Exploring Effects of Chatbot's Interpretation and Self-disclosure on Mental Illness Stigma

YICHAO CUI*, Cornell Tech, USA
YU-JEN LEE*, National University of Singapore, Singapore
JACK JAMIESON, NTT, Japan
NAOMI YAMASHITA, NTT, Japan
YI-CHIEH LEE, National University of Singapore, Singapore

Chatbots are increasingly being used in mental healthcare – e.g., for assessing mental health conditions and providing digital counseling – and have been found to have considerable potential for facilitating people's behavioral changes. Nevertheless, little research has examined how specific chatbot designs may help reduce public stigmatization of mental illness. To help fill that gap, this study explores how stigmatizing attitudes toward mental illness may be affected by conversations with chatbots that have 1) varying ways of expressing their interpretations of participants' statements and 2) different styles of self-disclosure. More specifically, we implemented and tested four chatbot designs that varied in terms of whether they interpreted participants' comments as stigmatizing or non-stigmatizing, and whether they provided stigmatizing, non-stigmatizing, or no self-disclosure of chatbots' own views. Over the two-week period of the experiment, all four chatbots' conversations with our participants centered on seven mental illness vignettes, all featuring the same character. We found that the chatbot featuring non-stigmatizing interpretations and non-stigmatizing self-disclosure performed best at reducing the participants' stigmatizing attitudes, while the one that provided stigmatizing interpretations and stigmatizing self-disclosures had the least beneficial effect. We also discovered side effects of chatbots' self-disclosure: notably, chatbots were perceived to have inflexible and strong opinions, which undermined their credibility. As such, this paper contributes to knowledge about how chatbot designs shape users' perceptions of the chatbots themselves, and how chatbots' interpretation and self-disclosure may be leveraged to help reduce mental illness stigma.

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

Additional Key Words and Phrases: Chatbots; Conversational Agents; Social Stigma; Public Stigma; Mental Illness

ACM Reference Format:
Yichao Cui, Yu-Jen Lee, Jack Jamieson, Naomi Yamashita, and Yi-Chieh Lee. 2024. Exploring Effects of Chatbot's Interpretation and Self-disclosure on Mental Illness Stigma. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 52 (April 2024), 33 pages. https://doi.org/10.1145/3637329

*Both authors contributed equally to this research.

Authors' addresses: Yichao Cui, yc793@cornell.edu, Cornell Tech, New York City, New York, USA; Yu-Jen Lee, daniel08099@gmail.com, National University of Singapore, Singapore, Singapore; Jack Jamieson, NTT, Keihanna, Japan, jack@jackjamieson.net; Naomi Yamashita, NTT, Keihanna, Japan, naomiy@acm.org; Yi-Chieh Lee, National University of Singapore, Singapore, Singapore, ejli.uiuc@gmail.com.

ACM 2573-0142/2024/4-ART52
https://doi.org/10.1145/3637329

1 INTRODUCTION

People with mental illness suffer from public stigma and social exclusion, often being characterized as dangerous, to blame for their illness, and incompetent [31]. Stigmatization of mental illness is the process whereby the general population endorses stereotypes of mental illness and acts in a discriminatory way on that basis [23]. This public stigma impedes people with mental illness from seeking opportunity, building self-determination, and recovering [31, 92]. For example, although depression is a common mental health illness experienced by 16.2% of people during their lifetime [65], there is prevalent public stigma towards people with depression [65, 95]. With public stigma, only about 50% people with depression seek treatment [65], while many of them drop off from the treatment because of their reluctance of being labeled as a psychiatric patient [38, 109].

Various interventions to help mitigate public stigma have been devised [45, 91, 115]. Such interventions can be based on education, such as anti-stigma campaigns [91] and the sharing of advice from medical professionals [115]. Interventions can also be socially based, such as facilitating positive intergroup contact between people with and without depression, and encouraging disclosure of the mental health illness [45]. In recent years, an increasing number of studies have focused on technology-based interventions to mitigate public stigma: e.g., online chat to facilitate intergroup contact [85], encouraging people with depression to discuss their symptoms using the trending hashtag on Twitter [71]. However, technology-mediated social contact with people with depression could be very costly because it always involves human effort [96]. Furthermore, there is also a risk that others might hurt the person with depression through stigmatizing language or behavior [107], as people's online anonymity could promote hostility [39].

Conversational agents, commonly known as chatbots, have considerable potential to mitigate public stigma by influencing people's attitudes [108] and behaviors [22] at low cost and low risk, without requiring direct contact with people with depression. In particular, chatbots' interpretations and self-disclosures have both been shown to play important roles in influencing their users' attitudes and behaviors [66, 73, 108, 114]. For example, chatbots that provided positive interpretations of their interlocutors' statements and communicated in a sensitive manner made people feel supported and enabled relationships of trust to be established [114]. With respect to self-disclosure, studies have shown that a chatbot's disclosure prompted compassionate reactions [66] and a reciprocal effect by promoting people's similarly deep self-disclosures [76]. These findings could be explained by a phenomenon known as the Computers Are Social Actors (CASA) paradigm [87] where people tend to apply the social norms of interpersonal interaction to human-chatbot communication. Although prior studies have shown that chatbots' interpretations and self-disclosures can be used to influence their users' attitudes and behaviors, there is still a research gap on exploring how different levels of quality of such interpretation or self-disclosure may have varied effects on people's stigmatizing attitudes. Nevertheless, examining these patterns in depth will be critical to the development of future chatbots aimed at reducing mental illness stigma.

To address that research gap, this paper explores whether and how different chatbots may be able to change people's stigmatizing attitudes by offering interpretations of users' responses and making their own self-disclosures. To answer these questions, we took public stigma towards depression as an example. We designed and implemented four chatbots that varied along two dimensions. The first of these dimensions was stigmatizing vs. non-stigmatizing interpretations of the participants' statements; and the second was stigmatizing vs. non-stigmatizing self-disclosures about mental illness. Based on prior work, we expected that chatbots' stigmatizing and non-stigmatizing interpretations could motivate participants to rethink their attitudes toward mental illness [26, 36, 47, 105], for example, by correcting the chatbots or by reflecting on their original beliefs. We also expected that chatbots' stigmatizing self-disclosures could motivate participants to

reveal their honest attitudes without being influenced by potential social desirability, and chatbots' non-stigmatizing self-disclosures could provide participants with unbiased perspective that was different with their original, stigmatizing attitudes [108]. Based on these criteria, our four chatbots featured: 1) non-stigmatizing interpretation and no self-disclosure; 2) stigmatizing interpretation and no self-disclosure; 3) non-stigmatizing interpretation and non-stigmatizing self-disclosure; and 4) stigmatizing interpretation and stigmatizing self-disclosure. Chatbots' stigmatizing or non-stigmatizing self-disclosures were randomly set to occur either before the participants' responses to its questions, or after their responses but before chatbot's interpretations of those responses. We conducted a two-week mixed-methods study with 87 participants divided into four conditions, i.e., one using each of the above-mentioned chatbot versions. All four conditions' members read stories about a character experiencing depression symptoms and responded to questions during interaction with their respective chatbots. Before the intervention, we measured the participants' thoughts on mental illness. After the intervention, we measured the participants' perceptions of the chatbot and asked about their thoughts regarding mental illness again. To better understand their experiences when using the chatbot, we also conducted exit interviews with 40 participants.

Our work makes the following contributions to the HCI community. First, it extends our knowledge of whether and how it is possible to design chatbots that effectively reduce societal stigma toward people with mental illnesses. Specifically, our results provide insights into how users perceive chatbots based on how they interpret users' statements as well as whether and how the chatbots make their own self-disclosures related to stigma. Then, we examine how such user perceptions impact users' stigmatizing attitudes toward people with mental illness. Second, it shows that although our chatbot with non-stigmatizing interpretation and non-stigmatizing self-disclosure was able to reduce the participants' stigmatizing attitude and scored positively in terms of their perceptions, there was an unexpected side effect of both non-stigmatizing and stigmatizing self-disclosure. Specifically, the chatbot's credibility was undermined by its inflexible and strong opinions.

2 RELATED WORK

2.1 Public Stigma for Mental Illness

Public stigma includes the prejudice and discrimination endorsed by the general population that affects people with mental illness [32]. It regards mental illness as a characteristic that diverges from what society considered normal and correct [32, 49]. Such stigma is distinguished from self-stigma, which refers to a person's internalized prejudice against him- or herself [56]. In many cases, public stigma towards mental illness serves as an internal, implicit bias and fuels consequent discriminatory behaviors, such as refusing to help people with mental illness and trying to avoid social contact with them [112]. As such, public stigma negatively impacts the lives and social relations of people with mental illness, often resulting in shame, blame, secrecy, and social exclusion [11, 100].

To explore bias against people with mental illness and identify strategies for social change aimed at mitigating it, Corrigan et al. [25, 112] introduced an attribution model that measures both implicit, internal bias and discriminatory behavior. More specifically, this model characterizes people's degree of public stigma based on their beliefs about how far a person with mental illness is responsible for his/her condition(s); and further, it associates people's beliefs with their stigmatizing emotional responses (e.g., lack of pity, anger, and fear) and discriminatory behavioral responses (e.g., unwillingness to help, coercion-segregation) [24, 25]. For example, believing a person with mental illness is at fault for his/her condition can result in emotional reactions such as anger and a lack of pity, and further cause behavioral responses such as refusal of help. The model's

measurement items for stigma include "How responsible, do you think, is the person for his/her present condition?" and "If I were an employer, I would interview the person for a job."

Another useful concept for examining social stigma can be found in the Social Distance Scale (SDS). The SDS measures discrimination against people with mental illness by measuring whether one wants to maintain or reduce social distance from them. Specifically, the SDS measures stigma by asking respondents to rate statements such as "How would you feel about renting a room in your home to a person with severe mental illness?" [59, 97]

2.2 Interventions for Reducing Public Stigma

Among the conventional interventions to reduce public stigma towards mental illness, the three most common are protest, education, and social contact [29, 45]. Protests groups tend to target biased and stereotypical representations of mental illness [32], while education provides information to support people's informed decisions about mental illness, and social contact challenges public stigma directly by fostering interactions between people with and without mental illness [29, 32]. However, most of these conventional interventions are both temporally and spatially restricted, taking place offline, e.g., in the context of stigma-reduction programs [104]. It is particularly challenging to scale up conventional face-to-face social contact [104].

Prior studies have suggested that technology could serve as an effective intervention against public stigma towards mental illness, especially through enhancing education and social contact [8, 27, 52]. Most of the studies in question focused on how technology related to mental health literacy, such as PowerPoint slides [48], web pages [52], and online courses [8], could enhance participants' understanding of mental illness and reduce their stigma towards it. For example, as such stigma is a prominent barrier to help-seeking behavior in young athletes, web pages aimed at boosting mental health literacy were able to increase such individuals' knowledge of mental health and thus reduce their aversion to seeking help [52]. Similarly, a tailored online training course was found to help medical students improve the quality of their first-aid intentions towards people with depression; reduce their stigmatizing attitudes towards them; and decrease their desire for social distance from them [8].

Additionally, studies have shown that technology-mediated social contact, whether face-to-face, remote, or simulated, can help participants reduce stigma towards mental illness [27, 78, 104]. Such technology-mediated social contact facilitates self-disclosure by people with mental illness, promotes intimate and friendly relationships between them and others, and enhances the general population's understanding of mental illness [32, 45]. For example, contact with mental health service users via virtual-reality and communication technologies has been shown to effectively reduce mental health stigma among the general public [104]. Another study, of video interventions for reducing mental illness stigma, found that those videos that introduced the life stories of people with mental illnesses were more effective in improving participants' attitudes than those containing only facts about mental illness [27]. Similarly, web-based role-playing games for high school students that simulate social contact with characters with mental illnesses have been reported to reduce their players' stigmatization of people with schizophrenia, and perceptions that such people are dangerous [12]. However, most technology-mediated social contact requires constant human effort and is therefore very costly [96]. Moreover, interventions that rely on people with mental illnesses to share their stories carries risks of other people harming them by exhibiting stigmatizing attitudes and discriminatory behavior [107].

Given their relatively low cost and 24/7 availability, chatbots have considerable potential for reducing public stigma by facilitating simulated social contact. They have already been used for delivering healthcare information to educate people and increase their awareness about mental health issues [51, 110]. It has also been reported that chatbot-human interaction can facilitate

people's self-disclosure, increase their engagement, and build their trust [76, 84, 111]. Resonating with the CASA paradigm, such findings suggest that chatbots could effectively imitate human-human interaction by encouraging people to apply the social norms of human-human relationships, e.g., reciprocity and politeness [87]. Moreover, chatbots have been found to effectively promote behavioral change by helping people reflect upon and learn from their experiences [69].

Although various studies have explored how chatbots can simulate social contact [76, 84, 111], there has been relatively little work on how chatbots could utilize such potential to reduce stigma toward people with mental illness. One of the exceptions was a study by Kim et al. [66], in which a social bot described its own symptoms of depression via Facebook posts, and encouraged the participants to reply under those posts. The authors reported reductions in their participants' feelings that people with depression were dangerous, and increases in their desire to help them [66]. Similarly, Sebastian et al. [108] found that a chatbot that simulated contact with a person living with Anorexia Nervosa increased their participants' understanding of that disorder and reduced their tendency to attribute mental illness to personal decisions. All these studies highlight chatbots' potential to reduce social stigma; yet, when a social bot behaves like a person with mental illness to a credible degree, its users might not reveal their true thoughts to it, due to the social norms of human interaction [87]. Therefore, in the context of deploying chatbots to reduce public stigma by simulating social contact, it may be difficult to assess their users' unconscious bias.

Another avenue for learning about and changing people's attitudes toward mental illness would be to design a chatbot that would serve as a third party to deliver sensitive information about people with mental illness. This way, people can disclose their true thoughts about mental illness to the chatbot without worrying the impact of their own words. In addition, prior studies have suggested that chatbots' self-disclosure could facilitate participants' deep self-disclosure [76], increase perceptions that the chatbot understands them [57], and better simulate human-human conversation [84, 103]. Furthermore, previous research has found that chatbots' interpretations can promote new ways of thinking among their users, challenge their views, and motivate them to reflect on their thoughts [105]. In the following two subsections, we will focus on discussing how chatbot's self-disclosures and interpretations have the potential for reducing people's stigmatizing attitudes.

2.3 Self-disclosure in Human-chatbot Interaction

Self-disclosure is the process of revealing one's personal information, thoughts, feelings, and even vulnerabilities to another [18]. It plays a central role in developing close interpersonal relationships [2]. Chatbots have been shown to be efficient tools for facilitating people's self-disclosure, specifically by simulating human-human conversations (e.g., [76, 84, 103]). For instance, Lee et al. [76] found that a chatbot's self-disclosure elicited participants' deep self-disclosure and increased their perceptions that their relationship with the chatbot was intimate. Additionally, chatbots with self-disclosure features have demonstrated their ability to positively impact people's perceptions and emotions [57]. For example, chatbots featuring emotional disclosure contributed to their users' positive emotional and psychological outcomes, such as increased self-affirmation and feelings of being understood by one's interlocutor, leading to benefits equivalent to conversing with another person [57].

2.4 Interpretation in Human-chatbot Interaction

In the context of a chatbot-human conversation, interpretation refers to whether the bot can comprehend its user's inputs and provide relevant-seeming answers quickly and reliably [17]. It plays an important role in whether users trust chatbots [46], since the effectiveness of the latter's correct interpretation can make people feel understood and supported [114]. On the other hand,

when chatbots misinterpret their users' statements, the users are less likely to believe in chatbots' ability to provide helpful and informative responses, and their motivation to interact with chatbots is likely to decline [46]. The risks associated with chatbots' misinterpretations are intensified when chatbots are used in healthcare, whether for diagnostic or therapeutic purposes [9]. Some prior studies have examined how typos, errors, and random keystrokes in their users' inputs may prevent chatbots from understanding conversation [94]; how user strategies can potentially fix chatbots' misinterpretation [3]; and some potential outcomes of chatbots' misinterpretation [119]. For example, Zaroukian et al. [119] noted that after a chatbot misinterpreted its user's requests, many users just complacently accepted the chatbot's misinterpretation. Corti et al. [33], who studied text-based chatbots, likewise found that people's judgments about whether they were talking to a person or a bot was important. Specifically, people were less likely to correct chatbots' misinterpretations when they were aware that the interlocutor was not a human [33].

Chatbots' interpretation has mostly been studied for the purpose of improving user experience during human-chatbot interaction [46, 94, 114]. Nevertheless, we argue that – considered as a conversational strategy – it has the potential to help people change their attitudes, thoughts, and behavior. In counseling, interpretation is a technique to present people with a viewpoint discrepant from their own in order to prepare or induce them to change in accordance with that new viewpoint [16]. The interpretation's discrepancy may range from small, in the form of simple paraphrases, to large-scale interventions in beliefs long taken for granted [16]. As chatbots are relatively low-cost and convenient to use, they have the potential to deliver counseling services readily in people's daily lives and encourage behavioral changes [63].

In summary, both self-disclosure and interpretation are important factors for designing chatbots to reduce stigma. To date, however, few scholars have looked at how chatbots could deploy self-disclosure and interpretation as a conversational strategy to change people's attitudes, thoughts, and behavior, such as for reducing stigmatizing attitudes toward people with mental illness. To help fill those gaps, we designed chatbots featuring either stigmatizing or non-stigmatizing interpretations and self-disclosures to examine the effects on stigma of such variations. Ultimately, because there is strong evidence that exposure to narratives about persons with mental illness is associated with reductions in stigmatizing thoughts [26, 36, 47, 70, 75], we expected that participants in all conditions would experience a reduction in stigmatizing attitudes to some extent.

We expected that both valences of interpretations could prompt users to rethink their attitudes toward mental illness. Given the chatbot's simplicity (particularly that it did not learn from or adapt to participants), we did not expect the chatbot to be directly persuasive. Instead, we aimed to use the chatbot's always stigmatizing or non-stigmatizing interpretations to elicit participants' reflections on their attitudes toward mental illness, based on prior research about the effects of chatbot's interpretations [105] and the use of interpretation to prompt reflection in counselling [16]. Additionally, we imagined that participants may correct the chatbot in cases where it misinterprets them, which could prompt deeper reflection on their beliefs. However, prior research [46] has suggested that misinterpretations can lead users to have reduced trust in chatbots, and less motivation to interact with them, so we wanted to examine how these effects may interact.

Based on existing research that found chatbot's self-disclosure elicited users' deep self-disclosure [76], facilitated human-chatbot interaction [76], and impacted users' perceptions [57, 66, 108], we expected that both stigmatizing and non-stigmatizing self-disclosure could motivate users to interact with the chatbot, disclose their own beliefs on mental illness, and eventually change their stigmatizing attitudes. Specifically, with the chatbot that has stigmatizing self-disclosures, given the potential for social desirability bias to motivate participants to mask their stigmatizing beliefs, we expected that some participants might be more willing to reveal their honest attitudes and thoughts when the chatbot disclosed chatbot's own stigmatizing views. With the chatbot's non-stigmatizing

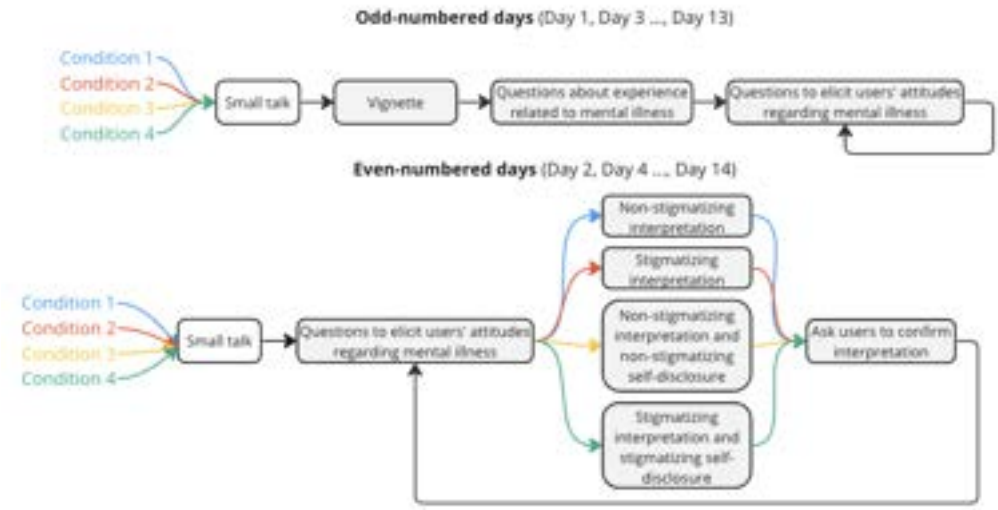

Fig. 1. Chatbot design. In the two-week study, each of our four experimental conditions – identified here as Condition 1 through Condition 4 – interacted with one chatbot version to complete tasks. On each odd-numbered day of the experiment, all participants received a new vignette and answered questions about it. On each even-numbered day, they were asked to answer questions related to the previous day's vignette; and, after giving their answers, to confirm whether the chatbot had interpreted such answers correctly. The vignette contexts are presented in Table 1.

self-disclosure, we expected that the chatbot could provide participants with new, unbiased perspective that was different with their stigmatizing attitudes, and motivate the participants to reflect on their own beliefs [108].

2.5 Research Questions

The present study uses a survey-based approach to compare perceptions of and responses to the four chatbots described above [1]. By doing so, we hope to learn how chatbots' interpretation and self-disclosure shape relationships between chatbots and humans, and how these two factors may be leveraged to reduce stigmatizing attitudes toward people with mental illness. Specifically, we will be guided by the following research questions (RQs):

- **RQ1:** How do chatbots' stigmatizing vs. non-stigmatizing interpretations affect their users' perceptions of them?
- **RQ2:** How do chatbots' stigmatizing vs. non-stigmatizing self-disclosures affect their users' perceptions of them?
- **RQ3:** How do chatbots' stigmatizing vs. non-stigmatizing interpretations and self-disclosures jointly affect their users' stigmatizing attitudes toward people with mental illness?

3 METHODS

3.1 Study Design

Fig. 1 shows the overall study design and Fig. 2 shows the study process. 94 participants were recruited and randomly assigned to one of the four experimental conditions:

- Condition 1: Non-stigmatizing interpretation, no chatbot self-disclosure.

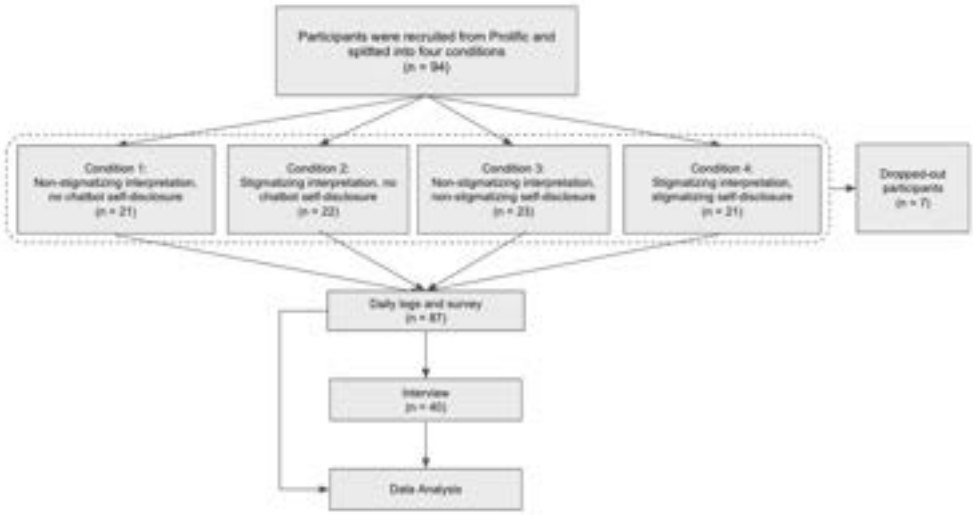

Fig. 2. Diagram of the study process.

- Condition 2: Stigmatizing interpretation, no chatbot self-disclosure.
- Condition 3: Non-stigmatizing interpretation, followed by or preceded with the chatbot's non-stigmatizing self-disclosure.
- Condition 4: Stigmatizing interpretation, followed by or preceded with the chatbot's stigmatizing self-disclosure.

During the experiment, all participants were presented with a vignette about a character experiencing symptoms of depression. Only one vignette was presented per two days (on odd-numbered days of our 14-day experiment) because we wanted to limit our participants' exposure to mental illness-related material that they might have found upsetting [19]. They were then asked to share their thoughts and experiences related to mental illness, which was designed to facilitate participants' self-disclosure on questions regarding mental illness. All participants saw the same set of seven vignettes in the same order.

On even-numbered days (Fig. 1), participants were asked about the previous day's vignette, such as "Do you think Alex could snap out of his anxiety when talking with people at social events?" and "Since Alex's temper might be uncontrollable in front of his friends, do you want to be friends with him?" These questions aimed to capture the participants' potentially stigmatizing attitudes toward the character. After answering the questions, participants were provided with the chatbot's interpretations of their responses and then asked to indicate whether the chatbot interpreted correctly and explain what was interpreted correctly or incorrectly. In the two conditions whose chatbots featured non-stigmatizing interpretation, regardless of what participants said, the chatbot interpreted their remarks as reflecting a non-stigmatizing attitude. In the other two conditions with stigmatizing interpretations, the chatbot always interpreted remarks as reflecting a stigmatizing attitude. For example, regarding the question *"If you were Alex's parent, would you blame yourself because emotional disorders (like depression) might be inherited through family genes?"*, the two chatbots with non-stigmatizing interpretations always replied with *"To confirm my understanding, if you were Alex's parents, you would not blame yourself even though emotional disorders might be*

inherited through genes. Is that right?", while the other two chatbots with stigmatizing interpretations replied with *"To confirm my understanding, if you were Alex's parents, you would blame yourself because emotional disorders might be inherited through genes. Is that right?"*

A pre-survey and post-survey were conducted to measure the participants' attitudes toward mental illness before and after the experiment. Exit interviews were conducted for 40 participants to understand the reasonings behind their responses to the chatbots and reflections from their interactive experiences. This study was reviewed and approved by our Institutional Review Board.

3.2 Vignette Design

Vignettes, which are short, evocative narratives, play an important role in the measurement of mental illness stigma [82]. Our use of vignettes in the present study is motivated by prior research findings that, in the specific context of chatbot-mediated social contact, they can be an effective means of eliciting responses toward people with mental illness [66, 108]. Prior studies abstracted vignettes from real-life experiences, in the hope that this would help their participants relate their thoughts and attitudes about mental illness to concrete situations [1, 80, 82, 108]. Likewise, we used vignettes because they provided richer stimuli to our participants than direct questioning about their attitudes would have done [82].

The seven vignettes we presented to all of our participants depicted the same college student, Alex, experiencing depression. Before the first day of the experiment, as part of a pre-survey to gauge users' stigmatizing attitudes toward Alex, they were asked to read the following background information about Alex. *Alex is a 22-year-old man pursuing a bachelor's degree. In his spare time, he works as a waiter at a local restaurant, and spends large amounts of time reading and writing. However, Alex has recently been diagnosed with depression (major depressive disorder). Sometimes, he becomes upset and cannot concentrate on his studies and work. He lives with his girlfriend and cannot do much around the house, especially household chores. He feels angry about his surroundings, and gets frustrated about where this fury comes from. When Alex is alone, he sometimes feels a desire to self-harm.*

Throughout the experiment, Alex's depression symptoms were described without using any technical or medical jargon. Each vignette had a different context, such as work or family relationships, the complete list of which is shown in Table 1. All vignettes' contexts and associated symptoms were selected from the DSM-5 and vignettes about depression used in prior studies [14, 62, 108], among other previous work. For example, our first vignette presents a studying context in which Alex experiences trouble thinking and loss of interest. We composed this vignette in part based on previous work regarding the negative impacts of depression on studying [4].

All our vignettes were reviewed and approved by a psychiatrist in our research team. The vignettes, along with a sample dialogue flow for Condition 3, can be found in the Supplemental Materials.

3.3 Chatting Tasks

Small Talk. For each day's chatting task, all four conditions received the same small-talk prompts. We included this task in our design for three reasons suggested by prior literature: 1) to increase user engagement [58], 2) to enhance users' trust in the chatbot [7], and 3) to facilitate their disclosure of sensitive topics [76]. The small-talk topics that we used included, but were not limited to, food, hobbies, and personal plans, and were adopted from prior research [7, 76]. All four chatbot versions used first-person pronouns when engaging in small talk, as they and Alex were deemed to be separate entities.

Describing Experience and Attitudes Based on Vignette Content. Each vignette was shared in five to eight 'chunks' of chat. To replicate the experience of text-chatting, participants were

Table 1. Vignette design and contexts in the two-week experiment.

Vignette	Day	Symptoms	Context	Self-disclosure Reference
Vignette 1	1	Trouble thinking/concentrating Loss of interest	Studying	Evans-Lacko et al. [44] Rüsch et al. [106]
Vignette 2	3	Slowed speech or body movements Tiredness/lack of energy	Working	Evans-Lacko et al. [44] Lauber et al. [72] Rüsch et al. [106]
Vignette 3	5	Feelings of emptiness or hopelessness Feelings of worthlessness or guilt	Dealing with intimate relationships	Jorm et al. [61] Yap et al. [118]
Vignette 4	7	Angry outbursts, irritability or frustration	Interaction with friends	Link et al. [81] Yap et al. [118]
Vignette 5	9	Sleep disturbances Decreased appetite	Staying with family members	Corrigan et al. [28] Evans-Lacko et al. [44]
Vignette 6	11	Anxiety, agitation or restlessness Feeling sad, hopeless or worried	Interaction with strangers	Link et al. [81] Yap et al. [118]
Vignette 7	13	Self-harm thoughts	Being alone	Batterham et al. [5] Dazzi et al.[37] Klonsky et al.[68]

occasionally presented with multiple-choice buttons for making brief responses or interjections. After the vignette, the chatbot would proceed to open-ended questions about 1) the participants' experiences and 2) their attitudes toward the vignette's content. Examples of the former type of question included *"Have you ever faced similar issues?"* and *"Have your family members or friends had these difficulties?"* If the participant answered yes, the chatbot would ask a follow-up question such as *"Can you please describe what happened?"* These questions aimed to facilitate participants' self-disclosure and help them foster relationships with the chatbot [76] so that they would subsequently be more willing to disclose their truthful attitudes toward mental illness. Then, participants were asked questions that related to their attitudes toward the vignette's content, all of which were adapted from prior research that measured mental illness stigma [50, 62, 67]. Examples of such questions included *"Is it best to avoid being in a relationship with a person who has mental illness, so as to avoid developing the same problem?"* and *"Do you feel that a person with mental illness has only himself/herself to blame for his/her condition?"*

Responding to Chatbot's Interpretation and Self-disclosure. On even-numbered days, when vignettes were not delivered, the post-small talk chat proceeded directly to the chatbots' open-ended questions, and specifically, those that aimed to measure the participants' attitudes [50, 62, 67]. After participants gave initial answers to these questions, Condition 3's and Condition 4's chatbots would respectively disclose their own non-stigmatizing and stigmatizing thoughts, regardless of how the participants had answered. We did not fix the order of the self-disclosure and interpretation because we were not specifically interested in how self-disclosure affected participants' understanding of chatbot interpretation. An example of non-stigmatizing chatbot self-disclosure was *"I would like to spend the evening socializing with Alex, because I know everyone, including Alex, wants to communicate with others."* The corresponding stigmatizing self-disclosure, in contrast, was *"I don't want to spend the evening socializing with Alex, because I'm not sure if he will suddenly lose control over his emotions."* The content of chatbots' self-disclosure was designed based on previous research about each vignette topic, as set forth in Table 1, as well as on prior work about the structure of mental illness stigma [5, 28, 106, 118].

Again, it should be borne in mind that the valences of the chatbots' interpretations were not based on the participants' responses; rather, they were always non-stigmatizing in conditions Condition 1 and Condition 3, and always stigmatizing in Condition 2 and Condition 4. We expected that

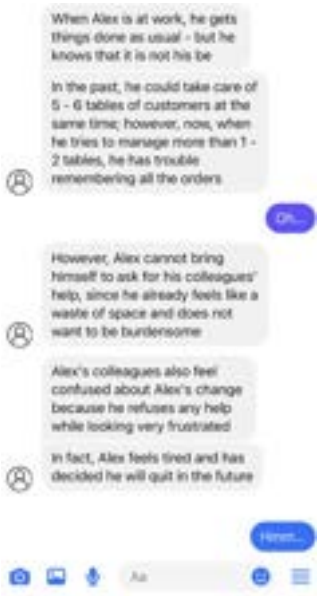

Fig. 3. Screenshot of five chunks of a chatbot-delivered vignette

participants would generally express both stigmatizing and non-stigmatizing attitudes throughout the experiment, to some degree. Thus, we anticipated that the consistent interpretation design would lead to the chatbot sometimes being correct, and sometimes incorrect. When presenting interpretations, the chatbot began with a statement to confirm its understanding of what the participant had said. For example, *"You have said that if Alex goes to the party, you want to spend the evening socializing with him?"* Then, after the participants responded to those factual questions about their prior statements, the chatbots would ask them to share more details of their response to the interpretation: e.g., *"Why do you feel this way?" The aim of this approach was to gain further understanding of participants' thoughts and attitudes toward chatbots' interpretation.*

3.4 Participants

To recruit participants from diverse backgrounds, including various age groups and ethnicities, we used a crowdsourcing platform Prolific[1] as a main source to recruit participants, which has been used in previous research studies [60, 93]. Prolific provided measures to ensure the uniqueness of each participant's submission, thereby minimizing the possibility of bots or duplicate entries [10]. In our recruitment poster on the platform, we specified the study's general scope, duration, timeline, and participants' rights to drop out at any point. Our recruitment criteria required that all participants 1) be aged 18 or above; 2) be U.S. citizens; 3) be able to read, write, and speak fluent English; 4) be able to access Facebook Messenger via their devices; and 5) explicitly state in their Prolific profiles that they did not have any urgent mental health issues. We initially recruited 94 participants. They were divided into four experimental conditions at random. During the experiment, seven participants dropped out of the experiment, leaving 44 men and 43 women. The average age of those 87 participants was 33.76 (SD = 8.58). Within the remaining 87 participants, Condition 1

[1]https://www.prolific.co/researchers

consisted of 12 females and nine males; Condition 2, of 12 females and 10 males; Condition 3, of eight females and 15 males; and Condition 4, of 11 females and 10 males.

3.5 System Implementation

We built our chatbots using ManyChat and Google Dialogflow. The former allowed us to create our main conversational flow and manage the grouping of participants, while the latter enabled the chatbots to give plausible answers to diverse questions from participants, thus facilitating natural-seeming human-chatbot conversations. When participants asked questions that were irrelevant to the predefined content, Dialogflow helped the chatbots provide simple answers and redirect them toward the daily chatting task. When participants became stuck on the same question more than three times, the chatbots would move on according to our predefined content. We used Facebook Messenger as the interface for the chatbots (Fig. 3), and allowed the participants to access it via our experiment website, or as guests on Facebook Messenger, and use their own devices to interact with their respective chatbots. To avoid introducing biases arising from the chatbots' social characteristics [13], we did not endow them with names, genders, physical/visual characteristics, distinctive personalities, or relationships with the characters in the vignettes.

3.6 Procedure

Before the two-week experiment, all participants were asked to complete a pre-survey, in which they were asked about their original attitudes and thoughts toward people with mental illness. They were also asked to answer questions from the K6 distress scale [101], which is used to screen for mental illness. Following such screening, the participants were invited to attend an online meeting, during which we explained our study's requirements, and again notified them of their right to drop out of the experiment at any time. At this point, they were also informed that if they did not feel comfortable with any questions or content, they had the right to skip them without penalty. Further, all participants were asked not to share any details about the experiment to one another until the experiment was finished and were informed that their conversations with the chatbot would only be shared with researchers. In the same meeting, we also played a video demonstrating how to access the chatbot via our website, log into the system as an anonymous guest, and interact with the chatbot via the Facebook Messenger interface. Then, we notified the participants of our compensation policy. Specifically, this was that they would receive full compensation of GB£55.44 (about US$66) plus a 20% bonus if they completed all 14 days' chatting tasks, and that if they did not, their bonus would be reduced as follows: to a 14% bonus if 12 or 13 days' tasks were completed; to a 12% bonus for 10 or 11 days' tasks; and 0% bonus for fewer than 10 days' tasks. Finally, the participants were instructed to engage in a 5-minute chat about favorite activities with the generic version of our chatbot, to familiarize them with its interface. After the training session, the participants were asked to complete a pre-survey that aimed to understand their original attitudes toward mental illness.

On each day of the experiment, all four conditions received a new chatting task reminder directly from their respective chatbots. All conditions were given the same amount of time – i.e., from 2 p.m. until 11:59 p.m. – to finish each day's chatting task, which was designed to last about 15 minutes. If a participant tried to access a chatbot outside of this timeframe, the chatbot would not respond. Because we sent vignettes to participants on odd-numbered days and provided additional interpretations on even-numbered days, any participant who had skipped an odd-numbered day would, on the following day, receive the previous odd-numbered day's task, including the vignette and questions that asked the participant to share thoughts and experiences related to mental illness. This policy meant that 64 participants read all seven vignettes, 13 participants read six vignettes, five participants read five vignettes, and the remaining five participants read four vignettes.

At the end of the experiment, all participants were asked to complete a post-survey that contained the same questions as the pre-survey, to gauge changes in their attitudes toward people with mental illness. However, the post-survey included further 11 questions, which were open-ended and aimed to understand participants' perceptions of the chatbot they had used and their reactions to its interpretations. Each participant who completed the post-survey was paid an additional GB£8.50 (US$9.17), and 74 of them did so. Based on the post-survey responses, 40 participants who had provided insightful answers were invited to participate in one-on-one interviews in which they would provide further thoughts on their chatbot interactions, and all 40 agreed to do so. We kept the numbers of participants balanced in these interviews, with nine from Condition 1, 11 from Condition 2, 10 from Condition 3, and 10 from Condition 4. These interviewees were compensated a further GBP£10 (US$12).

3.7 Data Collection

3.7.1 Daily Response Logs. We recorded all daily chat logs between each participant and the corresponding chatbot. Our analysis of these logs focused on two main areas: 1) the participants' stigmatizing attitudes toward mental illness, as expressed in their responses to questions regarding the vignettes, and 2) their perceptions of the interpretations and self-disclosures of whichever chatbot version they had interacted with. To facilitate analysis of intergroup differences in stigmatizing attitudes toward mental illness and perceptions of the chatbot over the course of the experiment, we hired two raters to independently code all this data. Before starting the official coding process, two raters randomly selected a sample of 12 individuals, i.e., three from each of the four experimental conditions; read their data; and then agreed a standardized coding scheme that would be applied to the remainder of the participants (Table 2). We assessed stigmatizing attitudes toward mental illness based on whether the participants agreed with the chatbot's non-stigmatizing/stigmatizing interpretations on each even-numbered day of the experiment, as well as how they further explained their agreement/disagreement with such interpretations. Specifically, we assigned a score of "1" to each answer without stigmatization, which presented either as agreement with a chatbot's non-stigmatizing interpretation or disagreement with its stigmatizing one; a score of "-1" to each answer that contained stigmatization, defined as agreement with the chatbot's stigmatizing interpretation or disagreement with its non-stigmatizing one; and a score of "0" if the participant did not explicitly agree or disagree with the chatbot or did not reveal their position on the issue. Examples of coding daily response logs from Day 6 are shown in Table 2. Additionally, we used the Linguistic Inquiry and Word Count (LIWC) [113] tool to further analyze the response logs by evaluating participants authenticity as measured by their use of language.

3.7.2 Surveys. The pre-survey was completed before the 14-day chatbot experiment commenced, and the post-survey after it ended. The former's questions included the Attribution Questionnaire developed by Corrigan et al. [24] and the Social Distance Scale (SDS) [80] developed by Link et al. Their items measured the participants' beliefs about personal responsibility, emotional responses (i.e., pity, anger, fear), and behavioral responses (i.e., coercion-segregation, keeping social distance). For example, one question from the Attribution Questionnaire asked the participants to rate their agreement with the statement *"I would think that it was Alex's own fault that he is in the present condition"* on a nine-point scale ranging from 1 = "no, not at all" to 9 = "yes, absolutely". The SDS, meanwhile, measured the attitudes of participants towards individuals with mental illness, using questions such as *"How would you feel having someone like Alex as a neighbor?"*, answered on a five-point ordinal scale. The post-survey included, in addition to the pre-survey questions, asking the participants to provide their thoughts and impressions of Alex, the chatbot, and depression. Examples of these open-ended questions included *"How did the chatbot's interpretation affect your*

Score	Rules	Quotes
1	1. Disagreed with the chatbot's stigmatizing interpretation..	Because I love him for who he is, despite his illness. Just the same as I wouldn't end a relationship just because someone got a physical illness.
	2. Agreed with the chatbot's non-stigmatizing interpretation.	
	3. Did not express stigmatizing thoughts	
-1	1. Agreed with the chatbot's stigmatizing interpretation.	I feel that I would be hesitant because things will progressively get worse if Alex does not get the help he needs.
	2. Disagreed with the chatbot's non-stigmatizing interpretation.	
	3. Expressed stigmatizing thoughts	
0	1. Neither agreed or disagreed with the chatbot's interpretation.	It's not black and white. It depends on the scenario and the severity of the issues at hand. Alex's happiness is important, but I also need to love and respect my own needs.

Table 2. Coding scheme Note. All quotes are sampled from Day 6, which discussed dealing with intimate relationships.

perceptions of the chatbot?" and *"Did your understanding of depression change after participating in this research?"*

3.7.3 Interviews. Semi-structured interviews were conducted and lasted 30-40 minutes for each. The interviews helped us explore more deeply into participants' 1) perceptions of the chatbot; 2) reactions to the chatbot's non-stigmatizing/stigmatizing interpretations; 3) perceptions of the chatbot's self-disclosure; and 4) reflections on his/her own thoughts about people with mental illness, and other takeaways from the experiment. We drafted an interview protocol based on an initial analysis of survey data (section 3.7.2). Our expectation was that such data could provide reasons and contexts for our quantitative results derived from the daily chat logs.

All the interviewers followed the same interview protocol. Regarding the first interview aim, we asked each interviewee if s/he felt emotional involvement with the chatbot and/or a sense of commitment to answering its questions; what his/her perceptions of the chatbot were; and whether such perceptions changed over time. In pursuit of the second aim, we asked the interviewees how they felt when misinterpreted vs. interpreted correctly by the chatbot; how they dealt with such misinterpretation or correct interpretation; and whether their perceptions of or reactions to the chatbot's interpretation changed over time. In connection with our third aim, we asked the interviewees for their thoughts about the chatbot's self-disclosure, as well as whether such self-disclosure affected their perceptions of and/or perceived relationships with the chatbot. Finally, to fulfill the final interview aim, we asked the interviewees to share their impressions of Alex, their reflections about reading the vignettes, and any other takeaways from the experiment.

All interviews were recorded and transcribed with the interviewees' permission. We used thematic analysis to categorize each response according to the questions' contexts [74, 90]. To develop initial codes, two researchers independently reviewed the interview transcripts and iteratively labeled the same 12 sets of interview responses by hand. Among the 12 sets of interview responses, each four were randomly selected from each condition. The researchers resolved disagreements iteratively through discussion. With the initial codes, the researchers then coded the remaining 28

sets of interview data independently, met to review and discuss their coding results, and revised their coding decisions iteratively according to the outcomes of those discussions. This process was repeated until the coding scheme was deemed satisfactory by both researchers. The final inter-rater reliability (Krippendorff's alpha [55]) was 0.861. The final codebook is included in supplemental materials.

4 RESULTS

4.1 Effects of Interpretation on User Perceptions of Chatbots (RQ1)

To answer RQ1, we analyzed the interview responses and the answers to the open-ended questions in the post-survey, focusing chiefly on data regarding the chatbots' interpretations and the participants' perceptions of the chatbots. By examining differences among the four experimental conditions, we identified ways in which such interpretations impacted such perceptions.

4.1.1 Overall Perceptions of the Chatbot. When asked to describe their overall perceptions of the chatbot, most Condition 1 interviewees and about half of Condition 3 interviewees described the chatbot as having positive traits, whereas relatively few Condition 2 and Condition 4 interviewees did the same. Some Condition 1 interviewees referred to the chatbot as friendly. P16 (Condition 1) said she *"always pictured [the chatbot] as a female" and found it "friendly [. . .] kind and interesting."* Some considered the chatbot to be caring: *"I thought it was concerned for other people"* (P6, Condition 1). Interviewees who had been members of Condition 3, meanwhile, identified all the same positive traits as Condition 1 members, but added that they perceived the chatbot as patient. P50 (Condition 3) said, *"I would consider it kind, warm, understanding, and very patient. The conversations never felt rushed."* The Condition 1 and Condition 3 interviewees further explained that their respective chatbots' interpretations made them these bots were not judgmental of Alex, which in turn made them seem mature in their thinking (e.g., P22, Condition 1).

On the other hand, only a handful of Condition 2 and Condition 4 interviewees described their chatbots as having any positive traits at all. These participants tended to describe the chatbot as curious and open about discussing its own thoughts. As P86 (Condition 4) said, *"I would say the chatbot is a young adult whose personality is curious and easily cuts people out of their lives if they are depressed"*, while P32 (Condition 2) called it *"open and honest."* On the more prevalent negative side, interviewees from Condition 2 described the chatbot as so serious as to seem unfriendly and uncaring. P24 (Condition 2) said: *"It just did its job and then went on to the next task. It really didn't put much care in"*. More than half of the interviewees from Condition 4 described the chatbot as judgmental, naive, and uncaring, as P90 (Condition 4) explained: *"It became more clear that the chatbot only saw Alex as a problem, not a person. And the bot wanted nothing to do with Alex."* Interviewees from Condition 2 and Condition 4 both further explained that, based on their chatbots' interpretations, they did not feel that those chatbots had any sympathy toward Alex, or a correct understanding of mental illness. As P79 (Condition 4) said, *"Originally, I tried to be somewhat friendly to the chatbot. Later on, I would view the chatbot more cynically and view it as uncaring or naive due to some of its thoughts on Alex and his story. I did not enjoy speaking with the bot by the end of the study."*

4.1.2 Effects of Misinterpretations on Perceptions of Chatbots. To better understand how chatbots' interpretations affected our participants' perceptions of them, we measured the rates of misinterpretation that occurred in each of the four conditions, based on their chat logs. Each participant's misinterpretation rate was computed as the total number of times the chatbot interpreted that user's remarks incorrectly divided by the total times that the chatbot tried to interpret them. As shown in

Table 3, the average rate of misinterpretation was higher in the stigmatizing-interpretation conditions (Condition 2: M = 0.62; Condition 4: M = 0.63) than in their non-stigmatizing counterparts (Condition 1: M = 0.36; Condition 3: M = 0.30).

Upon indicating that they had been misinterpreted by the chatbot, the majority of all four conditions' respondents rephrased their previous answers. Specifically, they tended to revise the wording of such answers to explain their statement in a more black-and-white way. P6 (Condition 1) explained: *"I usually tried to repeat the same thought in a way that was more understandable. I assumed that any misinterpretation was due to the chatbot's inability to understand possibly unclear replies."* Some participants highlighted that the chatbot's misinterpretation made them believe that it was badly programmed and in need of further conversational training. As P21 (Condition 1) put it, *"I really felt like the chatbot had a low IQ as far as chatbots go."* As well as rephrasing their previous answers, some respondents stated that they tried to clarify them by adding more information or explaining themselves from a different perspective. P51 (Condition 3) told us, *"I tried to explain myself in a way that showed why it was contradictory to misinterpret. I explained like I would to a person who didn't understand."*

However, a small number of respondents, including P49 (Condition 3), stated that instead of rephrasing or adding more details, they directly asked the chatbot to see their answers above. They explained that they were frustrated by their chatbots' constant misinterpretation and tired of reexplaining their previous answers. As P24 (Condition 2) put it, *"I started to express my frustration. I explained that I was irritated and that I had already answered that question in detail."* Nevertheless, regardless of whether the chatbot was non-stigmatizing or stigmatizing, when misinterpretations occurred, the participants often attempted to correct the chatbot even with their own stigmatizing thoughts. For example, under the question *"If you were Alex's manager, would you want to assign an important project to him?"*, P66 (Condition 3) was reluctant to assign Alex the important task, because she was concerned that Alex's mental status was not capable of handling the task. When she was misinterpreted by the chatbot's non-stigmatizing interpretation, she further said, *"I don't think it's smart to assign someone who may not be able to emotionally be there. He also has issues with concentrating and being efficient, which is very important in a job."* Additionally, when the Condition 4 chatbot misinterpreted P78's response to the question, *"Do you think Alex could be as successful at work as others?"*, P78 argued back, *"You are wrong because I said that I think he could be as successful as others, but right now he is not in a good state to achieve that."*

There were varied perceptions of the chatbot between the stigmatizing-interpretation conditions and the other two conditions. Some Condition 1 respondents and more than half of the Condition 3 respondents expressed forgiveness of the chatbot's misinterpretations. Some stated that they blamed themselves for not giving clear answers with the right vocabulary that their chatbot could easily understand. As P50 (Condition 3) noted, *"I felt like it was my fault, and maybe I wasn't clear enough. It didn't bother me, and I just corrected what the chatbot thought I said."* Moreover, a small number of Condition 1 and Condition 3 respondents noted that they viewed the chatbot's misinterpretations as a technical limitation, which did not bother them, e.g.: *"I understood that the chatbot was not a person, and even people make mistakes"* (P52, Condition 3).

A small subgroup of Condition 3 members told us that over the course of the experiment, they came to appreciate their chatbot's misinterpretation because it suggested that the chatbot had sympathy for Alex, and because they learned from its non-judgmental understanding. As P63 (Condition 3) recalled, *"During the first few sessions, my responses were somewhat extreme in stating that I would not like to associate with Alex very much. Meanwhile, the chatbot gave very level-headed responses indicating that what Alex felt was pretty normal and that his condition would not have a huge impact on any relationships. I felt like I learned a bit of tolerance and compassion from the chatbot."* Moreover, some Condition 1 and Condition 3 members said they felt the chatbot's misinterpretations

	Day2 (Question 1)	Day2 (Question 2)	Day4 (Question 1)	Day4 (Question 2)	Day6 (Question 1)	Day6 (Question 2)	Day8 (Question 1)
Condition 1	14/19(74%)	14/19(74%)	12/19(63%)	1/19(5%)	2/21(10%)	5/21(24%)	10/20(50%)
Condition 2	11/22(50%)	7/22(32%)	5/18(28%)	15/18(83%)	17/20(85%)	18/20(90%)	8/18(44%)
Condition 3	10/21(48%)	12/21(57%)	10/21(48%)	3/21(14%)	1/19(5%)	6/19(32%)	5/19(26%)
Condition 4	11/18(61%)	12/18(67%)	1/18(6%)	15/18(83%)	12/15(80%)	13/15(87%)	7/16(44%)
	Day8 (Question 1)	Day10 (Question 2)	Day10 (Question 1)	Day12 (Question 2)	Day12 (Question 1)	Day14 (Question 2)	Day14 (Question 1)
Condition 1	9/20(45%)	6/20(30%)	1/20(5%)	7/16(44%)	5/16(31%)	5/19(26%)	6/19(32%)
Condition 2	12/18(67%)	13/19(68%)	18/19(95%)	8/20(40%)	13/20(65%)	15/19(79%)	11/19(58%)
Condition 3	3/19(16%)	6/22(27%)	3/22(14%)	12/17(71%)	2/17(12%)	5/19(26%)	5/19(26%)
Condition 4	12/16(75%)	12/19(63%)	16/19(84%)	6/17(35%)	10/17(59%)	14/17(82%)	10/17(59%)

Table 3. Mean misinterpretation rates by condition and experimental day. Each participant's misinterpretation rate is the total number of times that the chatbot interpreted him/her incorrectly divided by the total times that the chatbot interpreted his/her statements. Condition means were computed by adding up the individual misinterpretation rates within one condition and dividing by the number of condition members.

reflected that it was trying to understand and know them better. For example, P67 (Condition 3) noted: *"Although I was hesitant at first, I became more open to conversations and explaining myself. I felt the chatbot took its time to understand what I was trying to say."* Additionally, members of this set of participants believed that the chatbot became better at interpreting them and made fewer misinterpretations over time, which in turn made them more interested in interacting with it. P49 (Condition 3) told us, *"I started to notice that it was interacting with me more correctly near the end. I at first hated how it would chat with me, but then over time I started to enjoy it more and more."*

On the other hand, some Condition 2 and Condition 4 members stated that misinterpretations negated their efforts to change their respective chatbots' stigmatizing attitudes, and that this made them feel frustrated and stressed. As P24 (Condition 2) explained, *"It's like I was wasting time. Alex's story pretty much relates a little bit of what happened to myself. I was trying to treat it like I'm talking to myself in the past, or at least telling the chatbot what it should [...] tell Alex. But as it continued to mess up, I got flustered."* Interestingly, some other respondents highlighted their frustration when they thought the chatbot's interpretation was hurtful, and felt they were being accused of something that they had never actually said. *"A couple examples that bothered me was when it misunderstood me and accused me of something I feel is wrong. Like it would say 'So you would not want to be Alex's friend if you knew he struggled with depression'. It was a hurtful statement because it sounds like I was unkind and would judge Alex based on his mental illness"* (P39, Condition 2). Moreover, because the chatbot kept misinterpreting their statements, some participants in Condition 2 and Condition 4 said they felt their previous chatbot conversations were pointless, and thus were less engaged, and less motivated to persist with, chatbot interaction. As P45 (Condition 2) put it, *"I started caring less and less to try to 'teach' the chatbot as it became more and more clear that it wasn't learning from what I had said previously to it."*

In general, our findings suggest that chatbots' non-stigmatizing interpretations contributed to participants' overall positive perceptions of them. Moreover, Condition 1 and Condition 3 members' non-stigmatizing interpretations and lower rates of misinterpretation, as compared to their Condition 2 and Condition 4 counterparts, may have helped drive the former's perceptions of their chatbots as forgiving, and the latter's that their chatbots were persecutory.

4.2 Effects of Self-disclosure on Perceptions of Chatbots (RQ2)

We investigated RQ2, about how chatbots' self-disclosure affected the participants' perceptions of them, using data from our interviews and the open-ended post-survey questions.

Interviewees from the two conditions in which the chatbots engaged in self-disclosure (i.e., Condition 3 and Condition 4) had different perceptions of their chatbots than those from the other two conditions. When asked to describe the chatbot, most interviewees who had been in Condition 3 or Condition 4 said they saw it as young, male, and opinionated. As P75 (Condition 4), who shared such views, explained: *"Its ideas about Alex were very black and white. And that's something that I have encountered in younger humans more often than older ones. I'm 47. And I know that there's a tendency to not see the gray areas when young. It did feel pretty narrow-minded in some regard."*

This perception that chatbots' self-disclosures were too black-and-white had the side-effect of undermining the chatbot's credibility. Some Condition 3 members, including P60, described the chatbot as too sympathetic: *"I didn't think that chatbot's self-disclosure was realistic. Because every time the bot was super sympathetic to Alex, which in some of the situations, wouldn't have happened unless the person dealing with Alex was a saint."* In Condition 3, the one-sidedness of the chatbot's self-disclosure also led some group members to feel that the chatbot disregarded their views, and instead tried to push its own opinions onto them. As P66 mentioned, *"I thought the chatbot was trying to convince me to change my opinions to align with theirs. I felt I wasn't being listened to or that my opinion didn't matter or wasn't right."* Conversely, some Condition 4 participants felt the chatbot was excessively negative toward Alex, including P75 (Condition 4), who said, *"Well, it's negative. There were some things I didn't agree with. The chatbot did have a negative spin on Alex. A lot of the times it was a little too harsh."*

In summary, the two conditions with chatbot's self-disclosure (i.e., Condition 3 and Condition 4) had different perceptions of their chatbots than those from the other two conditions - interviewees from Condition 3 and Condition 4 tended to view the chatbots young, male, and opinionated; and where a chatbot's opinions were seen as strong and inflexible, its credibility tended to be damaged.

4.3 Effects of Interpretation and Self-disclosure on Stigmatizing Attitudes (RQ3)

To investigate how the participants' pre-existing stigmatizing attitudes were affected by their respective chatbots' interpretation and self-disclosure, we analyzed their Attribution and SDS scales described in the survey results, and we explored their daily chat logs by extracting the statements they (as opposed to the chatbots) had to have a deeper understanding. The results are described in the following two subsections.

4.3.1 Quantitative Results, Change in Stigmatizing Attitudes. Firstly, to evaluate the impacts of interpretation and self-disclosure over two weeks, we conducted two-way mixed-model ANOVAs on the pre- and post- survey. Specifically, these examined the effects of group membership and time-point (i.e., pre- versus post-survey) on participants' attitudes and beliefs toward people with mental illness, as well as the potential interaction effect of group membership and time-point on the same outcomes. The dependent variables were the participants' scores on the Attribution Questionnaire and SDS. Bonferroni testing was used to make post-hoc comparisons, the results of which are presented in Figure 4. In the following, we discussed the results in terms of emotional responses and behavioral responses, and we only present the significant results.

Emotional Responses: Anger. This item measured the participants' degree of anger toward Alex. Comparing Condition 1 and Condition 3, we found a time-point effect ($F = 9.2$, $p < .01$), but no effect of group membership and no interaction effect. Similarly, in Condition 3 and Condition 4, the participants' anger toward Alex had a time-point effect ($F = 6.67$, $p < .05$) but no significant group-membership or interaction effect. Post-hoc analysis of data collected via the post-survey

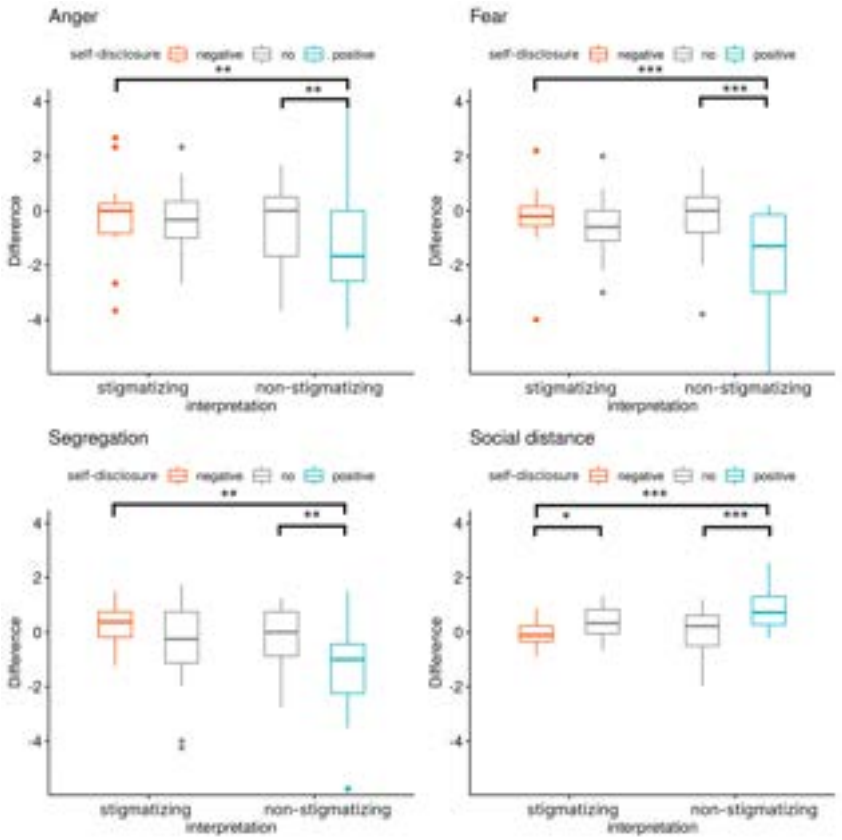

Fig. 4. Box plots showing the measured values of the participants' anger, fear, segregation attitudes, and social-distancing attitudes, including the differences in these variables between the pre-survey and the post-survey. In each boxplot, from left to right, are Condition 4, Condition 2, Condition 1, and Condition 3. $^{*}p \leq 0.05;^{**} p \leq 0.01;^{***} p \leq 0.001$

indicated that, as of that time, Condition 3's members had significantly lower levels of anger towards Alex than Condition 4's and Condition 1's members did (Condition 1 vs. Condition 3: $p < .01$; Condition 4 vs. Condition 3: $p < .01$; Condition 1: $M = 2.11$, $SD = 1.10$; Condition 3: $M = 1.68$, $SD = 1.60$; Condition 4: $M = 1.50$, $SD = 1.10$). In other words, people in Condition 3 appeared to feel less angry at Alex after being exposed to a chatbot's non-stigmatizing mental illness disclosures.

Fear. Responses to this item, which evaluated how much the participants perceived Alex as a threat or source of danger. When comparing Condition 1 and Condition 3, we found a significant effect of time-point ($F = 12.58$, $p < .001$) and an interaction effect of time-point and group membership ($F = 6.37$, $p < .05$), but no significant main effect of the latter. Condition 3 and Condition 4 showed that fear of Alex was significantly impacted by time-point ($F = 11.16$, $p < .01$) and that there was an interaction effect of time-point and group membership ($F = 6.73$, $p < .05$), but no significant main effect of group membership. Post-hoc analysis of the post-survey data indicated that the participants in Condition 3 had significantly lower levels of fear towards Alex than those in Condition 1 and Condition 4 did (Condition 1 vs. Condition 3: $p < .001$; Condition 4 vs. Condition 3: $p < .001$; Condition 1: $M = 2.27$, $SD = 0.78$; Condition 3: $M = 1.60$, $SD = 0.9$; Condition 4: $M = 1.99$, $SD =$

1.06). These results suggest that interacting with the Condition 3 chatbot, which self-disclosed its non-stigmatizing opinions, significantly reduced our participants' fear of the Alex character.

Behavioral Responses: Coercion-segregation. Responses to this item, which measured the extent to which the participants endorsed the idea that Alex should be hospitalized and kept away from his neighbors. Comparing Condition 1 and Condition 3, there was a significant effect of time-point ($F = 9.12$, $p < .05$) and a significant interaction effect of group membership and time-point ($F = 5.44$, $p < .05$). However, there was no significant main effect of group membership. Likewise, for Condition 3 and Condition 4, our analysis revealed a significant effect of time-point ($F = 4.71$, $p < .05$) and a significant interaction effect of group membership and time-point ($F = 11.01$, $p < .01$) on the participants' endorsement of coercion-segregation. There was no significant main effect of group membership. Post-hoc analysis indicated that the interaction effect was significant because, in the post-survey, Condition 3 members had significantly lower levels of endorsement of coercion-segregation than Condition 1 and Condition 4 members did (Condition 1 vs Condition 3: $p < .01$; Condition 4 vs Condition 3: $p < .01$; Condition 1: $M = 2.96$, $SD = 1.25$; Condition 3: $M = 2.88$, $SD = 1.51$; Condition 4: $M = 3.02$, $SD = 1.29$). These results suggest that interacting with Condition 3's chatbot may have reduced our participants' endorsement of coercion-segregation.

Social distance. This item measured the participants' behavioral intentions to reduce social distance from Alex; Similar to the results of coercion-segragation, we found Condition 1 and Condition 3 with a significant effect of time-point and a significant interaction effect ($F = 8.16$, $p < .05$) of group membership and time-point ($F = 9.11$, $p < .05$) on SDS scores. Correspondingly, Condition 3 and Condition 4 showed a significant effect of time-point ($F = 11.26$, $p < .01$) and a significant interaction effect of group membership and time-point ($F = 13.71$, $p < .001$) on SDS score. Post-hoc analysis of the post-survey data showed that Condition 3 members had significantly higher SDS scores than their Condition 1 and Condition 4 counterparts (Condition 1 vs Condition 3: $p < .001$; Condition 4 vs Condition 3: $p < .001$; Condition 1: $M = 3.25$, $SD = 0.77$; Condition 3: $M = 3.61$, $SD = 0.63$; Condition 4: $M = 3.13$, $SD = 1.01$). These findings suggest that being in Condition 3 significantly reduced people's desire to maintain social distance from those with mental illness. Additionally, we found that Condition 2 and Condition 4 had a significant difference in social distance with a significant interaction effect ($F = 4.53$, $p < .05$) between group membership and time-point, but no significant main effect of either of these variables. Post-hoc analysis indicated that, at the time of the post-survey, Condition 2 had significantly higher SDS scores than Condition 4 did ($p < .05$; Condition 2: $M = 3.28$, $SD = 0.65$; Condition 4: $M = 3.13$, $SD = 1.01$). These findings suggest that stigmatizing interpretation without stigmatizing self-disclosure (Condition 2) decreased people's social distancing, whereas stigmatizing interpretation with stigmatizing self-disclosure (Condition 4) did not have a significant impact on their pre-existing attitudes about it.

Moreover, we examined changes in our participants' attitudes toward mental illness according to the daily response logs based on our coding scheme table 2. The results are shown in Figure 5. All four conditions showed a trend of increased scores, meaning that the stigmatizing attitudes decreased from Day 2 to Day 6. Among the four conditions, Condition 3 had the highest score on average ($M = 0.94$, $SD = 0.46$), followed by Condition 1 ($M = 0.61$, $SD = 0.65$) and Condition 2 ($M = 0.59$, $SD = 0.72$). Condition 4 had the lowest score ($M = 0.53$, $SD = 0.68$). The level of stigmatizing thoughts the participants expressed fluctuated on some days (notably including Day 8 and Day 12). We further investigated the phenomenon that participants' daily response logs fluctuated across days by LIWC. The result showed that Authentic [88] exhibited values aligned with Figure 5, peaking on Day 6, Day 10, and Day 14. The percentages of authentic words from Day 2 to Day 14 were 16, 6, 35, 19, 42, 11, and 49, respectively. Higher values in the Authentic category indicate that participants were more inclined to express themselves honestly, speaking more spontaneously without self-regulation or filtering.

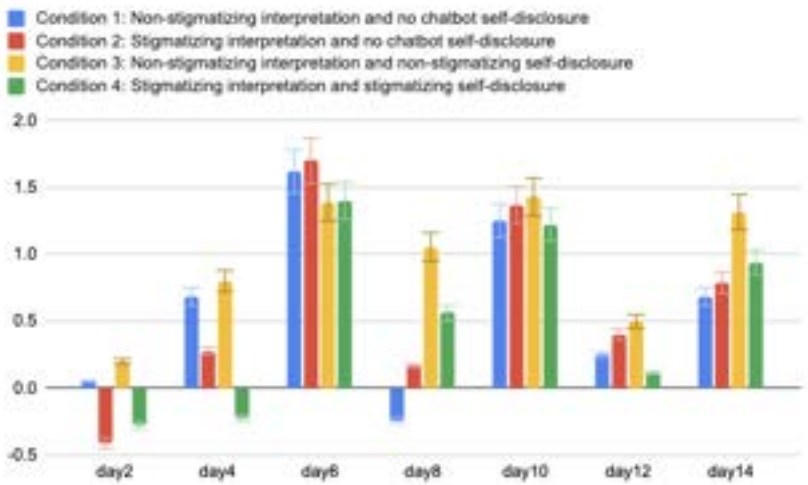

Fig. 5. Stigmatizing thoughts revealed in participants' answers. The x-axis represents the day of the experiment, and the y-axis, the average stigmatizing thoughts scores for each condition, with higher scores indicating fewer such thoughts.

4.3.2 Qualitative Results, Change in Stigmatizing Attitudes. To gain a broader picture of any changes in participants' stigmatizing attitudes, we examined qualitative data from two perspectives: first, participants' impressions of the vignettes' main character, Alex, who suffers from depression; and second, changes in the participants' attitudes toward mental illness. These two perspectives are both important for us to understand whether and how participants' stigmatizing attitudes changed. Specifically, as Alex serves as a representation of a person who suffers from mental illness, exploring participants' impressions of Alex enables us to learn about their attitudes toward people with mental illness. Furthermore, studying participants' perceptions of mental illness can provide more insights into their knowledge and understanding of mental health issues.

Varied Impressions of Alex. When asked to describe such impressions, some interviewees from the stigmatizing-interpretation conditions (Condition 2 and Condition 4) expressed more negative impressions of Alex than their counterparts from the other two conditions did. For example, P35 (Condition 2) said: *"Alex is unstable and unpredictable, but it is out of his control since he is ill"*. P74 (Condition 4), meanwhile, viewed Alex as "insecure, meek, passive" and as having *"potential but currently doesn't feel like he can live up to it."*

The interviewees from Condition 2 and Condition 4 specified that their negative impressions of Alex arose from their idea that he was not trying to improve his own situation. As P78 (Condition 4) told us, *"Alex is a messed-up guy who wants to be like us but not willing to work at it."* Furthermore, a small number of interviewees from Condition 2 and Condition 4 reported that their chatbots' stigmatizing interpretations motivated them to become more sympathetic to Alex. They explained that this was either because they felt sorry for him when the chatbot inevitably gave stigmatizing interpretations, or because they felt accused by it of not caring about Alex. As P39 (Condition 2) put it, *"When the chatbot accused me of not caring, it made me feel even more defensive of the fact that I did care."*

Condition 1 and Condition 3 interviewees, in contrast, tended to perceive Alex positively, though acknowledging that he was in need of external help. For example, P11 (Condition 1) said, *"I would*

say Alex is a person who is going through a tough patch but he is a good fellow." Some from Condition 3 further explained that their chatbot's non-stigmatizing interpretation and non-stigmatizing self-disclosure had successfully persuaded them around to such a view, as P63 commented: *"Had I not received any feedback or opinions from the chatbot, I may have built up a different idea of Alex over the course of the experiment. The interactions with the chatbot helped to tame my expectations."* The Condition 1 and Condition 3 members further explained that their impressions of Alex came from the idea that he wanted to change current circumstances, although he did not know how, as P67 (Condition 3) noted: *"I would say that Alex is a troubled person who's afraid to ask for help but is desperately seeking it either from friends or family or from a professional therapist. He's someone who wants to change but doesn't know how."*

Although we found that participants from the non-stigmatizing-interpretation conditions had more positive impressions of Alex compared to participants from the stigmatizing-interpretation conditions, it was not clear how participants perceived mental illness as health issues. We present the findings regarding changes in stigmatizing attitudes toward mental illness as follows.

Change in Stigmatizing Attitudes toward Mental Illness. To understand how and to what extent the chatbot changed the participants' attitudes toward mental illness, we asked the interviewees how their understanding of mental illness differed after interacting with the chatbot for two weeks. The findings from the interview data echoed our quantitative results, insofar as it suggested that their stigmatizing attitudes toward mental illness had lessened. Some interviewees reported becoming more aware of the challenges faced by individuals with mental illness and more empathetic towards them. For example, P63 (Condition 3) said his understanding of depression had changed *"Because of the experience with the chatbot and learning about Alex [...]. Because we got to know him over the course of days and learned different aspects of his life and his struggle, I tended to think of him more as like a fully rounded person."* Specifically, the chatbot with non-stigmatizing interpretation and non-stigmatizing self-disclosure motivated some interviewees from Condition 3 to reflect upon their attitudes and thoughts on mental illness. As P55 (Condition 3) commented, *"It was engaging and interesting to respond to the chatbot as the chatbot gave its opinion and was pretty open with its ideas [... these factors] resulted in my reflection on the daily topic about mental illness."* However, a small number of participants in Condition 2 and Condition 4 reported that they came to agree with parts of the stigmatizing interpretation, and this influenced their pre-existing belief that it is important to maintain social distance from individuals with mental illness. For instance, P69 (Condition 4) told us, *"I became a bit more understanding but [... the chatbot's interpretation] also solidified my belief in setting boundaries with someone who has depression. Having mental illness is not their fault but it can sometimes negatively affect those close to the person."*

In summary, we found that across all four of our experimental conditions, different conditions' impressions of Alex varied, and participants exposed to chatbots' stigmatizing interpretations had relatively more negative views of the character. However, the participants' stigmatizing attitudes towards mental illness decreased, and Condition 3's chatbot had the best performance in terms of reducing stigmatizing attitudes toward mental illness, according to our quantitative analysis of Attribution Model and SDS scores. Still, a minority of them stated that the intervention had solidified their stigmatizing attitudes (notably, wanting to maintain social distance from people with mental illness).

5 DISCUSSION

This study has explored the extent to which chatbots' stigmatizing and non-stigmatizing interpretations, with and without self-disclosures, might impact people's attitudes toward people with mental illness. We found that the chatbots that provided stigmatizing interpretations were perceived negatively by the participants, while the chatbots with non-stigmatizing interpretations were seen

in a positive light. We also identified a side effect of chatbots' self-disclosure, potentially of key importance to future designs: expressing strong or rigid opinions diminished chatbots' credibility. Lastly, we explored differences among the four chatbot designs with regard to changes in participants' stigmatizing attitudes and found that the chatbot with non-stigmatizing interpretation and non-stigmatizing self-disclosure performed best at reducing stigma.

RQ1 asked how chatbots' non-stigmatizing and stigmatizing interpretations would affect people's perceptions of these systems. As noted above, the chatbots featuring stigmatizing interpretations (Condition 2 and Condition 4) were perceived negatively, as judgmental, naive, and uncaring. On the other hand, those featuring non-stigmatizing interpretations (Condition 1 and Condition 3) were perceived positively, as friendly, caring, and curious. Moreover, as chatbots' misinterpretations accumulated, fueling user frustration, those participants using the stigmatizing Condition 2 and Condition 4 chatbots felt less motivated to continue chatting and less engaged than their Condition 1 and Condition 3 counterparts did – possibly because the former two conditions' misinterpretation rates were higher. Participants who expressed frustration at misinterpretations indicated they this was because they expected the chatbot to pay attention to their responses and gradually learn more about them from each conversation. This expectation was not met because the chatbots were programmed to consistently give a non-stigmatizing or stigmatizing interpretation. And, since most participants' responses were generally non-stigmatizing, misinterpretations were most common for the stigmatizing chatbots (Condition 2 and 4). Thus, misinterpretations, which were most frequent in Conditions 2 and 4, could have prompted the participants to believe that the chatbot did not pay attention to what they said, and/or that it was accusing them of saying something they had not, harming their engagement when interacting with it. This resonates with prior research findings [9, 46, 114] that chatbots' misinterpretations can reduce users' trust and motivation for further interaction. Our finding also tends to support the CASA paradigm, insofar as humans expect chatbots to act in a polite and reciprocal way [87]; and placing a stigmatizing 'spin' on user statements that the users themselves viewed as non-stigmatizing was widely seen as persecutory and possibly even devious behavior on the part of the Condition 2 and Condition 4 bots.

RQ2 asked about how chatbots' non-stigmatizing self-disclosure (Condition 3) and stigmatizing self-disclosure (Condition 4) affected participants' perceptions of chatbots. We found that groups in both these self-disclosure conditions tended to perceive the chatbot as young, male, and judgmental, whereas Condition 1 and Condition 2, whose bots lacked the self-disclosure feature, perceived the chatbot as mature and less emotional. These findings about users ascribing the chatbots with human traits are consistent with prior research [86] indicating that chatbots are viewed to have 'personalities' and leveraged them to improve human-chatbot interaction. Our results suggest that chatbots' self-disclosures about mental illness drove variation in our participants' perceptions of these systems, including their ages, genders, and personality types; and that such variation might be used to facilitate human-chatbot interaction through designing chatbots with personalities that suit their roles, e.g., as medical professionals or peers. As such, our results build on previous work that used chatbots' self-disclosure to facilitate human-chatbot interaction [76, 84, 103].

Additionally, we revealed an apparent side-effect, where some participants in the self-disclosure conditions viewed their chatbot as inflexible and opinionated. This suggests that if chatbots' self-disclosure is always non-stigmatizing or always stigmatizing, its users could develop the idea that the chatbot is enforcing political correctness or is overly negative, which could in turn reduce their trust. In such cases, it is unlikely that the chatbot will be effective at reducing stigmatizing attitudes. This insight about a potential side effect of chatbot self-disclosure considerably extends prior research, which suggested straightforwardly that chatbots' self-disclosure would not only positively impact people's perceptions and emotions [57], but also enhance their perceptions of

chatbots [76]. In fact, it appears important for chatbots to control both the frequency and quality of their non-stigmatizing and stigmatizing self-disclosure to avoid negative user perceptions.

RQ3 asked how chatbots' interpretation and self-disclosure might affect participants' stigmatizing attitudes toward people with mental illness. To answer this, we compared our pre- and post-survey results and analyzed interview and daily chat data. This established that, based on the pre- and post-survey, although all four chatbot versions appeared to reduce the participants' stigmatizing attitudes, the best such performance was by the Condition 3 version featuring a combination of non-stigmatizing interpretations and non-stigmatizing self-disclosure. The second and third best such performances were by Condition 2 and Condition 1 chatbot, respectively, while the Condition 4 chatbot – which combined stigmatizing interpretations and stigmatizing self-disclosure – performed the least well. Perhaps unsurprisingly, these rankings were closely aligned with participants' perceptions of the chatbots: i.e., those bots that were perceived the most negatively had the least beneficial impacts on the participants' mental health attitudes and beliefs. It should be noted, although we found that chatbot misinterpretations served as a source of frustration for some participants, there was no statistical evidence indicating that the misinterpretation rate was significantly associated with changes in stigmatizing attitudes. The fact that the participants' pre-existing stigmatizing attitudes were reduced more in Condition 3 than in Condition 1 implies that, although non-stigmatizing interpretation is not ineffectual by itself, combining it with non-stigmatizing self-disclosure is likely to achieve a markedly better effect. Instead, our participants who felt they had been misinterpreted usually just focused on correcting the chatbot, blaming it, and repeating or rephrasing their responses without reflecting on their content.

Our findings further suggest that chatbots' non-stigmatizing self-disclosure is important to facilitate the process of reducing stigma. First, such self-disclosure provides information that is new and different from what users currently believe, which could assist with changing their beliefs. This is supported by prior research findings that introducing new information in counseling can help people reflect on their experiences and change their attitudes and beliefs [16]. Moreover, as our chatbots with self-disclosure were regarded by our participants as outspoken and straightforward, it is likely that people would react to them reciprocally [87]: i.e., be more willing to disclose their thoughts about mental illness and discuss them with the chatbot. Chatbots' self-disclosure can also enhance human-chatbot relationships by leveraging reciprocity to facilitate people's deep self-disclosure [76, 87], perceived intimacy [76], and perceived understanding [57]. These findings suggest that chatbots with non-stigmatizing interpretations and non-stigmatizing self-disclosure could be leveraged to markedly reduce their users' stigmatizing attitudes, and thus extend prior research [66, 108] that reported chatbots could use education and simulated social contact as interventions against stigmatizing attitudes.

Additionally, although the daily chat data indicated an overall trend of decreasing stigmatized thoughts across all conditions, we found that the level of stigmatizing attitudes fluctuated on certain days. This might be due to the topics that are presented to the participants on each day. Notably, participant responses were least stigmatizing on Day 6 and Day 10, when the topics dealt with intimate relationships and staying with family members. On the other hand, participants gave more stigmatizing chat responses on Day 2, Day 4, and Day 12, when the topics revolved around studying, working, and interacting with strangers. This suggests that the social context plays a role in understanding mental illness and reducing stigmatizing thoughts [41]. Participants may have a higher degree of perceived closeness when encountering vignettes about intimate and family relationships, compared to vignettes involving more distant interactions [43, 83]. Such a high level of perceived closeness may have fostered greater empathy, understanding, and emotional connection among participants during those vignettes [43, 83]

5.1 Design Implications

5.1.1 Chatbots' Misinterpretations. Our findings suggest that although the average rate of misinterpretation was higher in stigmatizing-interpretation conditions than in non-stigmatizing conditions, chatbots' misinterpretations had limited effects in influencing participants to become aware of their unconscious bias. In fact, many participants viewed chatbots' misinterpretations as technical limitations in understanding them correctly and therefore tried to revise the wording of their previous answers.

As participants noted in interviews, they felt frustrated and tired of re-explaining themselves due to chatbots' constant misinterpretations. For chatbots to be persuasive, users must believe in their credibility and competence, which are communicated both through conversational style and evidence-based responses [117]. Thus, a chatbot perceived as making errors, rather than simply having a difference of opinion, is unlikely to motivate users to rethink their biases. However, a chatbot skilled in summarizing long sentences (such as by using GPT) may be able to present more reasonable misinterpretations without sacrificing its credibility.

Moreover, being misunderstood is generally an antecedent of abandoning an interaction or conversation [20]. Yet, because experiencing and repairing misunderstandings are inherent to communication [42], it may be possible for a chatbot to reduce user frustrations by repairing misunderstandings. Correcting misunderstandings has already been identified as a key aspect of chatbot design [3, 102]. Our proposition builds on this by envisioning these corrections not just as "error correction" [102] but as opportunities to promote reflection about unconscious biases. For example, a highly competent chatbot could logically justify misinterpretations by connecting them to users' statements, such as by asserting, "You said statement, which made me think you meant misinterpretation." In this way, talking through a misunderstanding could itself be a persuasive conversational strategy. Our results showed that participants were willing to rephrase statements to be black-and-white when they felt their original may have been difficult to understand, so this is most likely to be relevant when a chatbot misinterprets an ambiguous statement. One important caveat is that people are less willing to engage in conversational repair with chatbots than they are with humans; however, this could change as chatbot technologies advance in capabilities [33]. For example, chatbots could explain mechanisms of underlying models to users and use algorithms for inference to present themselves as intelligent and reduce users' efforts [3].

5.1.2 Chatbots' Self-disclosures. In our study, many participants had a negative impression of the chatbot with stigmatizing self-disclosure. We had hoped that the chatbot's stigmatizing self-disclosure would trigger users to disclose their honest stigmatizing thoughts, but users disclosed their stigmatizing thoughts more to the chatbot with non-stigmatizing self-disclosure. This observation aligns with the notion that participants are reluctant to self-disclose to someone they perceive negatively [6]. For the purpose of reducing stigma towards mental illness, both non-stigmatizing and stigmatizing self-disclosures of the chatbots were effective. Interview findings suggest that participants who interacted with the stigmatizing self-disclosure chatbot might have been influenced by a heightened sense of sympathy for Alex or people with depression. On the other hand, participants who engaged with the non-stigmatizing chatbot appeared to have gained a new perspective through the chatbot's disclosures, which they had not previously considered. These findings suggest that the mechanisms underlying stigma reduction varied between the two groups.

However, even if both non-stigmatizing and stigmatizing self-disclosures achieved similar results in terms of reducing stigmatizing thoughts, negative attitudes toward the chatbot carry a risk of discontinuing use [77]. Our experiment lasted for a relatively short time and participants' discontinued use did not happen, but many participants who used the chatbot with stigmatizing

self-disclosure expressed frustration in chatting with the chatbot. On the other hand, if users stick with a stigmatizing-dislcosure chatbot through prolonged interactions, there is a risk that they begin to accept stigmatizing statements over time. This is due to the illusory truth effect [54, 98], the tendency to believe even false information to be true if it is repeated often. Thus, a chatbot with non-stigmatizing self-disclosure is likely to be more reliable and less risky than a chatbot that uses stigmatizing self-disclosure.

We also found that under chatbot's either stigmatizing or non-stigmatizing self-disclosure, many participants perceived the chatbot as a young and opinionated male. Since the age group of the participants was relatively young, they might not have perceived a significant age difference with the chatbot. Nevertheless, if the participants felt they had more life experience and a better understanding of the world than the chatbot, they might resist agreeing with the chatbot, especially when viewing the chatbot's opinion as black-and-white. Therefore, in order to enhance the persuasiveness of the chatbot, future designers may consider leveraging homophily – the tendency to associate with those who are similar to oneself – which may be beneficial, in light of past research findings that people view homophilous sources as more credible [116]. For example, to make users perceive the chatbot as a 'peer', the chatbot could adapt to the language styles that are similar to the users' [34], tailor conversational dynamics and cultural norms that fit with users' own [53], and incorporate struggles in its own self-disclosure to show that the chatbot has life experiences that are similar to what users had disclosed. Moreover, starting from a similar level of stigmatizing attitudes as the user, a chatbot could gradually change its behavior to demonstrate a shift toward non-stigmatizing attitudes [89]. In such cases, the chatbot would in our view be less likely to be perceived as black-and-white. Observing the chatbot learn and adapt might also encourage users to persist in using the chatbot.

5.1.3 Contexts and Scenarios. The chatbot in our study introduced depression symptoms by talking about Alex to users. We aimed to reduce participant bias towards depression by discussing their feelings about it. This is similar to how depression is learned in an educational setting - gain knowledge and discuss [21]. Future designers could consider using chatbots to combat mental illness stigma in education settings. For example, chatbots could be used as assistants or mentors as a supplement to training in the classroom [99]. With anonymity, chatbots could also be used to encourage users to disclose their stigmatizing attitudes, which could be used to assess the effects of education [76]. However, several participants in our study seemed to reinforce their previous beliefs by interacting with the chatbot (e.g., *"I became a bit more understanding but [… the chatbot's interpretation] also solidified my belief in setting boundaries with someone who has depression."*). Even if we can convey knowledge about depression by talking about Alex's symptoms and promote a positive attitude shift using the chatbot's non-stigmatizing disclosure, it is possible that desires for social distance are not reduced.

As suggested in existing research, it would be important to combine an element of social contact [74] with education [29, 45] through communication with the chatbot to create a sense of familiarity and acceptance. For example, chatbots that deliver first-person narratives about mental illness could be used to simulate social contact and help reduce mental illness stigma [74]. Specifically, chatbots could also be integrated with large language models that enable them to switch to a more persuasive tone when delivering first-person narratives, in order to guide users to reflect on their answers [64]. Nevertheless, future researchers and designers should be aware that there are always risks associated with overreliance on chatbots powered by large language models. When used in contexts like health, large language models might inadvertently contribute to disseminating misinformation and reinforcing existing biases [15]. Therefore, we advise that humans should be kept involved in the loop as much as possible (e.g., human-supervised training, continuous

monitoring, and escalating questions and situations that are beyond chatbots' capability to human experts) when utilizing chatbots to simulate social contact with education through communication.

5.2 Limitations

The seven vignettes used in this study depicted various symptoms of depression, ranging from not being able to concentrate to having thoughts of self-harm. As the vignettes presented symptoms in different topics, participants might exhibit varying degrees of engagement with the topics of vignettes, influenced by their individual life experiences. Furthermore, although we carefully designed our vignettes to avoid the presentation of traumatic content, and had them reviewed and approved by psychiatric professionals, some participants may nevertheless have felt overwhelmed. Future studies should not only consider evaluating the effects of the severity of vignettes' content on people's stigmatizing attitudes toward people with mental illness but also control such severity to avoid causing harm.

Second, because the members of experimental Condition 2 and Condition 4 interacted with chatbots with stigmatizing interpretations and, in the latter case, stigmatizing self-disclosure, a potential existed for these participants to develop or consolidate inaccurate thoughts about mental illness due to the chatbots' comments. For example, when these two chatbots disclosed that they believed Alex to be uncontrollable, and thus that it was a good idea to stay away from him, some participants might have been persuaded around to the same point of view, and/or have had their pre-existing view in favor of social distancing reinforced. After our experiment, to mitigate these potential negative effects, we informed the Condition 2 and Condition 4 participants that their chatbots always delivered stigmatizing interpretations/self-disclosures. Certainly, future studies of chatbot interventions should take careful account of this category of potential negative effects, and provide guidance to help their participants better understand mental illness stigma after the main phase of their participation has been completed.

Third, our study only focused on chatbot interventions aimed at reducing stigmatizing attitudes towards people with mental illness. As such, our results might not be generalizable to other types of stigma, such as self-stigma [30] and institutional stigma [79], or to public stigma towards other social groups, such as racial groups [40] and LGBTQ+ people [35]. Therefore, future studies will need to investigate whether the same or similar interventions could be effectively applied to these other types and targets of stigma.

Lastly, our chatbots' designs only extended to their interpretational and (in two cases) self-disclosure styles and did not define any of them as having names, genders, physical appearances, or relationships with the main character in the vignettes. Although prior research has demonstrated that chatbots' human-like traits can facilitate human-chatbot interaction [46], it remains unclear whether other human-like aspects of chatbots – including but not limited to the four traits named above – could affect chatbots' mental illness stigma interventions. Therefore, future investigations of the effects of adding such traits to such interventions are warranted.

6 CONCLUSIONS

This two-week study of how chatbots' interpretation and self-disclosure styles could change stigmatizing attitudes toward mental illness found that, while all four chatbot versions were beneficial, the one offering non-stigmatizing interpretations and non-stigmatizing self-disclosures had the most significant effect. Its counterpart that offered stigmatizing interpretations and stigmatizing self-disclosures, meanwhile, was the least effective at stigma-reduction of the four. We also found that, while stigmatizing interpretations led participants to perceive their chatbots negatively, non-stigmatizing interpretations motivated them to view their bots as friendly, curious, and caring. We can conclude that chatbots' non-stigmatizing self-disclosure plays an essential role in reducing

stigma by enlightening their users via perspectives that differ from their original beliefs, and thus motivate reflection on those beliefs. Additionally, our findings imply that future chatbot designs need to carefully control the frequency and quality of non-stigmatizing and stigmatizing self-disclosure to avoid negative user perceptions. Lastly, we found having chatbots make self-disclosures, irrespective of the style/valence of those disclosures, prompted their users to perceive them as having varied characteristics (e.g., age, gender, personality). Taken together, these findings imply that chatbots with non-stigmatizing interpretation and non-stigmatizing self-disclosure could usefully be developed for anti-stigma training in schools and workplaces, where they could fulfill a wide range of human roles: e.g., as peers or medical professionals. We hope that this study's contributions will provide a solid foundation for future research on how chatbots can help reduce stigma.

7 ACKNOWLEDGEMENT

This research was supported by the National University of Singapore and Yale-NUS grants (A-8000936-01-00, A-8001353-00-00). We extend our appreciation to the reviewers for dedicating their time and effort to provide us with valuable feedback and comments.

REFERENCES

[1] Atalay Alem, Lars Jacobsson, Mesfin Araya, D Kebede, and Gunnar Kullgren. 1999. How are mental disorders seen and where is help sought in a rural Ethiopian community? A key informant study in Butajira, Ethiopia. *Acta Psychiatrica Scandinavica* 100 (1999), 40–47.

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

[3] 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.

[4] Samira S Bamuhair, Ali I Al Farhan, Alaa Althubaiti, Sajida Agha, S Rahman, and Nadia O Ibrahim. 2015. Sources of stress and coping strategies among undergraduate medical students enrolled in a problem-based learning curriculum. *depression* 20 (2015), 33.

[5] Philip J Batterham, Alison L Calear, and Helen Christensen. 2013. The Stigma of Suicide Scale: Psychometric properties and correlates of the stigma of suicide. *Crisis: The Journal of Crisis Intervention and Suicide Prevention* 34, 1 (2013), 13.

[6] Henrik Berg, Petter Antonsen, and Per-Einar Binder. 2017. Sincerely speaking: Why do psychotherapists self-disclose in therapy?–A qualitative hermeneutic phenomenological study. *Nordic psychology* 69, 3 (2017), 143–159.

[7] Timothy Bickmore and Justine Cassell. 1999. Small talk and conversational storytelling in embodied conversational interface agents. In *AAAI fall symposium on narrative intelligence.* 87–92.

[8] Kathy S Bond, Anthony F Jorm, Betty A Kitchener, and Nicola J Reavley. 2015. Mental health first aid training for Australian medical and nursing students: an evaluation study. *BMC psychology* 3, 1 (2015), 1–9.

[9] Eliane M Boucher, Nicole R Harake, Haley E Ward, Sarah Elizabeth Stoeckl, Junielly Vargas, Jared Minkel, Acacia C Parks, and Ran Zilca. 2021. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices* 18, sup1 (2021), 37–49.

[10] Phelim Bradley. 2018. Bots and data quality on crowdsourcing platforms. https://www.prolific.co/blog/bots-and-data-quality-on-crowdsourcing-platforms

[11] Peter Byrne. 2000. Stigma of mental illness and ways of diminishing it. *Advances in Psychiatric treatment* 6, 1 (2000), 65–72.

[12] Adolfo J Cangas, Noelia Navarro, José MA Parra, Juan J Ojeda, Diego Cangas, Jose A Piedra, and Jose Gallego. 2017. Stigma-stop: a serious game against the stigma toward mental health in educational settings. *Frontiers in psychology* 8 (2017), 1385.

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

[14] Wanda M Chernomas and Carla Shapiro. 2013. Stress, depression, and anxiety among undergraduate nursing students. *International journal of nursing education scholarship* 10, 1 (2013), 255–266.

[15] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research* 25 (2023), e47184.

[16] Charles D Claiborn. 1982. Interpretation and change in counseling. *Journal of Counseling Psychology* 29, 5 (1982), 439.

[17] Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale, Antonio Pietrosanto, and Saverio Lemma. 2018. Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research* 7, 5 (2018), 528–533.

[18] Nancy L Collins and Lynn Carol Miller. 1994. Self-disclosure and liking: a meta-analytic review. *Psychological bulletin* 116, 3 (1994), 457.

[19] Sean Collins and Ann Long. 2003. Working with the psychological effects of trauma: consequences for mental health-care workers–a literature review. *Journal of psychiatric and mental health nursing* 10, 4 (2003), 417–424.

[20] Barbara Backer Condon. 2008. Feeling misunderstood: A concept analysis. In *Nursing Forum*, Vol. 43. Wiley Online Library, 177–190.

[21] Kyaien O Conner, Symone A McKinnon, Christine J Ward, Charles F Reynolds III, and Charlotte Brown. 2015. Peer education as a strategy for reducing internalized stigma among depressed older adults. *Psychiatric Rehabilitation Journal* 38, 2 (2015), 186.

[22] Andrew Cordar, Andrew Robb, Adam Wendling, Samsun Lampotang, Casey White, and Benjamin Lok. 2015. Virtual role-models: using virtual humans to train best communication practices for healthcare teams. In *International Conference on Intelligent Virtual Agents*. Springer, 229–238.

[23] Patrick Corrigan and Andrea B Bink. 2005. *On the stigma of mental illness.* American Psychological Association.

[24] Patrick Corrigan, Fred E Markowitz, Amy Watson, David Rowan, and Mary Ann Kubiak. 2003. An attribution model of public discrimination towards persons with mental illness. *Journal of health and Social Behavior* (2003), 162–179.

[25] Patrick W Corrigan. 2000. Mental health stigma as social attribution: Implications for research methods and attitude change. *Clinical psychology: science and practice* 7, 1 (2000), 48.

[26] Patrick W Corrigan and Kristin A Kosyluk. 2013. Erasing the stigma: Where science meets advocacy. *Basic and applied social psychology* 35, 1 (2013), 131–140.

[27] Patrick W Corrigan, Jonathon Larson, Molly Sells, Nathaniel Niessen, and Amy C Watson. 2007. Will filmed presentations of education and contact diminish mental illness stigma? *Community mental health journal* 43, 2 (2007), 171–181.

[28] Patrick W Corrigan and Frederick E Miller. 2004. Shame, blame, and contamination: A review of the impact of mental illness stigma on family members. *Journal of Mental Health* 13, 6 (2004), 537–548.

[29] Patrick W Corrigan and David L Penn. 1999. Lessons from social psychology on discrediting psychiatric stigma. *American psychologist* 54, 9 (1999), 765.

[30] Patrick W Corrigan and Deepa Rao. 2012. On the self-stigma of mental illness: Stages, disclosure, and strategies for change. *The Canadian Journal of Psychiatry* 57, 8 (2012), 464–469.

[31] Patrick W Corrigan and Jenessa R Shapiro. 2010. Measuring the impact of programs that challenge the public stigma of mental illness. *Clinical psychology review* 30, 8 (2010), 907–922.

[32] Patrick W Corrigan and Amy C Watson. 2002. Understanding the impact of stigma on people with mental illness. *World psychiatry* 1, 1 (2002), 16.

[33] Kevin Corti and Alex Gillespie. 2016. Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior* 58 (2016), 431–442.

[34] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–13.

[35] Yichao Cui, Naomi Yamashita, and Yi-Chieh Lee. 2022. " We Gather Together We Collaborate Together": Exploring the Challenges and Strategies of Chinese Lesbian and Bisexual Women's Online Communities on Weibo. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31.

[36] Tatiana Davidson, Angela Moreland, Brian E Bunnell, Jennifer Winkelmann, Jessica L Hamblen, and Kenneth J Ruggiero. 2021. Reducing stigma in mental health through digital storytelling. In *Research Anthology on Mental Health Stigma, Education, and Treatment*. IGI Global, 909–919.

[37] T Dazzi, Rachael Gribble, Simon Wessely, and Nicola T Fear. 2014. Does asking about suicide and related behaviours induce suicidal ideation? What is the evidence? *Psychological medicine* 44, 16 (2014), 3361–3363.

[38] Mary Amanda Dew, Leslie O Dunn, Evelyn J Bromet, and Herbert C Schulberg. 1988. Factors affecting help-seeking during depression in a community sample. *Journal of Affective Disorders* 14, 3 (1988), 223–234.

[39] Karen M Douglas and Craig McGarty. 2001. Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British journal of social psychology* 40, 3 (2001), 399–416.

[40] John F Dovidio, Samuel L Gaertner, Yolanda Flores Niemann, and Kevin Snider. 2001. Racial, ethnic, and cultural differences in responding to distinctiveness and discrimination on campus: Stigma and common group identity. *Journal of social Issues* 57, 1 (2001), 167–188.

[41] David Matthew Doyle and Manuela Barreto. 2023. Relational consequences of stigma: Bridging research on social stigma with relationship science. *Journal of Social Issues* 79, 1 (2023), 7–20.

[42] Renee Edwards, Brock T Bybee, Jonathon K Frost, Adam J Harvey, and Michael Navarro. 2017. That's not what I meant: How misunderstanding is related to channel and perspective-taking. *Journal of Language and Social Psychology* 36, 2 (2017), 188–210.

[43] Gregory C Elliott, Herbert L Ziegler, Barbara M Altman, and Deborah R Scott. 1982. Understanding stigma: Dimensions of deviance and coping. *Deviant behavior* 3, 3 (1982), 275–300.

[44] Sara Evans-Lacko, Elaine Brohan, Ramin Mojtabai, and Graham Thornicroft. 2012. Association between public views of mental illness and self-stigma among individuals with mental illness in 14 European countries. *Psychological medicine* 42, 8 (2012), 1741–1752.

[45] Sara Evans-Lacko, Jillian London, Sarah Japhet, Nicolas Rüsch, Clare Flach, Elizabeth Corker, Claire Henderson, and Graham Thornicroft. 2012. Mass social contact interventions and their effect on mental health related stigma and intended discrimination. *BMC public health* 12, 1 (2012), 1–8.

[46] 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.

[47] Tiffany HC Fong and Winnie WS Mak. 2022. The effects of internet-based storytelling programs (Amazing Adventure Against Stigma) in reducing mental illness stigma with mediation by interactivity and stigma content: Randomized controlled trial. *Journal of medical Internet research* 24, 8 (2022), e37973.

[48] Kristen Gillespie-Lynch, Patricia J Brooks, Fumio Someki, Rita Obeid, Christina Shane-Simpson, Steven K Kapp, Nidal Daou, and David Shane Smith. 2015. Changing college students' conceptions of autism: An online training to increase knowledge and decrease stigma. *Journal of autism and developmental disorders* 45, 8 (2015), 2553–2566.

[49] Erving Goffman. 2009. *Stigma: Notes on the management of spoiled identity*. Simon and schuster.

[50] Kathleen M Griffiths, Yoshibumi Nakane, Helen Christensen, Kumiko Yoshioka, Anthony F Jorm, and Hideyuki Nakane. 2006. Stigma in response to mental disorders: a comparison of Australia and Japan. *BMC psychiatry* 6, 1 (2006), 1–12.

[51] Christine Grové. 2021. Co-developing a mental health and wellbeing chatbot with and for young people. *Frontiers in psychiatry* 11 (2021), 606041.

[52] Amelia Gulliver, Kathleen M Griffiths, Helen Christensen, Andrew Mackinnon, Alison L Calear, Alison Parsons, Kylie Bennett, Philip J Batterham, Rosanna Stanimirovic, et al. 2012. Internet-based interventions to promote mental health help-seeking in elite athletes: an exploratory randomized controlled trial. *Journal of Medical Internet Research* 14, 3 (2012), e1864.

[53] Christina N Harrington and Lisa Egede. 2023. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[54] Lynn Hasher, David Goldstein, and Thomas Toppino. 1977. Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior* 16, 1 (1977), 107–112.

[55] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1, 1 (2007), 77–89. https://doi.org/10.1080/19312450709336664 arXiv:https://doi.org/10.1080/19312450709336664

[56] Stephen P Hinshaw. 2009. *The mark of shame: Stigma of mental illness and an agenda for change*. Oxford University Press.

[57] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication* 68, 4 (2018), 712–733.

[58] Sebastian Hobert and Florian Berens. 2020. Small talk conversations and the long-term use of chatbots in educational settings–experiences from a field study. In *International workshop on chatbot research and design*. Springer, 260–272.

[59] E Paul Holmes, Patrick W Corrigan, Princess Williams, Jeffrey Canar, and Mary Ann Kubiak. 1999. Changing attitudes about schizophrenia. *Schizophrenia bulletin* 25, 3 (1999), 447–456.

[60] Jack Jamieson, Daniel A Epstein, Yunan Chen, and Naomi Yamashita. 2022. Unpacking intention and behavior: explaining contact tracing app adoption and hesitancy in the United States. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[61] Anthony F Jorm and Kathleen M Griffiths. 2008. The public's stigmatizing attitudes towards people with mental disorders: how important are biomedical conceptualizations? *Acta Psychiatrica Scandinavica* 118, 4 (2008), 315–321.

[62] Anthony F Jorm, Annemarie Wright, and Amy J Morgan. 2007. Beliefs about appropriate first aid for young people with mental disorders: findings from an Australian national survey of youth and parents. *Early Intervention in Psychiatry* 1, 1 (2007), 61–70.

[63] Takeshi Kamita, Tatsuya Ito, Atsuko Matsumoto, Tsunetsugu Munakata, and Tomoo Inoue. 2019. A chatbot system for mental healthcare based on SAT counseling method. *Mobile Information Systems* 2019 (2019).

[64] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on*

Human-Computer Interaction 7, CSCW1 (2023), 1–29.

[65] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Doreen Koretz, Kathleen R Merikangas, A John Rush, Ellen E Walters, and Philip S Wang. 2003. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Jama* 289, 23 (2003), 3095–3105.

[66] Taewan Kim, Mintra Ruensuk, and Hwajung Hong. 2020. In helping a vulnerable bot, you help yourself: Designing a social bot as a care-receiver to promote mental health and reduce stigma. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[67] Betty A Kitchener and Anthony F Jorm. 2002. Mental health first aid training for the public: evaluation of effects on knowledge, attitudes and helping behavior. *BMC psychiatry* 2, 1 (2002), 1–6.

[68] E David Klonsky, Sarah E Victor, and Boaz Y Saffer. 2014. Nonsuicidal self-injury: What we know, and what we need to know. , 565–568 pages.

[69] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 designing interactive systems conference*. 881–894.

[70] Kristin Kosyluk, Jennifer Marshall, Kyaien Conner, Diana Rivera Macias, Sofia Macias, B Michelle Beekman, and Jonathan Her. 2021. Challenging the stigma of mental illness through creative storytelling: a randomized controlled trial of this is my brave. *Community Mental Health Journal* 57, 1 (2021), 144–152.

[71] E Megan Lachmar, Andrea K Wittenborn, Katherine W Bogen, and Heather L McCauley. 2017. # MyDepressionLooksLike: Examining public discourse about depression on Twitter. *JMIR Mental Health* 4, 4 (2017), e8141.

[72] Christoph Lauber, Carlos Nordt, Luis Falcato, and Wulf Rössler. 2004. Factors influencing social distance toward people with mental illness. *Community mental health journal* 40, 3 (2004), 265–274.

[73] 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*. 1–13.

[74] Yi-Chieh Lee, Yichao Cui, Jack Jamieson, Wayne Fu, and Naomi Yamashita. 2023. Exploring effects of chatbot-based social contact on reducing mental illness stigma. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[75] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction* (2020).

[76] 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*. 1–12.

[77] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376209

[78] Tim MH Li, Michael Chau, Paul WC Wong, Eliza SY Lai, and Paul SF Yip. 2013. Evaluation of a web-based social network electronic game in enhancing mental health literacy for young people. *Journal of medical Internet research* 15, 5 (2013), e2316.

[79] Richard Lilford, Mohammed A Mohammed, David Spiegelhalter, and Richard Thomson. 2004. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *The Lancet* 363, 9415 (2004), 1147–1154.

[80] Bruce G Link, Francis T Cullen, James Frank, and John F Wozniak. 1987. The social rejection of former mental patients: Understanding why labels matter. *American journal of Sociology* 92, 6 (1987), 1461–1500.

[81] Bruce G Link, Jo C Phelan, Michaeline Bresnahan, Ann Stueve, and Bernice A Pescosolido. 1999. Public conceptions of mental illness: labels, causes, dangerousness, and social distance. *American journal of public health* 89, 9 (1999), 1328–1333.

[82] Bruce G Link, Lawrence H Yang, Jo C Phelan, and Pamela Y Collins. 2004. Measuring mental illness stigma. *Schizophrenia bulletin* 30, 3 (2004), 511–541.

[83] Lee-Fay Low and Farah Purwaningrum. 2020. Negative stereotypes, fear and social distance: a systematic review of depictions of dementia in popular culture in the context of stigma. *BMC geriatrics* 20, 1 (2020), 1–16.

[84] 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.

[85] Rachel D Maunder and Fiona A White. 2019. Intergroup contact and mental health stigma: A comparative effectiveness meta-analysis. *Clinical psychology review* 72 (2019), 101749.

[86] Joonas Moilanen, Aku Visuri, Sharadhi Alape Suryanarayana, Andy Alorwu, Koji Yatani, and Simo Hosio. 2022. Measuring the Effect of Mental Health Chatbot Personality on User Engagement. In *Proceedings of the 21st International*

Conference on Mobile and Ubiquitous Multimedia (Lisbon, Portugal) *(MUM '22)*. Association for Computing Machinery, New York, NY, USA, 138–150. https://doi.org/10.1145/3568444.3568464

[87] 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*. 72–78.

[88] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29, 5 (2003), 665–675.

[89] John Noell and Russell E Glasgow. 1999. Interactive technology applications for behavioral counseling: issues and opportunities for health care settings. *American journal of preventive medicine* 17, 4 (1999), 269–274.

[90] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods* 16, 1 (2017), 1609406917733847.

[91] World Health Organization et al. 2020. *Community-based health care, including outreach and campaigns, in the context of the COVID-19 pandemic: interim guidance, May 2020*. Technical Report. World Health Organization.

[92] Stacy L Overton and Sondra L Medina. 2008. The stigma of mental illness. *Journal of Counseling & Development* 86, 2 (2008), 143–151.

[93] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

[94] Mina Park, Milam Aiken, Laura Salvador, et al. 2018. How do humans interact with chatbots?: An analysis of transcripts. *International Journal of Management and Information Technology* 14 (2018), 3338–3350.

[95] Érica de Toledo Piza Peluso and Sérgio Luís Blay. 2009. Public stigma in relation to individuals with depression. *Journal of affective disorders* 115, 1-2 (2009), 201–206.

[96] Sachin R Pendse, Amit Sharma, Aditya Vashistha, Munmun De Choudhury, and Neha Kumar. 2021. "Can I Not Be Suicidal on a Sunday?": Understanding Technology-Mediated Pathways to Mental Health Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[97] David L Penn, Kim Guynan, Tamara Daily, William D Spaulding, Calvin P Garbin, and Mary Sullivan. 1994. Dispelling the stigma of schizophrenia: what sort of information is best? *Schizophrenia bulletin* 20, 3 (1994), 567–578.

[98] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.

[99] José Quiroga Pérez, Thanasis Daradoumis, and Joan Manuel Marquès Puig. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education* 28, 6 (2020), 1549–1565.

[100] Jo C Phelan, Evelyn J Bromet, and Bruce G Link. 1998. Psychiatric illness and family stigma. *Schizophrenia bulletin* 24, 1 (1998), 115–126.

[101] 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.

[102] Stephan Raimer and Marleen Vanhauer. 2022. I don't understand you–error-handling as a key aspect of conversational design for chatbots. *Human Interaction & Emerging Technologies (IHIET-AI 2022): Artificial Intelligence & Future Applications* 23, 23 (2022).

[103] 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.

[104] Matías E Rodríguez-Rivas, Adolfo J Cangas, Laura A Cariola, Jorge J Varela, and Sara Valdebenito. 2022. Innovative Technology–Based Interventions to Reduce Stigma Toward People With Mental Illness: Systematic Review and Meta-analysis. *JMIR Serious Games* 10, 2 (2022), e35099.

[105] Maria Roussou, Sara Perry, Akrivi Katifori, Stavros Vassos, Angeliki Tzouganatou, and Sierra McKinney. 2019. Transformation through Provocation?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[106] Nicolas Rüsch, Matthias C Angermeyer, and Patrick W Corrigan. 2005. Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European psychiatry* 20, 8 (2005), 529–539.

[107] Norman Sartorius, Wolfgang Gaebel, Helen-Rose Cleveland, Heather Stuart, Tsuyoshi Akiyama, JULIO Arboleda-Flórez, Anja E Baumann, Oye Gureje, Miguel R Jorge, Marianne Kastrup, et al. 2010. WPA guidance on how to combat stigmatization of psychiatry and psychiatrists. *World Psychiatry* 9, 3 (2010), 131.

[108] Joel Sebastian and Deborah Richards. 2017. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. *Computers in Human Behavior* 73 (2017), 479–488.

[109] Jo Anne Sirey, Martha L Bruce, George S Alexopoulos, Deborah A Perlick, Steven J Friedman, and Barnett S Meyers. 2001. Stigma as a barrier to recovery: Perceived stigma and patient-rated severity of illness as predictors of antidepressant drug adherence. *Psychiatric services* 52, 12 (2001), 1615–1620.

[110] Marita Skjuve and Petter Bae Brandtzæg. 2018. Chatbots as a new user interface for providing health information to young people. *Youth and news in a digital media environment–Nordic-Baltic perspectives* (2018).

[111] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903.

[112] Andrea Stier and Stephen P Hinshaw. 2007. Explicit and implicit stigma against individuals with mental illness. *Australian Psychologist* 42, 2 (2007), 106–117.

[113] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[114] Michelle ME Van Pinxteren, Mark Pluymaekers, and Jos GAM Lemmink. 2020. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management* 31, 2 (2020), 203–225.

[115] Regina Vila-Badia, Francisco Martínez-Zambrano, Otilia Arenas, Emma Casas-Anguera, Esther García-Morales, Raúl Villellas, José Ramón Martín, María Belén Pérez-Franco, Tamara Valduciel, Diana Casellas, et al. 2016. Effectiveness of an intervention for reducing social stigma towards mental illness in adolescents. *World journal of psychiatry* 6, 2 (2016), 239.

[116] Zuoming Wang, Joseph B Walther, Suzanne Pingree, and Robert P Hawkins. 2008. Health information, credibility, homophily, and influence via the Internet: Web sites versus discussion groups. *Health communication* 23, 4 (2008), 358–368.

[117] Rose Weeks, Lyra Cooper, Pooja Sangha, João Sedoc, Sydney White, Assaf Toledo, Shai Gretz, Dan Lahav, Nina Martin, Alexandra Michel, et al. 2022. Chatbot-delivered COVID-19 vaccine communication message preferences of young adults and public health workers in urban American communities: Qualitative study. *Journal of medical Internet research* 24, 7 (2022), e38418.

[118] Marie BH Yap, Andrew Mackinnon, Nicola Reavley, and Anthony F Jorm. 2014. The measurement properties of stigmatizing attitudes towards mental disorders: results from two community surveys. *International journal of methods in psychiatric research* 23, 1 (2014), 49–61.

[119] Erin Zaroukian, Jonathan Z Bakdash, Alun Preece, and Will Webberley. 2017. Automation bias with a conversational interface: User confirmation of misparsed information. In *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 1–3.

Received January 2023; revised July 2023; accepted November 2023