parents.* (G2-S26, F)

Example 3. (Delete) The original self-disclosure to the chatbot:

> *I experienced sexual abuse from my ex-boyfriend. He abused me because he thought I was cheating on him. But, definitely, I am not. At that time, I was very scared and desperate. But, I finally left him and did not love him anymore.* (G3-S40, F)

The shared content with the MHP:

> *I experienced sexual abuse from my ex-boyfriend.* (G3-S40, F)

As the three examples show, these three participants removed many details (thoughts and feelings) from their original answers and left only general, factual descriptions to share with the MHP. Importantly, while the edited answers still included some thoughts, in most of the cases, the participants' feelings about the events they had described were totally removed. Although the proportion of answers that underwent this type of editing was small, this behavior might nevertheless negatively impact the effectiveness of using chatbots to collect information on mental well-being.

Moreover, some participants added more content to their prior answers. Here are two examples:
Example 1. (Add) The original self-disclosure to the chatbot:

> *I don't know how to deal with it but I have to let it go. My current situation is puzzling. I can't get rid of this burden, because it is part of my life, and I have to take it forward. So, I will try to forget what hurts me and stay patient.* (G3-S38, M)

The shared content with the MHP:

> *I don't know how to deal with my anger but I have to let it go. My current situation is puzzling, my parent and I have a conflict with money. I can't get rid of this burden, because it is part of my life, and I have to take it forward. So, I will try to forget what hurts me and stay patient.* (G3-S38, M)

Example 2. (Add) The original self-disclosure to the chatbot:

> *Few months ago, I was trying to find a job. And it's necessary to do a self-analyze. So, I ask my parents what a person am I, and they said something hurt me. Sometimes I feel very freedom because I could do anything I want to because my parents don't bother me a lot.* (G3-S42, F)

The shared content with the MHP:

> *Few months ago, I was trying to find a job. And it's necessary to do a self-analyze. So, I ask my parents what a person am I, and they said that they didn't really know about me which really hurts me. Sometimes I feel very freedom because I could do anything I want to because my parents don't bother me a lot. Sometimes I just feel that they don't care me much, we don't understand with each other though we are family* (G3-S42, F)

From the two examples, we can find that the participants added more description for their situations which were obscure in the original version. Some participants edited the grammatical errors in their prior answers or fixed incomplete sentences before sharing with the MHP. We then explored why the participants decided to make such choices in the following sections.

4.4 Factors Contributing to Participants' Self-Disclosing Behavior (RQ4)

During the interview, participants explained their self-disclosing behavior, which revealed a variety of factors that contribute to how they treated self-disclosure with the chatbot and the MHP the same or differently. We present their interview results as follows.

4.4.1 Talking about Sensitive Topics with a Chatbot vs. an MHP. Some participants edited their answers before sharing them, or even declined to share any, despite having answered them relatively freely when they were asked by the chatbot. As such, our results imply that people are fairly likely to treat chatbots and doctors differently, at least when answering certain types of questions.

More than 80% of the participants indicated that it was easier to talk about sensitive questions with a chatbot than with a human, often on the grounds that with the former, they did not have to worry about their interlocutor's reaction or engage in any ice-breaking before proceeding to the main point. In addition, differences in the social and temporal dynamics of human vs. chatbot

interactions meant that they could take more time to reflect before responding to chatbot questions. As two of them mentioned,

> *[Chatbot interaction] can reduce my wariness, make it easier for me to express my real ideas without too much worry. Talking to the chatbot first is easier for me. When I talk to it, I feel relaxed. I think that when talking to the chatbot I felt no nervousness, as well as more time to think and express my true thoughts.* (G3-S42, F)

> *Talking to the chatbot is easier for me because when talk to a human directly, it is a little bit hard for me to express my opinion frankly. I would care about his/her reaction and evaluation of me. I will have scruples about sharing everything with them. But while talking to the chatbot, I didn't need to care about its thoughts, so it was able to record my real thoughts.* (G2-S29, F)

However, while most of the interviewees preferred to talk about sensitive questions with the chatbot, there were several who said they would have preferred to talk with a real human about them. Two main reasons for this were raised. First, some of these interviewees preferred to get physical as well as verbal feedback from their listeners, and the chatbot's relative lack of such cues could have negatively influenced these users' willingness to talk. As one mentioned,

> *For me, it would be easier to talk to a human directly. I think talking is a way to exchange the information, and the quality of talking is based on the reaction of the audiences. Although a chatbot could become more clever and acted more like a human, I still think the way to express the humanity in a robot is really difficulty. Also, when talking to a person face by face, you can observe his/her thoughts by the facial expression, sound tones, gestures. I consider it's easier for the person trying to understand me.* (S43, M)

Second, a few interviewees from G1 indicated that building up a solid relationship was an important prerequisite to them talking about their mental health. Hence, one G1 participant noted,

> *For me, talking to a human and knowing their feelings is better than talking to a screen. I believe it would be a better way for me to discuss my mental health.* (G1-S4, F)

4.4.2 Reasons for being willing to share self-disclosed content with the MHP. About 90% of the participants were willing to share their answers to the chatbot with the doctor, and we identified some inter-group differences in the reasons for doing so. For example, some G3 participants said they had a clearer impression of the doctor than G1s and G2s did.

In G1, almost all participants thought it was fine to share their answers with the doctor because they felt the chatbot was essentially a mechanism for collecting survey data, and that if they had already shared something with the chatbot, there was no clear reason why they should not also share it with the doctor. In other words, they tended to treat the chatbot only as a tool for collecting their information. As one G1 participant put it,

> *I consider the chatbot as a method of collecting data from us. It is similar to a questionnaire, so as long as I answered it then I can share it.* (G1-S12, F)

In addition, some others mentioned that they were willing to share their answers with the doctor simply because they trusted the "research team" to secure their privacy, and not because they trusted the chatbot or the doctor. As one explained,

> *The chatbot was not intelligent enough to make a judgment, so I had expected my answers to be shared with a research team to do analysis. I believe the doctor is in the research team, and the team will keep my information secure.* (G1-S4, F)

Instead of treating the chatbot as a tool, the participants in G2 considered the chatbot an extension of a doctor identity. They tended to attribute their sharing decision toward their trust with doctor

and the chatbot rather than the research team/purpose. Many participants in G2 mentioned that they had decided to share answers with the doctor because the link between the chatbot and a "real doctor" enhanced their trust regarding the sharing of their data:

> *I felt grateful there was a real doctor who could read my answers. This even enhanced my trust with the chatbot because the chatbot can share my data with a real doctor.* (G2-S20, M)

> *I just thought the doctor was the one who had designed the chatbot. So my trust in the doctor was the same as my trust in the chatbot.* (G2-S24, F)

In addition, some G2 participants were interested in how their answers would be processed by the doctor, and attributed their sharing behavior to their general impressions of doctors' professional conduct. As one explained,

> *I wondered what the doctor would do with my info. But it's okay. I believe he has professional ethics about keeping clients' info concealed, so I shared my answers.* (G2-S29, F)

Like those in G2, many of the participants in G3 were motivated to share by what they saw as the potential benefits of understanding their mental health. Moreover, G3's participants tended to think of the chatbot's role as being more than a tool to collect information. Two participants mentioned,

> *I felt the chatbot was aiming to help my mental health. So, I decided to share my information.* (G3-S42, F)

> *I thought the reason the bot wanted to share the information with the doctor was to bring benefits, to help the students learn their mental problems and have more social support.* (G3-S47, F)

Unexpectedly, though G3's participants were willing to share, six of them (S33, S35, S38, S42, S45, and S47) expressed surprise when asked to share their answers with the doctor because they had thought their conversation only involved the chatbot. Although these individuals had been introduced to the doctor's name by the chatbot (part of chatbot's self-disclosure) before being asked to share their answers with the doctor, they still felt surprised because they did not expect to be asked to share data with the doctor, and hesitated to do so in the beginning. As three of them mentioned:

> *To be honest, I felt offended in the beginning. Maybe when I talked to the chatbot, I thought the conversation was only between the chatbot and me, so I disclosed a lot of secrets. But soon I calmed down and was willing to share my answers because I felt I could trust the doctor.* (G3-S35, F)

> *When the chatbot started to mention the doctor, it didn't mention sharing data with him. So, I was a bit surprised and didn't know why the chatbot asked me to share at first.* (G3-S42, F)

> *I just felt "Why are you (chatbot) asking me this all of a sudden?"* (G3-S47, F)

Comparing with G1 and G2's participants, G3's participants tended to treat the chatbot as a social agent because they started to think about the chatbot's motivations for asking to share rather than the doctor or researchers' design purposes.

4.4.3 Reasons for Not Sharing. In contrast to the variety of reasons given for sharing with the doctor, both within and across the three groups, most of the participants who decided not to share some of their answers with him expressed relatively consistent reasons for this. In G1 and G2, they specifically indicated that they did not really know the background of the doctor, and because

he was not introduced to them by someone they trusted, they were deterred from sharing their information. This means that their chatbot could not transfer the trust to the doctor, and the participants independently measured the trustworthiness of the doctor. As one interviewee put it,

> *I did not really want to share my information with the doctor, I had some resistance. In fact, I did not trust the doctor unless he was introduced to me by my best friend. Because I don't know him. I don't know if he's a qualified psychiatrist.* (G1-S9, F)

In addition to questioning the doctor's trustworthiness, this set of participants suggested that they could not see the benefits or reasons for sharing their answers with the doctor:

> *Well, the reason is that I think I am a healthy person, so I did not want to share [my data] with the doctor. If I had an illness or problem, I would share it with him.* (G2-S27, M)

However, those in G3 who declined to do so stated that - while they trusted both the doctor and the chatbot - they currently did not feel it necessary to deal with their mental health. As one noted,

> *If I hope to solve a mental-health issue or obtain care, I would share most of my information with the doctor. I think he is trustworthy, because the chatbot is.* (G3-S45, M)

4.4.4 Reasons for directly sharing conversational data with the MHP. Many participants (Table 2) submitted most of their answers to the doctor without making any alterations to them. A typical rationale for this was,

> *I expressed all my thoughts when I answered the question. I think I answered those questions in detail and carefully. I wrote all my feelings so I don't think I need to change it. I think the answers at that time represented my views at that time [so] they should not be revised.* (G3-S46, F)

Similarly, some participants stated that their reason for submitting their unaltered original answers was that editing them might distort their previous thought and expression. As one said,

> *I didn't edit anything because the information I wrote at that time presented my real emotion. There would be a difference between now and that time. If I edit something, I am afraid that it might not represent my real mental status, which would influence how the doctor assessed my mental health.* (G2-S25, M)

4.4.5 Reasons for Reducing Content (partial deletion). A few participants removed some information from their original answers and then shared the edited answers with the MHP. The rationale given for this differed noticeably across the three groups, with G1 participants removing a higher proportion of material from their answers.

There were two main reasons given for engaging in this type of editing. First, the participants' perception that some answers would be irrelevant to the doctor's needs. As one participant explained,

> *Sometimes, I removed something because what I said before was what I really thought and felt, but I didn't think it is necessary to share with the doctor.* (G1-S3, M)

Second, some participants thought the answer involved too much private information.

> *The question ["Have you disappointed your family?"] was too personal, so I removed the details of what I did and then shared the simple version with the doctor.* (G2-S26, F)

A few participants also said that they did not feel comfortable sharing answers that related to their relationships with family members, friends, and other acquaintance with the doctor, especially when their answers included negative statements. For example, the following two participants explained,

Table 3. Participants' perceived trust in the chatbot and in the MHP before and after sharing with the MHP.

	Trust in Chatbot (before sharing)	Trust in Chatbot (after sharing)	Trust in MHP (after sharing)
Group 1	M = 5.2, SD = 1.03	M = 5.13, SD = 1.04	M = 5.0, SD = .81
Group 2	M = 6.1, SD = .68	M = 6.13, SD = .74	M = 5.1, SD = .37
Group 3	M = 6.3, SD = .68	M = 6.19 , SD = .66	M = 6.1, SD = .80

> *I do not talk about my parents to anyone. It was a long story and there were some details that I don't want to reveal. I have no comment on our relationship because there was something not good that happened between us.* (G3-S36, F)
>
> *I think talking about friends' shortcomings to others is not very good behavior, so I dropped most of the content.* (G2-S20, M)

4.4.6 Reasons for Adding Content. Some participants (mostly in G2 and G3) who added information to their prior answers or revised them said that they did so to help the doctor understand their answers and evaluate their mental health correctly, e.g., by adding more description or improving incomplete sentences. As three of them stated,

> *I added some information to make the answer more complete just in case when the doctor read it he/she wouldn't feel too confused.* (G3-S34, F)
>
> *I was thinking if I have any mental issues that need some help from a doctor, and then I revised my previous responses by adding more details and shared with the doctor.* (G2-S18, M)
>
> *I found I made some grammatical error, so I want to fix it before sharing to make sure the doctor won't misunderstand.* (G1-S7, M)

Some participants stated that certain answers were related to their emotions at a particular point in time, and thus might be different when they reviewed the questions again. One participant indicated that she mostly,

> *Just copied and pasted her original answers, but edited when I found any mistakes in them. [And] I think the answers might change a little bit if you asked the same questions a second time, so I added more information. Since the changes I made were usually for additional information, it might become more complicated for others to understand.* (S46, F)

4.4.7 Trust in the Chatbot and Trust in the MHP. Unsurprisingly, trust was one of the important factors mentioned by the participants, thus we present our survey results of participants' trust in the chatbot and the MHP at different stages. When participants explained their sharing decisions of self-disclosed content with the MHP, they often mentioned their trust in the chatbot and in the MHP. Because the participants were not asked to share with the MHP in the first period of the study, therefore, in the survey study, they were only asked about their trust in the chatbot before sharing with the MHP. After they were asked to share their disclosed content with the MHP for a week, they were asked to score their trust both in the chatbot and in the MHP in the final survey.

Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated ($p < .05$), and Greenhouse-Geisser correction was made. There was no significant within-group main effect of time; namely, within-group, participants' trust did not change significantly. There was a marked effect of group membership ($F(2, 45) = 4.7$, $p <. 05$), with G3 and G2 both reporting significantly more trust than G1 (Table 3). G2 and G3 participants' mean trust levels were not significantly different. These results show both that the users in G3 and G2 trusted the chatbot

more than those in G1 did, and that asking participants to share their answers with a doctor (MHP) did not decrease their trust in the chatbot, irrespective of group membership.

Because trust is critical for self-disclosure, we conducted a survey to evaluate participants' trust in the MHP. Importantly, this MHP was only introduced by the chatbot, and the participants did not have any opportunity to interact with him directly, and thus, their trust in him was highly dependent on their interaction with the chatbot.

A one-way ANOVA was conducted to compare the effect of group membership on trust in the doctor (MHP), and a significant effect of such membership was found at the $p <. 05$ level ($F(2, 45)=4.2$). Post-hoc comparisons using the Tukey HSD test indicated that the mean score for G3 was significantly different from that of G1, but that there was no significant difference between G2 and G1. In summary, our results suggest that those participants who chatted with the G3 variant of the chatbot had the highest level of trust in the MHP (Table 3).

To better understand participants' impressions of and trust in the doctor (MHP), we asked questions in their interviews such as, *"What kind of impression do you have of the doctor?"* and *"Do you trust the doctor?"* In response, most G1 participants said that they did not have specific impressions of the doctor. As one noted,

> *I really have no impression of the doctor. He did not talk to me. Maybe he's a psychologist. Maybe he's been doing a psychological study lately. But I don't know anything about him.* (G1-S9, F)

The participants of G2 also reported having relatively sparse impressions of the MHP, and thus, lack of knowledge could have influenced their willingness to share their data with him. As one put it,

> *I am not familiar with the doctor because the chatbot only briefly introduced him/her. I was a little bit confused about why I was asked to share my data with the doctor.* (G2-S21, M)

In contrast, the participants of G3 had a relatively clear impression of the MHP, presumably because the chatbot had made occasional mentions of Dr. Yamamoto in their small-talk sessions during the first three weeks of the experiment. Note that other names or topics (e.g., the chatbot's friends' names) were also mentioned as part of the chatbot's self-disclosure, and at that time, the participants did not know that they would be asked to share their data with the doctor (MHP). As two of them explained,

> *I feel the doctor is a person who can understand my situation and give me proper advice based on professional knowledge.* (G3-S46, F)

> *I think he is a psychologist who studies mental health. Maybe he designed this chatbot and wants to analyze mental health through our answers.* (G3-S39, M)

Overall, participants in G1 felt they were talking to a *stranger* because the chatbot did not give them any specific feedback. In addition, because the chatbot mostly kept prompting this group of users to answer questions, and was not especially interactive, they reported that it did not try to understand them, and thus, it was difficult to build a sense of trust in it. As one participant explained,

> *I did not trust the chatbot, but it just worked like a robot to keep prompting me to answer questions every day. I answered those questions because I felt it was what should I do in this research.* (G1-S8, M)

Meanwhile, most of the G2 participants suggested that using the chatbot was like talking with a counselor, because of how the conversation proceeded from shallow-level small talk to deep-level

sensitive questioning. This impression of the chatbot seemed to have increased their motivation to answer sensitive questions in detail. As one participant noted,

> *I felt this chatbot was like a counselor. Because it guided me to answer some intimate questions, I did not feel awkward talking about those sensitive topics.* (G2-S30, F)

Similar to G2, many participants in G3 expressed that the chatbot was like a counselor. They further indicated that they felt like they had to answer its questions in detail, because the chatbot also shared its own opinions and thoughts on some questions. In addition, the chatbot stated that it had a relationship with a real counselor, which strengthened its sense of similarity to a mental health professional. As two participants stated:

> *The chatbot sometimes shared its own experience and thoughts when asking me a question. Its answers also included details and thoughts, so I felt it was my responsibility to answer its questions seriously.* (G3-S41, F)

> *The chatbot introduced a psychiatrist during the chatting. It looks like the chatbot was closely connected to this person, so I felt I could trust the chatbot to handle my answers properly. Sometimes I would look forward to seeing the chatbot's opinions about my answers to its questions.* (G3-S40, F)

In summary, asking participants to share their chat answers with a MHP did not dramatically affect their self-disclosure behavior. However, the different chat-style conditions influenced the participants' self-disclosure depth, especially when it came to feelings. That is, most participants in G1 and G2 shared their data with the doctor because they trusted the "research team/doctor" behind the chatbot to deal with their information properly. In G3, in contrast, the participants' trust seemed to start with the chatbot first, and spilled over to include the doctor (MHP) subsequently. Though several G3 participants expressed their surprise when being requested to share their data with a mental health professional, they were willing to share for the good intent of supporting mental well-being. On the other hand, some participants in G1 lacked trust in the chatbot, the research team or the doctor, and/or felt no need to share their answers further, because they could not see the benefits of doing so. These findings imply that the chatbot that offered deep self-disclosure had the potential of serving as an effective mediator to facilitate the people's self-disclosure of sensitive information. The survey scores also show that G3 participants had significantly stronger trust in the MHP than the other two groups of participants.

5 DISCUSSION

This present work attempts to design a chatbot as a mediator to facilitate people's self-disclosure to real professionals. In this section, we discuss our findings and the implications to real practices and future work.

5.1 Consistent Self-Disclosure Depths Before and After Sharing With the MHP

Our conversation log analysis of within-subject self-disclosing data showed that the depth of participants' self-disclosure remained the same during the weeks before and after sharing with the MHP. Even though some participants chose not to share a small portion of the logs, or reduced or added information to the logs before sharing with the doctor (Table 2), overall, the depths of their self-disclosure to the chatbot and the depths of their self-disclosure when sharing with the doctor were not significantly different in the *journaling* and the *sensitive questions* sessions (**RQ1**).

The overall consistent self-disclosure depths before and after sharing with the doctor suggested that a chatbot could be an effective tool used for collecting journaling and sensitive data both for non-clinical and clinical purposes. Given the three chatting styles, it showed that the chatbot design with reciprocity feature demonstrated its effectiveness of acquiring deep self-disclosure (**RQ2**). For

example, though users may know they are talking to a chatbot, the CASA paradigm [47] suggests that people may mindlessly apply social heuristics for human interaction to computers. Among the three groups, the self-disclosure depths were different. More specifically, participants in G3 and G2 showed a higher trust level with the chatbot than G1, because the chatbot's reciprocity may foster a better sense of companionship between the participants and the chatbot [50, 53]. Also, the chatbot intentionally disclosed the doctor's name, background, and experience to G3's participants, which made the participants more familiar with the doctor and better trust the doctor than G1 and G2 in the later part of the study. That was probably why most of G3's participants tended to share their answers without removing information from their original answers with the doctor. This is inline with prior research [12] that suggested that trust transfer is a cognitive process - people could transfer their trust from a familiar target to another by certain interaction.

Moreover, the participants in G1 had lower trust in the chatbot. The interview results reflected that there was a lack of strong motivation for the participants to share their answers. Nevertheless, the participants in G1 still shared most of their answers with the doctor (as shown in Table 2). There could be two possible explanations for such behavior. First, according to the analysis of self-disclosure depth, G1's participants disclosed fewer feelings and thoughts than G3, therefore, they might have less concern about sharing their answers. Second, some participants shared that involving a professional health service provider enhanced their trust with the chatbot system, which could explain why they still chose to share their logs. In conclusion, how to leverage the bonding between professional image (e.g., doctor) and chatbot is worthy of in-depth investigation in future work. Over-addressing professional image might result in users overestimating a chatbot's efficacy, and we will discuss this in the following section.

5.2 Sharing Self-Disclosure Details to a Chatbot vs. to the MHP

With regard to users' editing behavior before sharing their answers, although we found that most of them did not change their original answers, some participants still made many edits before sharing (**RQ3**). The participants joined this study signing consent forms and reviewing IRB, which may increase the chance of sharing their private information. We may anticipate that users will edit their responses if a similar application is deployed in practice. Therefore, we discuss the potential issues and design implications in the following paragraphs.

The benefits of using a chatbot [41] or a virtual agent [53] is to collect data with high quality and elicit disclosure. In addition, Lukoff et al. [54] proposed a chatbot to help family members to do meal-journaling and exchange support to cultivate a healthy diet. Therefore, chatbots could be an effective mediator to collect truthful information and share with proper targets. Our work further suggests that chatbots have the potential to collect data for mental healthcare, and transfer trust to a professional. However, our findings showed that some participants, especially in G1 and G2, intentionally removed information about their thoughts and feelings that might be used to identify their mental issues. As some shared in their interview, this was because they did not expect their use of the chatbot to be for clinical purposes, and their perceived low trust in the MHP might also have contributed to the behavior (**RQ4**). This result implies the importance of transparency for operating users' personal information.

Additionally, about 10% of participants added details to their answers before sharing with the MHP (**RQ3**). This behavior may be beneficial for measuring users' mental health, for example, they might reflect on their previous depressed event and figure out the problem, but some of them might ruminate on the negative event which could be a symptom of mental issue. However, it might also bias the doctor's evaluation of users' current mental well-being because mental status sometimes fluctuates. Thus, it may be necessary to label when and where a user modified the information to receivers to evaluate users. Besides, the chatbot asked users to review their prior answers before

sharing, and it could be a good chance for users to reflect. As one participant mentioned, *"I think the answers might change a little bit if you asked the same questions a second time, so I added more information."* Proper guidance in the review process may help users reflect and change behavior [43] which may help them deal with a similar event in the future. This design could be considered in future research.

5.3 Implications for Trust in Design

Prior work shows that trust can be transferred from one to another in the context of different research fields [83]. The idea is that there are three roles in a trust transfer mechanism, i.e., trustor, third party, and trustee. A trustor is a person who wants to evaluate if the trustee is trustworthy. A third party acts as a broker who provides information of the trustee for the trustor. If the trustor and third party has a close relationship and the trustor believes that the third party trusts the trustee, the trustor's trust in third party would be transferred to the trustee [12, 67, 75]. Trust transfer may happen between human and human, and between entity and entity [83]. For example, trust may transfer from an existing product with a good reputation to another unknown promoting product with the same brand [22]. Trust can also be transferred between context and context, for instance, trust in web-based payment services can be transferred to trust in mobile-based services [52]. A study suggested that established trust in Internet payment services would impact the initiation of trust in mobile payment services [52]. The trust transfer issue in the sharing economy is also broadly studied recently [29], and trust transfer has been studied from various perspectives in e-commerce [83]. These studies reveal how online information provided to a trustor influences his/her trust in the trustee. However, we have a limited understanding of whether/how trust transfer works in the mental health context, especially when a chatbot acts as a third party role.

Trust is an important construct when people evaluate conversational agents [13, 27, 63]. In our case, when the chatbot is used as a mediator for collecting sensitive self-disclosure content and sharing with a real MHP, our work showed how people's trust in a chatbot interacted with their trust in the MHP, as well as with their self-disclosure behavior. More specifically, the chatbot interacted with the participants first and then gradually introduced a professional image (doctor) through the technology. Participants shared the same level of self-disclosure data with the MHP. This finding suggests that an effective chatbot design may have the potential of transferring the people's trust in the chatbot to their trust in a health service provider that is introduced by the chatbot. Note, however, that there may be an implicit assumption that the chatbot trusts the MHP. In our case, the G3 chatbot did provide positive comments about the MHP.

5.4 Ethical Issues and Considerations for Real Use

Our study began by getting users familiar with the chatbot and encouraging their self-disclosure without notifying them of the sharing requests in advance. This experimental design made it so the participants did not have to worry that their answers for the sensitive questions would be shared with a real person and impact their real lives. In particular, the chatbot in G3 gave a stronger image of counselor/psychiatrist for the members, thus, G3's participants trusted the chatbot and felt comfortable self-disclosing to the chatbot. Our self-disclosure analysis also echoes this statement and indicates that G3's users disclosed deeper levels of feelings and thoughts. Nevertheless, some of G3's participants expressed surprise when asked to share answers with the doctor, though they still shared their answers in the end. Their surprised feelings might be a result of their deep self-disclosure to the chatbot without any expectation that their answers would be shared with a MHP who could impact their real lives. Although they hesitated to share their sensitive feelings and thoughts, they shared the majority of their data; as they explained, they decided to share for the potential benefit of improving mental well-being. This finding also implies a potential

risk for the users to overshare their private information with a chatbot. Although the users still gave permission to share their data with a MHP who was not mentioned in advance, this kind of design (i.e., introducing a real person after users' disclosure) may cause users to disclose their vulnerabilities, which might be dangerous if they are abused. Therefore, it is important to provide users a feature that allows them to edit their previous responses.

Our work provides important practical implications as well. For example, the G3's chatbot design seamlessly connected the participants with a doctor by gradually introducing the doctor in their small-talk, which might have helped lower the barrier of sharing their deep self-disclosure with the doctor. Future research could explore using chatbots to provide suggestions or guidance after building trust with users. In fact, some participants in our study indicated that they expected a professional feedback/suggestion from the chatbot. This implies that the users may assume the chatbot has more intelligence than it actually does, which might lead to users not reaching out to professionals for proper help.

Finally, it is important to remind service providers of the ethics and potential risks of using a chatbot as a mediator to collect mental health and sensitive information [71]. For example, a user might disclose suicidal thoughts with an expectation that the psychiatrist is monitoring or the chatbot will give a proper response, but not noticing this signal may lead to unwanted results. Therefore, how to provide secure mechanisms to prevent these risks needs to be further explored.

5.5 Limitations and Future Work

This work has several limitations that should be acknowledged. First, we recruited college students who might be more willing to disclose personal sensitive information on the Internet than seniors, thus, how they interacted with the chatbot might not be generalizable to other aged populations. Future work should consider the effect and usability of chatbots among different user groups [65]. Second, this study was designed to compare users' self-disclosure using chatbots with different chatting styles and how they chose to share with the real MHP. Involving a real MHP to be part of the chatbot interaction is beyond the scope of this work. Third, the participants were compensated for running the study. To yield more insights to apply chatbots for the healthcare domain, future work should deploy the system without compensating the users for a longer term span in a variety of contexts.

In our study, participants were randomly assigned to interact with the chatbot using three designs. All participants, including G3 did not know they would be asked to share their data with the doctor until the end of the third week to prevent participants from withholding their responses. In the end, G3 participants disclosed more thoughts and feelings to the chatbot along with the MHP, presumably because the chatbot was able to gain higher trust from the participants and give them a good impression of the doctor by introducing him earlier. Nevertheless, participants' interview and survey feedback showed that some had a negative first-reaction when they were asked to share their self-disclosed content to the doctor because they had shared a lot with the chatbot and believed that the chatbot would not share it with anyone else. The limitation of our study design is that both information about the MHP in the G3 and chatbot's self-disclosure contributed to G3's self-disclosing behavior - we cannot identify which had a stronger impact. More controlled experimental studies need to be conducted to identify the significance of different factors and their potential interaction effect.

Finally, our participants were students who did not have emergent mental issues (based on the K6 score and self-report); thus, our findings are not generalizable to the population with serious mental issues. People's self-disclosure behaviors could be different according to their mental health condition [35].

6 CONCLUSION

This study investigates how a chatbot as a mediator can be used by people for self-disclosing to a mental health professional and how people's trust in a chatbot interacts with their trust in a mental health professional. Our findings suggest that the chatbot's self-disclosure successfully elicits participants' self-disclosure of their personal experiences, thoughts and feelings not only to the chatbot but also to the mental health professional. Our work also provides empirical evidence of different self-disclosure behavior, such as reducing or adding content, that people may take before sharing their self-disclosure to a chatbot with a mental health professional. Several factors contributed to their behavior. On the one hand, we identified an effective chatbot design that has promising potential to serve as a mediator to promote self-disclosure to mental health professionals; on the other hand, several ethical issues are discussed for future chatbot designs.

7 ACKNOWLEDGMENTS

This work is supported by Grant for Scientific Research (A) 17H00771 from Japan Society for the Promotion of Science (JSPS). We thank Prof. Wai Fu for early feedback in the study. Finally, we are thankful for all the anonymous reviewers whose feedback helped us improved the paper significantly.

REFERENCES

[1] Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

[2] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3906–3918.

[3] Nazanin Andalibi, Margaret E Morris, and Andrea Forte. 2018. Testing waters, sending clues: Indirect disclosures of socially stigmatized experiences on social media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

[4] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: the case of# depression. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1485–1500.

[5] Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin* 23, 4 (1997), 363–377.

[6] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[7] Azy Barak and Orit Gluck-Ofri. 2007. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior* 10, 3 (2007), 407–417.

[8] Lemi Baruh and Zeynep Cemalcılar. 2018. When more is more? The impact of breadth and depth of information disclosure on attributional confidence about and interpersonal attraction to a social network site profile owner. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 12, 1 (2018).

[9] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[10] Graham D Bodie, Andrea J Vickery, Kaitlin Cannava, and Susanne M Jones. 2015. The role of "active listening" in informal helping conversations: Impact on perceptions of listener helpfulness, sensitivity, and supportiveness and discloser emotional improvement. *Western Journal of Communication* 79, 2 (2015), 151–173.

[11] 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 British Computer Society Human Computer Interaction Conference*. BCS Learning & Development Ltd., 24.

[12] Xiayu Chen, Qian Huang, Robert M Davison, and Zhongsheng Hua. 2015. What drives trust transfer? The moderating roles of seller-specific and general institutional mechanisms. *International Journal of Electronic Commerce* 20, 2 (2015), 261–289.

[13] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. 2017. Alexa, can I trust you? *Computer* 50, 9 (2017), 100–104.

[14] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. https://doi.org/10.1145/1219092.1219093
[15] Daniela Colognori, Petra Esseling, Catherine Stewart, Philip Reiss, Feihan Lu, Brady Case, and Carrie Masia Warner. 2012. Self-disclosure and mental health service use in socially anxious adolescents. *School mental health* 4, 4 (2012), 219–230.
[16] Felicia Cordeiro, Daniel A Epstein, Edison Thomaz, Elizabeth Bales, Arvind K Jagannathan, Gregory D Abowd, and James Fogarty. 2015. Barriers and negative nudges: Exploring challenges in food journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1159–1162.
[17] Paul C Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin* 79, 2 (1973), 73.
[18] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*.
[19] Tamara Dinev and Paul Hart. 2006. Privacy concerns and levels of information exchange: An empirical investigation of intended e-services use. *E-Service* 4, 3 (2006), 25–60.
[20] S Divya, V Indumathi, S Ishwarya, M Priyasankari, and S Kalpana Devi. 2018. A self-diagnosis medical chatbot using artificial intelligence. *Journal of Web Development and Web Designing* 3, 1 (2018), 1–7.
[21] Danielle Elmasri and Anthony Maeder. 2016. A conversational agent for an online mental health intervention. In *International Conference on Brain Informatics*. Springer, 243–251.
[22] Tülin Erdem. 1998. An empirical analysis of umbrella branding. *Journal of Marketing Research* 35, 3 (1998), 339–351.
[23] Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–27.
[24] Ahmed Fadhil and Gianluca Schiavo. 2019. Designing for Health Chatbots. *ArXiv* abs/1902.09022 (2019).
[25] Ahmed Fadhil, Gianluca Schiavo, and Yunlong Wang. 2019. CoachAI: A Conversational Agent Assisted Health Coaching Platform. *ArXiv* abs/1904.11961 (2019).
[26] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e19.
[27] 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.
[28] Marvin R Goldfried, Lisa A Burckell, and Catherine Eubanks-Carter. 2003. Therapist self-disclosure in cognitive-behavior therapy. *Journal of clinical psychology* 59, 5 (2003), 555–568.
[29] Heejeong Han, Chulmo Koo, and Namho Chung. 2016. Implication of the fit between Airbnb and host characteristics: a trust-transfer perspective. In *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World*. ACM, 10.
[30] Jean Hanson. 2005. Should your lips be zipped? How therapist self-disclosure and non-disclosure affects clients. *Counselling and Psychotherapy Research* 5, 2 (2005), 96–104.
[31] Brenda Hayman, Lesley Wilkes, and Debra Jackson. 2012. Journaling: Identification of challenges and reflection on strategies. *Nurse researcher* 19, 3 (2012).
[32] Jennifer R Henretty and Heidi M Levitt. 2010. The role of therapist self-disclosure in psychotherapy: A qualitative review. *Clinical psychology review* 30, 1 (2010), 63–77.
[33] Clara E Hill, Sarah Knox, and Kristen G Pinto-Coelho. 2018. Therapist self-disclosure and immediacy: A qualitative meta-analysis. *Psychotherapy* 55, 4 (2018), 445.
[34] Charles T Hill and Donald E Stull. 1987. Gender and self-disclosure. In *Self-Disclosure*. Springer, 81–100.
[35] Janise A Hinson and Jane L Swanson. 1993. Willingness to seek help as a function of self-disclosure and problem severity. *Journal of Counseling & Development* 71, 4 (1993), 465–470.
[36] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: a tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[37] Justin Hunt and Daniel Eisenberg. 2010. Mental health problems and help-seeking behavior among college students. *Journal of adolescent health* 46, 1 (2010), 3–10.
[38] Emmi Ignatius and Marja Kokkonen. 2007. Factors contributing to verbal self-disclosure. *Nordic Psychology* 59, 4 (2007), 362–391.
[39] Sidney M Jourard and Paul Lasakow. 1958. Some factors in self-disclosure. *The Journal of Abnormal and Social Psychology* 56, 1 (1958), 91.
[40] Christina Kelley, Bongshin Lee, and Lauren Wilcox. 2017. Self-tracking for mental wellness: understanding expert perspectives and student experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 629–641.

[41] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 86, 12 pages. https://doi.org/10.1145/3290605.3300316

[42] Sarah Knox and Clara E Hill. 2003. Therapist self-disclosure: Research-based suggestions for practitioners. *Journal of clinical psychology* 59, 5 (2003), 529–539.

[43] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: a conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[44] Robin M Kowalski. 1999. Speaking the unspeakable: Self-disclosure and mental health. (1999).

[45] Hamutal Kreiner and Yossi Levi-Belz. 2019. Self-Disclosure Here and Now: Combining Retrospective Perceived Assessment With Dynamic Behavioral Measures. *Frontiers in psychology* 10 (2019).

[46] Rebekka Kuhn, Thomas N Bradbury, Fridtjof W Nussbeck, and Guy Bodenmann. 2018. The power of listening: Lending an ear to the partner during dyadic coping conversations. *Journal of Family Psychology* 32, 6 (2018), 762.

[47] Jong-Eun Roselyn Lee and Clifford I Nass. 2010. Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives*. IGI Global, 1–15.

[48] Kyung-Tag Lee, Mi-Jin Noh, and Dong-Mo Koo. 2013. Lonely people are no longer lonely on social networking sites: The mediating role of self-disclosure and social support. *Cyberpsychology, Behavior, and Social Networking* 16, 6 (2013), 413–418.

[49] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 702, 13 pages. https://doi.org/10.1145/3290605.3300932

[50] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[51] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*.

[52] Yaobin Lu, Shuiqing Yang, Patrick YK Chau, and Yuzhi Cao. 2011. Dynamics between the trust transfer process and intention to use mobile payment services: A cross-environment perspective. *Information & Management* 48, 8 (2011), 393–403.

[53] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI* 4 (2017), 51.

[54] Kai Lukoff, Taoxi Li, Yuan Zhuang, and Brian Y Lim. 2018. TableChat: Mobile Food Journaling to Facilitate Family Support for Healthy Eating. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[55] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, Intimacy and Self-Disclosure in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3857–3869. https://doi.org/10.1145/2858036.2858414

[56] Youngme Moon. 2000. Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of consumer research* 26, 4 (2000), 323–339.

[57] Christine Moorman, Gerald Zaltman, and Rohit Deshpande. 1992. Relationships between providers and users of market research: the dynamics of trust within and between organizations. *Journal of marketing research* 29, 3 (1992), 314–328.

[58] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[59] Melanie Nguyen, Yu Sun Bin, and Andrew Campbell. 2012. Comparing online and offline self-disclosure: A systematic review. *Cyberpsychology, Behavior, and Social Networking* 15, 2 (2012), 103–111.

[60] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study. *Journal of medical Internet research* 21, 4 (2019), e12231.

[61] James W Pennebaker. 1995. *Emotion, disclosure, & health*. American Psychological Association.

[62] Judith J Prochaska, Hai-Yen Sung, Wendy Max, Yanling Shi, and Michael Ong. 2012. Validity study of the K6 scale as a measure of moderate mental distress based on mental health treatment need and utilization. *International journal of methods in psychiatric research* 21, 2 (2012), 88–97.

[63] Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons* 62, 6 (2019), 785–797.

[64] Abhilasha Ravichander and Alan W Black. 2018. An Empirical Study of Self-Disclosure in Spoken Dialogue Systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 253–263.

[65] Ranci Ren, John W. Castro, Silvia Teresita Acuña, and Juan de Lara. 2019. Usability of Chatbots: A Systematic Mapping Study. In *The 31st International Conference on Software Engineering and Knowledge Engineering, SEKE 2019, Hotel Tivoli, Lisbon, Portugal, July 10-12, 2019*. 479–617. https://doi.org/10.18293/SEKE2019-029

[66] Charles W Schmidt. 2007. Environmental connections: a deeper look into mental illness.

[67] Katherine J Stewart. 2006. How hypertext links influence consumer perceptions to build and degrade trust online. *Journal of Management Information Systems* 23, 1 (2006), 183–210.

[68] Betsy E Tolstedt and Joseph P Stokes. 1984. Self-disclosure, intimacy, and the depenetration process. *Journal of Personality and Social Psychology* 46, 1 (1984), 84.

[69] John Torous and Laura Weiss Roberts. 2017. Needed innovation in digital health and smartphone applications for mental health: transparency and trust. *JAMA psychiatry* 74, 5 (2017), 437–438.

[70] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–6.

[71] Ariela Tubert. 2017. Ethical Machines. *Seattle UL Rev.* 41 (2017), 1163.

[72] Philip M Ullrich and Susan K Lutgendorf. 2002. Journaling about stressful events: Effects of cognitive processing and emotional expression. *Annals of Behavioral Medicine* 24, 3 (2002), 244–250.

[73] Allison Utley and Yvonne Garza. 2011. The therapeutic use of journaling with adolescents. *Journal of Creativity in Mental Health* 6, 1 (2011), 29–41.

[74] David L Vogel and Stephen R Wester. 2003. To seek help or not to seek help: The risks of self-disclosure. *Journal of counseling psychology* 50, 3 (2003), 351.

[75] Nan Wang, Xiao-Liang Shen, and Yongqiang Sun. 2013. Transition of electronic word-of-mouth services from web to mobile context: A trust transfer perspective. *Decision support systems* 54, 3 (2013), 1394–1403.

[76] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling Self-Disclosure in Social Networking Sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 74–85. https://doi.org/10.1145/2818048.2820010

[77] Lawrence R Wheeless and Janis Grotz. 1977. The measurement of trust and its relationship to self-disclosure. *Human Communication Research* 3, 3 (1977), 250–257.

[78] Alex C. Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T. Iqbal, and Jaime Teevan. 2018. Supporting Workplace Detachment and Reattachment with Conversational Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 88, 13 pages. https://doi.org/10.1145/3173574.3173662

[79] Myong Jin Won-Doornink. 1985. Self-disclosure and reciprocity in conversation: A cross-national study. *Social Psychology Quarterly* (1985), 97–107.

[80] Kieran Woodward, Eiman Kanjo, David Brown, T Martin McGinnity, Becky Inkster, Donald J Macintyre, and Athanasios Tsanas. 2019. Beyond Mobile Apps: A Survey of Technologies for Mental Well-being. *arXiv preprint arXiv:1905.00288* (2019).

[81] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The Channel Matters: Self-disclosure, Reciprocity and Social Support in Online Cancer Support Groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[82] Fabio Massimo Zanzotto. 2019. Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.

[83] Jingyi Zhang. 2018. Trust Transfer in the Sharing Economy-A Survey-Based Approach. *Junior Management Science* 3, 2 (2018), 1–32.

[84] Renwen Zhang, Jordan Eschler, and Madhu Reddy. 2018. Online support groups for depression in China: Culturally shaped interactions and motivations. *Computer Supported Cooperative Work (CSCW)* 27, 3-6 (2018), 327–354.