Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond

YUE YOU, College of Information Sciences and Technology, Pennsylvania State University
CHUN-HUA TSAI, College of Information Science and Technology, University of Nebraska Omaha
YAO LI, School of Modeling, Simulation, and Training, University of Central Florida
FENGLONG MA, College of Information Sciences and Technology, Pennsylvania State University
CHRISTOPHER HERON, Family and Community Medicine, Penn State Health
XINNING GUI, College of Information Sciences and Technology, Pennsylvania State University

Chatbot-based symptom checker (CSC) apps have become increasingly popular in healthcare. These apps engage users in human-like conversations and offer possible medical diagnoses. The conversational design of these apps can significantly impact user perceptions and experiences, and may influence medical decisions users make and the medical care they receive. However, the effects of the conversational design of CSCs remain understudied, and there is a need to investigate and enhance users' interactions with CSCs. In this article, we conducted a two-stage exploratory study using a human-centered design methodology. We first conducted a qualitative interview study to identify key user needs in engaging with CSCs. We then performed an experimental study to investigate potential CSC conversational design solutions based on the results from the interview study. We identified that emotional support, explanations of medical information, and efficiency were important factors for users in their interactions with CSCs. We also demonstrated that emotional support and explanations could affect user perceptions and experiences, and they are context-dependent. Based on these findings, we offer design implications for CSC conversations to improve the user experience and health-related decision-making.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI;**

Additional Key Words and Phrases: Symptom checkers, chatbots, consumer-facing health technology, conversation design

ACM Reference format:
Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond. *ACM Trans. Comput.-Hum. Interact.* 30, 4, Article 64 (September 2023), 44 pages.
https://doi.org/10.1145/3589959

This work is supported by Penn State College of IST's seed grant, No. 150000004308 INTR..
Authors' addresses: Y. You, F. Ma, and X. Gui, College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802; emails: {yxy340, fenglong, xinninggui}@psu.edu; C.-H. Tsai, College of Information Science and Technology, University of Nebraska Omaha, Omaha, NE 68182; email: chunhuatsai@unomaha.edu; Y. Li, School of Modeling, Simulation, and Training, University of Central Florida, Orlando, FL 32816; email: yao.li@ucf.edu; C. Heron, Family and Community Medicine, Penn State Health, University Park, PA 16802; email: cheron@pennstatehealth.psu.edu.

1073-0516/2023/09-ART64 $15.00
https://doi.org/10.1145/3589959

1 INTRODUCTION

Chatbot-based symptom checker (**CSC**) apps have become increasingly integrated into daily health care. In recent years, popular CSCs, such as Ada, Your.MD, and K Health, have been downloaded from major mobile app stores tens of millions of times [100]. CSCs provide possible diagnoses by communicating with users through *conversations* that are mediated by interactive interfaces [103] as users are guided through questions. The design engages the users in human-like conversations by utilizing the techniques of natural language processing and machine learning [36, 78]. For example, K Health employs conversational phrases, such as "*Thanks for telling me about your...*," in its probing questions.

The way CSC conversations are designed can influence user interpretations of the technology's responses, affecting their future interactions [118] with it. Conversational design is defined as "a design language based on human conversation" [2]. Generally speaking, CSC conversations can be designed to offer potential diagnoses [164], explanations in response to questions (e.g., a CSC app called Buoy), and examples of similar instances of illness (e.g., K Health). A suitable conversational style, i.e., "the way things are said" [143] can impact user perception and overall experiences significantly. In the case of CSCs, the conversational design of CSCs may influence the future medical decisions users make and the medical care they receive. Users can put their health at risk if they blindly trust the medical recommendations given by CSCs and if they lack sufficient medical knowledge 3485 [50]. Therefore, it is necessary to consider user perceptions of CSCs' current conversational design and how it can be improved.

Despite the recent proliferation of CSCs, the user experience of the conversational design of CSCs remains understudied. There have been limited previous studies, including our own previous study [150], on the topic, which reported issues with information overload [46], emphasized the value of explanations [150], and illustrated the importance of humanistic qualities [117]. Most prior studies on chatbot-based health apps have mainly focused on chatbots designed for mental health intervention or coaching [84, 88, 108, 115]. Different from these chatbots that aim at reducing users' depression and anxiety [55, 125] or promote a healthy lifestyle [43, 45], the focus of CSCs is to provide potential medical diagnoses. The design goals and conversational patterns of CSCs are thus distinct from those of mental health chatbots [14, 71], which may promote different expectations and needs [44]. As prior **Human-Computer Interaction** (**HCI**) studies have argued, healthcare systems should include human-like features [29, 107, 155] and improve their transparency and explainability [62, 101, 127, 140, 154], begging the question of how to tailor and improve CSC conversational design in consideration of these qualities.

To investigate and enhance the conversational design in users' interaction with CSCs, we performed a two-stage exploratory analysis using a human-centered design methodology. That is, we first conducted a qualitative interview study (**Study 1**) with the goal of identifying key user needs in engaging in conversations with existing CSCs. Informed by the results from the interview study, we then performed an experimental study (**Study 2**) with the goal of investigating potential CSC conversational design solutions by testing Study 1's design implications.

Specifically, **Study 1** (i.e., the interview study) revealed that people desire emotional support and doctor-like probing mode, explanations of medical information, and efficiency (i.e., acquiring diagnostic results more quickly) in their interactions with the CSC app. Then, based on Study 1, in **Study 2** (i.e., the experimental study), we designed a chatbot-based symptom checker with four different conversational styles (a baseline CSC, one with emotional support, one with explanation, and one with both) and one efficient view which allows users to hide the emotional support/explanation information presented for efficiency. We collected quantitative and qualitative feedback from our participants to explore the effects of these design solutions in a lab-controlled

study. We additionally measured user experience regarding transparency, usability, human likeness, and emotional support. Study 2 demonstrated that emotional support and explanations could affect user trust, transparency, emotional support, human likeness, information overload, and privacy, and that participants had different requirements for emotional support and explanations in different situations. Our results stressed the importance of affording user control in different contexts and based on different user traits (e.g., different conversation stages and different individual personalities).

We make multiple contributions to the HCI community. First, we investigate the conversational design of CSCs from a real-world user's perspective which is important but understudied; Second, we propose design solutions based on user needs and conduct a follow-up design study to evaluate the design effectiveness. Such a mixed-method study not only demonstrates research rigor [23], but also offers a more comprehensive portrayal of the studied phenomenon [111] and a set of cross-validated design implications; Third, our mixed research methods not only contribute to the conversational design of CSCs, but also have the potential to contribute to the user experience research of conversational agents, as existing methods assessing the user experience after the interaction with conversational agents are still in their infancy [15, 47]. Last, we demonstrate the importance of user control and the necessity of considering different situations when offering emotional support and explanations; and lastly, we discuss design implications for CSC conversations to improve the user experience and health-related decision-making.

2 BACKGROUND

With the digital healthcare industry and Information Technology developing at a rapid pace, countless healthcare chatbots have appeared in mobile app stores, addressing everything from dermatology to cardiology, endocrinology, neurology, and nutrition [20, 61, 67]. One kind of popular healthcare chatbot is the CSC that users use for self-diagnosis and self-triage. It is a consumer-facing tool that uses a conversation interface, which can assist with self-triage and offer users potential diagnostic results [135]. The Google Play Store has more than 200 CSC-related apps listed [156], and in 2019, CSCs were used more than 100 million times in the United States alone [1]. Some most popular and well-known CSCs include Ada, K Health, and Healthily. These apps are popular with relatively high ratings in mobile app stores (above 4.0/5.0). According to previous studies, this kind of self-diagnosis tools can potentially support self-care [81], reduce patients' anxiety [93], and provide education on when to seek non-emergency or emergency care [126].

When seeking diagnoses through a CSC, users communicate with the app's embedded chatbot by inputting their symptoms in response to the app's probing questions. Consultations with a CSC usually go through several stages [5]: first, users are asked to input their symptoms; next, the app solicits more details of the symptoms, as well as the user's medical history and other areas of related information; finally, the CSC gives the user a consultation report and closes the conversation with a reminder to be cautious, such as "*Please remember, this service is only for general information purposes and isn't a personal or medical diagnosis*" (Ada). During the three stages, some CSCs use humanizing language in their interactions. For example, before the consultation starts, some (e.g., K Health, and Ada) greet users with phrases, like "*Hi, XXX [the user's name]*", and during the consultation, some give users feedback in a friendly manner, such as "*Thanks for telling me about your...*" or "*That's good*" (K Health).

3 LITERATURE REVIEW

A chatbot is a computer program that utilizes a natural language user interface to interact with users [98, 103]. Chatbots can be classified into rule-based and AI-based. Rule-based bots use pre-defined or handcrafted algorithms to identify users' responses while AI-based bots utilize

machine-learning algorithms to analyze the responses and generate feedback based on their learning models. [64, 96]. Compared to traditional WIMP (Windows, Icons, Menus, and Pointers) interfaces, chatbots provide a familiar way for users to disclose information about themselves and are flexible enough to accommodate various user requests rather than having a fixed path for users to follow [17, 92, 146, 148]. In the field of HCI, researchers have explored the effects of different conversational styles on user engagement [120] and the data quality of the users' responses [76]. They've categorized conversational styles into "high-involvement" and "high-considerateness," or put another way, "causal" and "formal."

Due to their advantages, a set of existing healthcare applications (e.g., symptom checkers) uses chatbots to communicate with users. The conversational and interaction designs of CSCs can impact users' trust [165] and how they engage in their healthcare in the future [118]. In this section, we first report on the existing literature on the conversational design of the existing CSC apps. Then, we report on two widely discussed topics related to general chatbot design: human-like features and explainability.

3.1 Usability Issues in the Conversational Design of CSC Apps

A CSC app is a type of healthcare chatbot that can assist with self-triage and offers users potential diagnoses [135]. Several studies have pointed out the problems found in its conversational design, stressing the issues of information overload [46, 117] (e.g., too many pieces of information presented at once) and the importance of humanistic qualities (e.g., empathy) [117].

The offerings of medical explanations have also been explored. One study [140] presented an interactive dialogue interface to let users control the delivery of explanations. The authors used an experimental study to explore the users' perceptions of three types of symptom checker prototypes: interactive dialogue with explanations, static explanations, and no explanations. They found that the interactive dialogue led to a higher level of transparency and trust. In another study, a commercial CSC (Buoy Health [4]) incorporated explanations into its design, allowing users to acquire the reason for some of the questions by clicking a "*Why am I being asked this?*" button.

However, the research on CSC conversational design is still nascent. The existing research mainly focuses on the probing questions with few studies investigating what other features, such as human-like responses, users desire and how they can be provided. Also, under-explored is how users perceive CSC conversational design in everyday practice. Our work is motivated to fill these gaps.

3.2 Human-like Features in the Conversational Design of Chatbots

With the increased use of chatbots, their conversational design has become a primary research strand. A key body of research has debated whether the chatbot conversation should be human-like with "the attribution of human qualities, including consciousness, intentions, and emotions" [134]). Past studies have shown discrepancies in terms of how a chatbot should communicate and behave.

Some researchers have highlighted the strengths of providing the chatbot's conversational design with human-like characteristics, arguing that a chatbot should ask intelligent questions [70], demonstrate humor [70], and show empathy [24, 52, 155], as well as be emotionally supportive of users [114]. One group of researchers illustrated how the involvement of human-likeness and the use of empathy can positively affect users' attitudes toward chatbots, increasing how much they like the system [105, 107], and their satisfaction with it [145], as well as positively influencing their relationship with the technology [11], and improving their trust in chatbots overall [21, 26, 97]. For example, Chen et al. [26] demonstrated that human-like communication strategies could enhance users' trust in the system designed to support migrants' social integration, while other studies found that it increased the chatbots' effectiveness, reducing users' verbal abuse [29, 30],

improving users' moods [94, 128], motivating behavioral change [22, 110], and ameliorating the persuasiveness of the system [41]. According to one study, a human-like conversational agent can shape users' decisions toward sustainable mobility behavior [41].

At the same time, researchers have had concerns. They've found that human-like conversational design can lead users to have too high and unrealistic expectations of chatbots [133, 142] and cause frustration when the users realize it cannot respond like an actual human. Contradicting previous studies, Donkelaar [42] found no relation between human-like cues and users' satisfaction. In fact, too much emotional support was found to make the chatbot seem deceptive or invasive [53].

In the healthcare domain, however, most research has proposed that healthcare chatbots be human-like and convey empathy. Empathy can be defined as "an affective response more appropriate to the condition of another than to one's own" [112], and includes the understanding and the acknowledgment of the other's feelings [79]. Past studies have demonstrated that users desire empathetic interactions with chatbots used for chronic pain self-management [60] and health advice [36, 89]. Such interactions can mitigate the negative impact of social exclusion on users' moods [38]. This is especially true of mental health-related chatbots, in which empathy and human-likeness helped meet users' emotional needs [73, 118]. One exploratory study, however, suggested that less human-like chatbots might outperform the human-like ones in sensitive disclosures [130].

While prior studies have mainly considered the "whether or not" question for the human-likeness of chatbots' conversational design, they've failed to fully consider how and when empathy should be provided during user interactions. In addition, few studies have investigated the human-like design of CSCs. One study compared an empathic chatbot with an advice-only chatbot for self-diagnosis [36] and found that the empathic chatbot was preferred by most participants. Another study highlighted how the design of CSC apps can—positively or negatively—affect users' perceptions of the CSC's authority. Neither of these studies, however, took into account the timing of human-like responses. Therefore, our study aims at addressing this.

3.3 Explainablility of Chatbots and XAI

In recent years, the importance of explainability has been noted by a body of HCI researchers investigating conversational interfaces [72, 101, 161]. Explainability is defined as "generating decisions in which one of the criteria taken into account during the computation is how well a human could understand the decisions in the given context [102]" and has been shown to increase transparency, accountability, and trust of intelligent agents [123]. Researchers have proposed that chatbots should provide explanations and relevant information in various situations, such as psychotherapy [132], group decision supports [136], conversational breakdown repairment [12], and education [160]. Explanations could help users better establish their mental models of chatbots [82], improve their perception of system control [157], and better facilitate their satisfaction and user experience [80, 113, 121, 144]. For example, [157] explored the effect of different justification styles for a movie-recommendation chatbot. The authors demonstrated how telling users why the chatbot delivered their recommendations could help them interpret the recommender's underlying rationale, thus increasing the users' perceived transparency, control, and trust. In the healthcare domain, it is also critical to ensure that the health information provided by chatbots is transparent and explainable [62, 127, 138, 154].

To provide explanations, previous researchers suggest that a chatbot may describe its information source [74], working mechanisms, competencies, and/or limitations [31]. Previous studies have discussed different explanation styles presented through the interface, such as visual example-based explanations [72, 82], context explanations [69, 69], causal explanations [27, 106], and social explanations [116]. Mimicking human conversations and considering target user groups have also been emphasized [123].

Explainability is also related to the very idea of explainable AI (XAI), which facilitates the human understanding and trust of AI technologies [74]. Most notably, previous research in XAI has stressed user-centered explanations. A set of researchers have suggested designing user-centered explanations to explore user needs [86, 87, 153]. Explanations should also be tailored to different users' specific needs as well as their interests and refer to their mental states [9, 131]. That is, the detailed levels of the explanation [57], its complexity, ways of presentation, and prioritization of information should be personalized to different users [131]. When it comes to chatbots, incorporating XAI could boost user trust in the chatbots [138, 154] by explaining their information provisions and decisions [71].

Despite the importance of explainability, there is a limited body of work that has explored it as part of the multifaceted CSC user experience. Two studies have pointed out that CSCs should provide explanations and information for data sources [46], probing questions [66], prediction accuracy [46], and its decision model, (i.e., how potential diagnoses were generated) [46, 66]. For example, Hwang et al. [66] pointed out that Ada should explain the types of data used for the model training and how the diagnostic recommendations were generated. Another study highlighted how explanations should be tailored to the specific needs of different users by considering them and their existing level of medical knowledge [159]. While these studies have offered insight into what types of explanations CSCs should provide, they did not consider how the explanations should be offered to users. Deeply rooting the conversational explanation in XAI [129, 139], our works aim at studying the nuanced needs of users for explanations in CSCs.

4 RESEARCH DESIGN

We completed two consecutive studies. In Study 1, we conducted 25 interviews with the first group of participants to empirically explore the user experience of the existing CSC conversational design. Building on the user needs uncovered in Study 1, for Study 2, we designed four types of conversational styles and looked at how they influenced user experience by recruiting a different group of participants (N = 34). We obtained IRB approval from our university prior to the studies.

5 STUDY 1: UNCOVERING USERS' NEEDS FOR EXISTING CSCS' CONVERSATIONAL DESIGN

This interview study aimed to investigate users' CSC conversational needs and experiences and generate design implications for the follow-up experimental study (Study 2).

5.1 Methods

To investigate how users perceive the conversational design of CSCs, we conducted 25 semi-structured interviews with CSC users in the United States. Since many CSC apps are widely used in the U.S. (e.g., K Health, Ada), this was the ideal setting for our research. We recruited participants through our institution's Studyfinder and through social media. We used a survey to screen participants for the following criteria: (1) over the age of 18; (2) spoke English or Mandarin (due to the research team's language skills); (3) previous use of a CSC to self-diagnosis; with (4) the last time of use less than one year ago. We selected 25 eligible participants in total. They were aged from 20 to 55 and had a diversity of occupations, including landscape designer, student, and engineer. Most participants were in their 20s and 30s. The participants had used various CSC apps, including Ada, K health, and your.MD (renamed Healthily). Most participants had used more than one CSC app, and had consulted them for a variety of symptoms, such as headache, backache, and acne. For a more detailed summary of the participants' demographic information, please see Table A.2 in the Appendix).

We started our interview with general questions, such as what CSC app(s) the participant had used and why. We then focused on their experiences with the conversational design of the CSCs, asking more specific questions, including how the participant felt about chatting with the app and what kind of conversational style they preferred. The interview questions are shown in Appendix D. Each interview took between 30 minutes to 1 hour.

We then manually transcribed all the interviews (Mandarin transcripts were also translated into English for further analysis and reporting) and adopted a thematic analysis through an inductive approach [16]. Three researchers participated in the whole analysis process. We first immersed ourselves in the transcripts by reading back and forth to familiarize the whole dataset. Next, each of us generated an individual list of codes focusing on the whole dataset using the software DeDoose. Then, we compared and combined these codes after several rounds of discussions, resulting in a final set of over 40 initial codes. After acquiring the initial codes, we analyzed the relationships between codes and searched for candidate themes and sub-themes. During this process, we also refined these generated themes using rounds of discussions to ensure internal homogeneity and external heterogeneity. We finally acquired a thematic map with three themes: (1) the need for emotional support and doctor-like probing mode, (2) the desire for explanations during interactions, and (3) the need for efficiency.

Before each interview, we obtained the participant's informed consent. Participants could withdraw from the interview and decline to answer any questions at any time. When reporting our findings, we use U1, U2, and so on to denote different participants.

5.2 Findings

Our analysis reflected how our participants experienced the conversational design of CSCs. Through the interviews and thematic analysis, we found the following three themes about user needs for the CSCs' conversational design: first, our participants (N = 6) desired emotional support and a doctor-like probing mode, e.g., human-like greetings and caring language; second, they (N = 16) wanted explanations for both the probing questions and the diagnostic recommendations; and third, some participants (N = 8) sought efficiency in their interactions with the app. The detailed interview results about these three themes are presented in the next few sections.

5.2.1 Desiring Human-like Emotional Support and Doctor-like Probing Mode. We found that our participants desired emotional support, such as an opening with a greeting and an ending with caring language and treatment suggestions (i.e., what the user should do to ease the symptoms). These parts of the CSC interaction were usually regarded as non-instrumental by our participants and needed to have a comforting effect. For example, U10 told us:

"*The chatbot's way of asking questions is just like usual [casual] conversations. It is not very mechanical and strict. Like K Health would ask me, "how do you feel today?" The beginning is simple, like "how are you feeling today", then it said, "let's begin". It makes people feel [like] there's some empathy. It's like, "we can figure this out together.*"" (U10)

Here, U10 expressed that by using human-like greetings, which were similar to those found in daily conversation, the app was showing caring and empathy. For U10, the greetings made the conversation less mechanical and strict.

After acquiring diagnostic results, some participants desired further suggestions and comfort, which is not a function of the current CSCs. For example, U11 wanted a response similar to a concerned doctor's instructions at the end of the diagnosis that would give her a warm feeling. She commented:

"*For example, doctors usually tell you that you should have a rest or something like that, but the app didn't say this kind of thing well. For example, when I'm uncomfortable, the app should tell me at the end, [something] like "keep warm, and don't catch a cold".*" (U11)

In addition to emotional support, participants also sought a doctor's style of professionalism. Participants tended to compare the sentence structure and the sequence of questions asking about symptoms in CSC's probing process with that of a doctor.

First, our participants complained that some CSCs (e.g., K Health) repeatedly used the same sentence structure when asking questions, which they considered stiff and mechanical. For example, U8 criticized how K Health always used the similar phrases and sentence structure to solicit symptom information, which was much different from a doctor's clinical probing, since doctors would not use same phrases to ask all questions. This made the app feel more like a rating system collecting her data and calculating the probability of potential diseases than a doctor diagnosing her. She remarked:

"*K health asked questions with the same style, like this, "Do any of these terms describe your headache? How uncomfortable does your headache make you feel?" I think that's why I think it's more like a machine instead of a human because it's like a rating system. It's like, the questions are collecting your data, instead of talking with you just like a human would. I think a doctor wouldn't ask you these questions, like "How uncomfortable does your headache make you feel?" and have you choose from "mild, severe..."*." (U8)

Second, similar to the probing process in clinical practice, the CSC apps (e.g., K Health and Ada) in our study used probing questions in a progressive order. Users paid attention to the sequence of probing questions, which also impacted their perceptions of the conversational design. Participants were satisfied with the progressive sequence that was presented in some CSC apps (e.g., K Health and Zuoshou Doctor). One reason is that users thought the progressive sequence, i.e., from general to specific, was in line with how doctors sequence probing questions, as illustrated by U3's comments:

"*I think the current conversation is good, because it's like asking progressive questions. It starts with very simple, general questions and then asks for more detail. It's like human. So, you feel like you're texting with a real doctor. Basically, I think all questions they ask are like human-like. I think they are comfortable to me.*" (U3)

In this example, U3 expressed a preference for the progression from general to specific because this kind of sequence is similar to that found in clinical practice, making the conversation human-like and more comfortable for him.

5.2.2 Expecting Explanations. In our findings, we discovered that some participants repeatedly sought explanations throughout their interactions with the app. For example, participants wanted to know why certain questions were asked and were curious about how the technology generated the diagnostic results as these questions and diagnostic recommendations were seemly irrational based on their knowledge.

In the middle of their interaction with the CSC, some participants were confused by the questions being asked about their symptoms. For example, U17 wanted to check her headache using two CSCs: Ada and K Health. Since K Health asked a question (i.e., "did your headache come slowly or quickly?") that, from her perspective, was seemingly irrelevant, she wondered why K Health asked this question; she believed that K Health should have asked for information about specific pain areas as Ada did. She explained:

"*K Health is more like a machine and it asked you "did your headache come slowly or quickly?" That's a really weird question. It didn't ask me anything about the pain areas like Ada did. Ada asked me if I have a sore temple and if I have pain around my eye socket, so those are the symptoms related to my problem. So I really think it [K Health] should have explained why it asked such questions.*" (U17)

In addition, some participants craved an explanation for why certain questions were asked again and again. For example, U20 wanted to know why Ada repeatedly asked questions about her neck

since she had already replied that she did not have neck issues. This made her feel like the app did not record her previous answers. She said:

"*So for Ada, especially, I must have answered the same question, like five or six times. And they kept asking about [the same] things. One example was they asked me if I had any issues with my neck and I said no. And then throughout the next five minutes, they asked me, like five or six questions, about my neck. And I just kept having to put no and I felt like, "did [the app] accept that I put no?" I was wondering why they were still asking, "Do you have spots on your neck?" or "Do you have pain in your neck?""*" (U20)

Furthermore, our participants were concerned about a large number of questions. The example below shows the importance of explaining why CSC apps have to ask so many questions.

"*K health, its conversation is long, with many questions, and it feels very long. So it's not very enjoyable. Instead, every time Ada asks questions, there's a help button for you to see what's the meaning behind the question. I like that feature, which is helpful.*" (U5)

Here, U5 complained about the large number of K Health questions that had no explanations provided. She appreciated the help button offered by Ada, which gave her the information she wanted.

Our participants also desired explanations at the end of their conversation about what symptoms they input that led to the final diagnostic results. For example, U22 suggested that the CSC tell users how the app generated such recommendations, including what specific symptoms led to its suggestions. She told us:

"*I kind of wish that after [the conversation] they would give me the results, they would say, "and this is how we arrived at the results. It would be nice that at the end of this, [it said] it could be this, could be that..." it said this is the process which our symptom checker app used in order to determine the diagnoses, because of these three or four or five reasons, or symptoms. I just wish the symptom checker, after they give you the results, would tell you this is our formula for giving for that.*" (U22)

Similarly, U16 shared her preference for explanations on why K Health diagnosed her with premenstrual syndrome, since it was different from her own guess of pregnancy. Interestingly, she believed that the probing questions asking about menstruation, nausea, and vomiting were in line with her symptoms, but the result turned out to be inaccurate. Thus, she wanted explanations for the diagnostic result she received. As she put it,

"*I tested twice, it said I had premenstrual syndrome. It didn't give me the right result, which is that I thought I might be pregnant. I feel it is so stupid. If you tell a doctor that your menstruation has not come, and you have recently suffered from nausea and vomiting, I think the doctor would say you are [maybe] pregnant. And then the app [K Health] asked me some questions [about menstruation, nausea and vomiting] that seemed to be in the right direction. I think the questions are okay, but the final result was wrong, I guess. It would be better if it could explain the result.*" (U16)

5.2.3 Wishing for Efficiency in the Probing Process. Our participants also wanted the CSCs' probing process to be concise and efficient, with a formal conversational style.

Some participants stressed that anthromorphasizing the CSC's conversational style at the probing stage could influence users' trust negatively and make the conversations verbose. They also felt that inappropriate human-like responses to the users' input could interrupt the whole diagnostic process. As U9 said,

"*"Thank you and we are generating the result page for you, just a moment." That's perfectly ok. But when you are asking a question, don't give the patient any encouragement. Because you know, first of all, it will just slow down their question-answering process, which should be a very efficient process. Second, if you give the user encouragement, like, "great job," it makes them feel "oh, I'm okay". This is*

a really wrong illusion and then on the result page, like in my result page, it shows me several severe diseases that makes the encouragement look bad and weird." (U9)

In this example, U9 thought that human-like encouragement should be excluded from the conversation for two reasons: first, such encouragement could interrupt the flow of the conversation, compromising the diagnostic efficiency; second, such encouragement could give the illusion that the user was not ill, but the app did offer her several potential diagnoses toward the end. For U9, the misguided human-like responses negatively affected her trust in the app.

Participants also felt that the human-likeness at the probing stage led to long-winded conversations from the user's perspective. For example, U8 complained:

"*I see it says things like, "let me ask you a few questions about your headache." I think it tries to use human-like language but maybe because there are so many questions, like around 20, it makes me want to quit.*" (U8)

From U8's perspective, the human-like language, along with the large number of the questions asking about her symptoms, producing a very long conversation, thereby making her too impatient to complete the whole probing process.

5.3 Summary

Through our interview study, we discovered the users' CSC conversational design needs: emotional support as well as doctor-like probing mode, explanations, and efficiency. Based on these three factors, we designed a CSC with four possible conversation styles as well as a toggle button on its interface to meet users' needs.

First, our findings revealed that the users desired emotional support, such as greetings, at the beginning of the conversation (e.g., "*how do you feel today?*") and further treatment suggestions (e.g., "*you may exercise more, eat more fruits*") at the end. Therefore, it could be user-friendly to provide emotional support for users during their conversations. Thus, we designed our first conversation style based on the idea of providing emotional support, through greetings (e.g., "*How are you feeling today?*"), assurances (e.g., "*don't worry!*"), and further care suggestions (i.e., what the user should do after receiving potential diagnostic results). Our findings reported that users disliked the repetitive sentence structure of the CSCs' probing questions. Instead, they desired flexible sentence structures (i.e., doctor-like probing mode). Therefore, we employed diverse question structures. For example, asking "*do you have XXX [symptom]?*" and "*may I ask if you have XXX [symptom]?*"

Second, our findings showed that the users wanted explanations for the questions during the conversation and for the diagnostic recommendations at the end. Therefore, our second conversational style provided the rationale for each probing question and the potential diagnoses prompted by the app. This is to say, for each probing question, the CSC app would tell the user why it asked such questions, and for the potential diagnostic results, it would explain what inputted symptoms led to the diagnostic recommendations.

Third, as emotional support and explanations were both desired by our participants, we designed a conversational style that combined these two types of information together. This conversational style not only provides emotional support (such as greetings, assurances, and care suggestions), but also offers explanations for each probing question and diagnostic results.

Finally, our findings highlighted our participants' appreciation of emotional support and explanations, as well as their wish for efficiency in the interactions. To test a potential way to allow users to balance human-likeness, explanations, and efficiency, we proposed to design a toggle button on the CSC's interface which offers user control for the presence of emotional support and explanations prompted by CSC apps. In other words, users can choose to hide or unhide caring words and explanations by turning on or off the toggle button at any time.

6 STUDY 2: TESTING PROPOSED DESIGN SOLUTIONS THROUGH AN EXPERIMENTAL STUDY

Based on our findings from Study 1, we designed a lab-controlled experimental study to explore potential design solutions. Informed by our Study 1 findings that users desire emotional support, explanations, and efficiency, we developed a CSC with four different conversational styles: baseline, emotional, explainable, and combined.

We then investigated how each style influenced the user experience by recruiting another set of participants. We designed and developed a symptom checker with the four different conversational styles and had our participants seek potential diagnoses by interacting with the four conversational styles. After experiencing each style, participants filled out a questionnaire that measured their experience. We also had follow-up interviews with the participants to further understand their perceptions of the different designs.

6.1 System Design

The interface of our proposed CSC was developed using JavaScript and HTML/CSS language. We chose to design the symptom checker with a chatbot-based feature because conversational interfaces have been widely adopted in online applications and can provide natural user interactions [139]. The conversational flows for each conversational style were built with pre-designed rules using Flow XO [6], a chatbot platform used to create online chatbots. A conversational flow would listen for a set of keywords as a trigger. Different inputs from the user would trigger a different series of actions of the proposed CSC.

We embedded the conversational flows into the CSC's web interfaces through the API provided by Flow XO. This framework enabled us to develop easy-to-use user interfaces in a time-efficient manner. We finally created four web pages. Each of them represented one of the four conversational styles of the CSC. For each web page, we designed and embedded different conversational flows to implement each conversational style (i.e., baseline, emotional, explainable, and combined) based on our empirical findings from Study 1 (see Table 1 for details) and prior research literature. More specifically, we referred to previous XAI studies [150, 152] to design the explanations and referred to literature on human-like design [38, 41, 89, 141] and communication skills in clinical practice [13, 122] to generate the emotional support responses.

Table 1. Proposed Design based on Study 1

Empirical findings from Study 1	Our proposed design	Conversation styles
Participants desired emotional support, such as greetings	Start with addressing and greeting the user; show caring and comforting words	Emotional style
Participants preferred further treatments	End the conversation with treatment options	Emotional style
Participants liked questions with diverse sentence structures	Use diverse asking patterns, such as "May I ask" and "Do you have"	Emotional style
Participants were confused about irrelevant, repetitive, and redundant questions	Provide explanations for each probing question	Explainable style
Participants were concerned with diagnostic recommendations	Provide explanations for diagnostic results	Explainable style
Participants wanted both emotional support and explanations	Offer both emotional support and explanations	Combined style
Participants wished for efficiency in interactions	Toggle button	Emotional, explainable, and combined style

The web interface of the proposed CSC had three components (see Figure 1): (1) an input box at the bottom that allowed users to input their symptoms; (2) clickable buttons displaying select

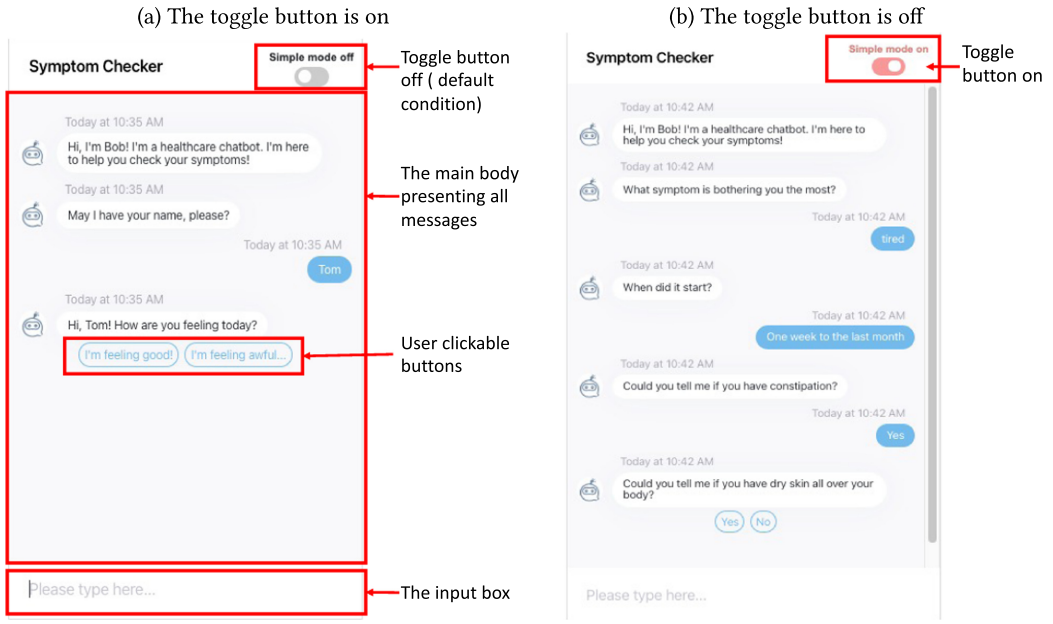

Fig. 1. Interface of designed CSC.

options (i.e., users can simply click the button to choose their options); and (3) the main body in the middle that presented all the responses from the checker as well as the users' input. Messages from the checker and users were distinguished using different colors. For each conversational style, we generated a web link and delivered it to participants, so that participants can open the link and use the symptom checker with the specific conversational style on their computers.

Efficiency View: With the exception of the baseline style, we designed a **toggle button** (see Figure 1) on the top right of the interface for all other conversational styles; users could hit the button at any time to decide whether the checker should provide caring language and/or explanations depending on the conversational style. If the text over the button showed "Simple mode off", the user received caring language and/or explanations all the time; if the user felt overwhelmed by the caring words and/or explanations, the user could click the button, and the text would become "Simple mode on", at which point, the user would receive neither the caring words nor the explanations.

We elaborated below the details about how each conversational style was designed and how their conversational flows were implemented in Flow.XO.

Style 1: Baseline style. The baseline style had a traditional, standard conversational design without any emotional support cues or explanations. The conversational flows of this style only included questions asking about users' symptoms and provided recommendations in formal language, such as "*Please tell me the main symptom that bothers you most" and "Please tell me if you have a fever*". We did not design a toggle button on the interface for this style.

Style 2: Emotional style. This style provided emotional support and human-like features for the user (see Figure 2) during each conversational stage, informed by Study 1 findings and previous literature. To be specific, our empirical findings from Study 1 (see Table 1 for details) indicated that users desired greetings, friendly addresses, caring language, and potential treatment suggestions (i.e., what users should do to treat their symptoms, e.g., how to ease the symptoms and whether to see a doctor). A set of prior literature also emphasized the human-like design [38, 41, 89, 141] as

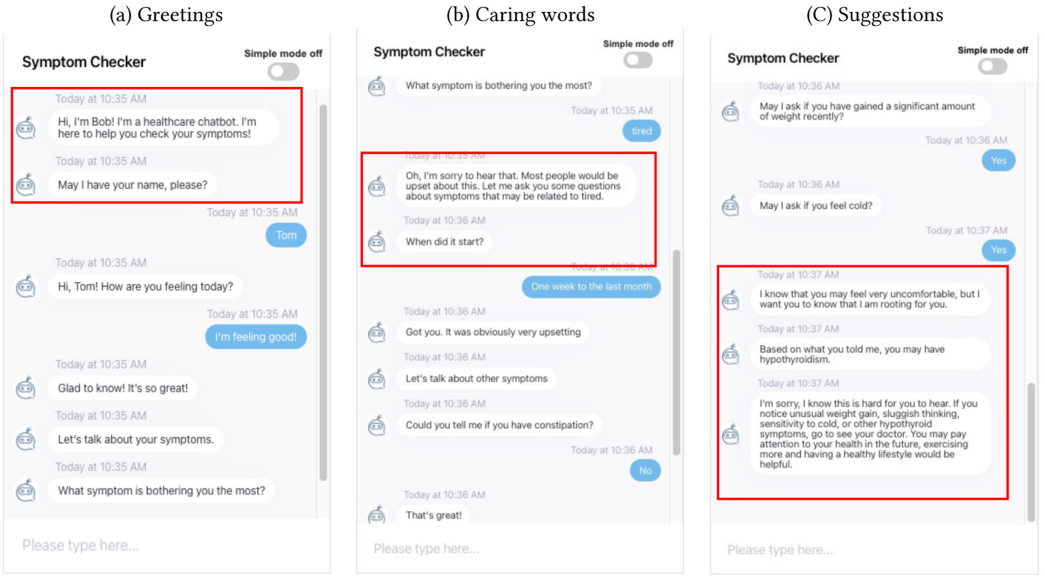

Fig. 2. Designed CSC with emotional style.

well as communication skills in clinical practice [13, 122]. Therefore, in the conversational flows of this style, we designed friendly addresses, greetings, caring words, encouragements, and potential treatment suggestions to offer comfort and encouragement.

The conversation flows of this style start with addresses and greetings, as our findings from study 1 reported that users desired these human-like languages. Some examples include "*Hi, I'm Bob! I'm a healthcare chatbot! I'm here to help you check your symptoms.*","*May I have your name please?*", and "*Hi, XXX! How are you feeling today?*" If the user answered "*I'm feeling good*", the checker would say "*Glad to know! It's so great!*"; otherwise, the checker would reply with comforting words, such as "*Thank you for sharing your feelings. Don't worry, I'm here for you!*"

For the probing stage in the conversation, we designed encouragements and comforting language. This style responded with human-like cues before asking follow-up questions for each symptom, such as "*Let me ask you a question about your tiredness.*" or "*I'm going to ask you more questions about...*" After each response from the user, the checker would offer encouragements, such as "*Don't worry! It's going to be okay,*" or comforting words, such as "*Oh, I'm sorry to hear that, most people would be upset about this.*"

At the end of the conversation, after showing the potential diagnoses, caring language was provided, such as "*Don't worry! Colds usually get better on their own in a few days*" or "*I know this is hard for you to hear*". After this, there were treatment instructions to encourage the users to engage in a healthy lifestyle and take actions to treat their symptoms, such as "*Please contact your doctor or call 911 immediately. You can pay attention to your health in the future by exercising more, and having a healthy lifestyle would be helpful.*"

Our Study 1 findings also showed that users desired a flexible structure of questions. Hence, distinct from the more formal question style of the baseline and the explainable design, the emotional style asked questions and delivered recommendations in a human-like, flexible way. For example, the checker would ask "*What symptom is bothering you the most?*" to solicit the user's main symptom, and then "*Could you please tell me if you have a fever?*" or "*May I ask you if you have a fever?*" to probe if the user had any other symptoms, followed by "*Thank you! Based on what you told me,*

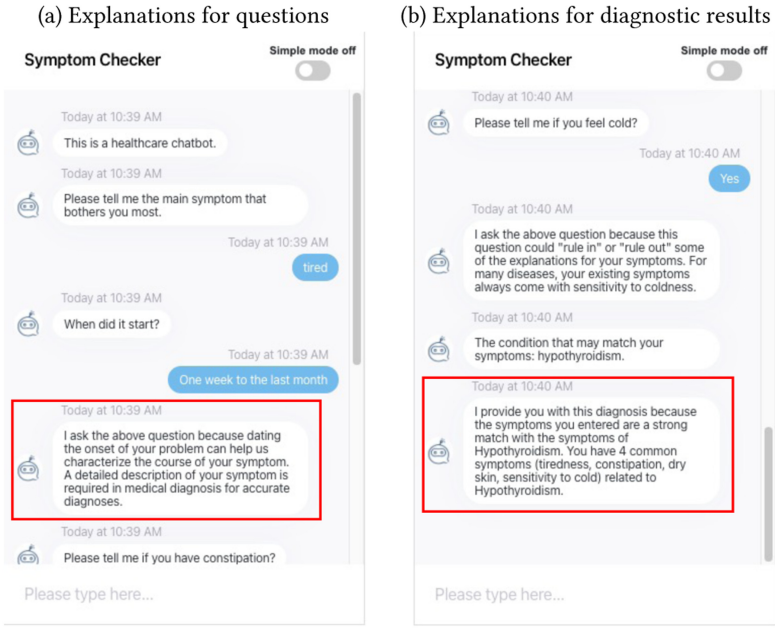

Fig. 3. Designed CSC with explainable style.

you may have a common cold." We also designed a toggle button for users to control whether or not they want to receive encouragements, comforting language, or treatment instructions.

Style 3: Explainable style. Explainable style was designed based on Study 1 findings and previous XAI studies [150, 152]. As the Study 1 findings reported, users were confused by the probing questions and diagnostic results, engendering a need for explanations. Informed by this, the explainable style focused on the rationale of each probing question and recommendation (see Figure 3). These explanations were designed based on the information from the WebMD website [3] and previous literature [150, 152]. They were also verified by a family doctor.

To provide a rationale for each probing question, this style explained why the checker asked each question to users. For example, in the case of a common cold, the checker prompted the user, "*Please tell me if you have a fever (Yes/No)*". The reason for this question is that having a fever can lead to different diagnoses based on the user's previous inputs. Thus, the checker would provide an explanation along the lines of, "*I ask the above question because this question could "rule in" or "rule out" some of the explanations for your symptoms. For many diseases (such as Influenza), your existing symptoms always come with a fever.*"

To explain why the checker provided the potential diagnoses, this style offered a personalized summary in the conversational flows, stating which symptoms led to the recommendation. In the case of the common cold, the checker would offer an interpretative summary at the end, like, "*I provide you with this diagnosis because the symptoms you entered are a strong match with the symptoms of a common cold. You have 3 common symptoms (throat pain, runny or stuffy nose, and cough) related to a common cold.*" We also designed a toggle button for users to control whether or not they wanted to receive explanations for this style.

Style 4: Combined style. Our findings from Study 1 implied that users desired both emotional support and explanations. Thus, the combined style not only applied human-like cues to its conversation, but also provided the rationale behind the probing questions and recommendations for

users. We designed a toggle button for users to control whether or not they wanted to receive caring words and explanations.

6.2 Evaluation

6.2.1 Participants. We recruited participants for the experiment through social media and Studyfinder, a participant recruitment tool at our university. Participants were eligible if they were (1) at least 18 years of age; (2) currently living in the United States; (3) spoke English; and (4) had laptops or computers to use the CSC we designed.

We recruited 34 qualified participants for our experimental study. Their ages ranged from 19 to 72 years old (M = 36.85, SD = 13.32; M = Mean, SD = Standard Deviation). The participants included 16 females, 17 males, and one person who is non-binary. They had a diversity of education levels as well as professions, such as statistician, software engineer, package handler, and preschool teacher. Their demographic information is shown in Table A.2 in the Appendix.

6.2.2 Study Design. The experiment was conducted from October 3, 2021 to October 25, 2021. The research had been approved by Penn State University's IRB. All the data obtained from the participants was securely stored and destroyed upon the completion of the project. All sessions of the experiment were conducted and video-recorded via Zoom. The experimental procedure is shown in Figure 4.

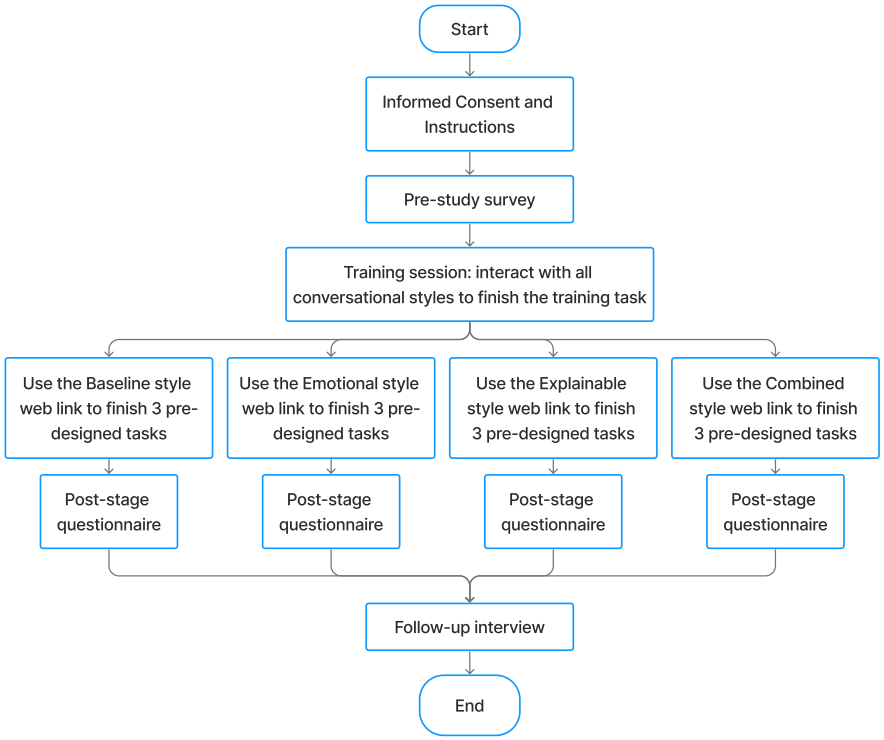

Fig. 4. Experimental procedure of each participant (all tasks and order of conversational styles were randomized).

Before each experiment, we sent the electronic consent form to the participants and obtained their agreement. We then collected the participants' basic demographic information, such as name,

age, profession, and education level in the pre-study survey. We also included two questions on a seven-point scale to control for the participants' previous experiences with CSC apps and conversational agents. The average score of the familiarity with symptom checkers, i.e., "I feel I am familiar with symptom checkers" was relatively neutral (M = 5.19, SD = 1.35). The average score of familiarity with chatbots, i.e., "I feel I am familiar with chatbot agents or apps (conversational agents or apps)" was also neutral (M = 5.16, SD = 1.58). This indicated that most participants had some knowledge of symptom checkers and chatbots. In addition, we also included one question asking whether participants desired emotional support from another person. The average score of the desire for emotional support was relatively high (M = 6.06, SD = 0.67) on a seven-point scale. This indicated that the majority of participants wanted emotional support from other people.

The research procedures included a within-subject experiment in which we let each user interact with four types of conversational styles and a short follow-up interview. Following the within-subject design, we let each user interact with each conversational style for a training session and three pre-designed tasks (these tasks and the order of conversational styles were randomized following a Latin square design). In the training session, participants were required to use the proposed CSC to finish the training tasks to ensure they were familiar with the checker and all the conversation styles. When assigning tasks to each participant, we followed a Latin square design to minimize potential ordering bias. We also followed a Latin square design to let each user interact with each conversational style. After interacting with each conversational style for all tasks, the participants needed to fill in a post-stage questionnaire. The questions were also shuffled each time they were used to avoid order bias.

A semi-structured interview was conducted at the end of the user study to acquire more in-depth and comprehensive insights about how each participant experienced the prototype (the proposed CSC). We focused on their experiences with different conversational styles. The interview questions are shown in Appendix E. Each interview took between 30 minutes to 1 hour. We compensated each participant with a $20 Amazon gift card.

The total process for each participant (including the interactions with all designed conversational styles, filling out post-stage questionnaires, and the follow-up interview) lasted between 45 and 100 minutes (M = 60.01, SD = 10.5). The participants spent around four minutes in each conversational style: Baseline (M = 3.21, SD = 1.25), Emotional (M = 4.52, SD = 2.05), Explainable (M = 3.52, SD = 1.60), and Combined (M = 4.30, SD = 2.55).

Participants (N = 34) were presented with the same scenarios for each conversational style. Each participant interacted with all conversational styles to finish three tasks in the pre-designed scenarios. These scenarios were adapted from validated patient vignettes presented in a previous study [85] and were classified into three categories of triage urgency [35, 85]. The scenarios were designed to cover different kinds of health conditions to ensure that the users had sufficient experience in various cases. The three types were (1) Self-care: the health issue or condition can get better on its own; seeing a doctor is likely not required; (2) Non-emergency: the health issue cannot get better on its own; seeing a doctor is likely required but not emergent; and (3) Emergency: the health condition requires immediate attention; the user likely needs to contact 911 or go to the Emergency Department. To offer the correct diagnostic results for different symptom inputs, we generated three decision tree models for each of the scenarios (see Appendix B), which we then verified with a family doctor, through previous research literature [10, 58, 151] and Ada (the most widely-used and accepted symptom checker in the app market [7]). Based on the pre-designed scenarios below, we asked participants to consecutively interact with each of the four types of conversational styles to find out potential diagnoses. The detailed scenarios and tasks are as follows.

Task 1: Mary (female, 23 years old) is an undergraduate student who wants to self-triage using a chatbot-based symptom checker. She has had the main symptom of a runny nose for more than one week but less than one month. She also has other symptoms: a sore throat, and cough, but she doesn't have a fever. Please use the symptom checker to find out her potential diagnosis.

Task 2: John (male, 29 years old) is a teacher who wants to self-diagnose using a chatbot-based symptom checker. He has been tired for more than one month, which bothers him the most. He also has other symptoms, such as constipation, dry skin, weight gain, and frequently feeling cold. Please use the symptom checker to find out his potential diagnosis.

Task 3: Tom (male, 30 years old) is a software engineer who wants to self-triage using a symptom checker. His main symptom is located in his lower right abdomen. He has not had similar pain before. He also has other symptoms like vomiting, a lack of appetite, nausea, and a fever. He doesn't have any other symptoms. Please use the symptom checker to find out the potential diagnosis.

6.2.3 Measurements and Data Analysis. The goal of Study 2 is to investigate potential CSC conversational design solutions by testing Study 1's design implications. Therefore, here, we used mixed research methods to measure how participants experienced the design solutions (the CSC we developed based on Study 1's design implications) to explore how to meet users' needs regarding the CSC's conversational design. We first utilized a statistical analysis to analyze quantitative questionnaire data and the use log (that recorded all users' conversations with the CSC). Then we used an inductive thematic analysis on the qualitative interview data to complement and explore the quantitative findings [111, 158].

A post-stage questionnaire was designed to collect quantitative data to measure the user experience and effectiveness of each conversational style. Since Study 1 revealed that users expected explanations, emotional support, and efficiency when interacting with CSC apps, our questionnaire was specifically developed to let participants evaluate these three aspects regarding the different conversation designs, which are particularly important for CSC apps. To measure user experiences with explanations, we utilized four constructs (i.e., effectiveness, trust, transparency, and satisfaction), which are widely-used and identified as key aspects to measure the effectiveness of explanations [42, 104, 149]. We further collected three constructs of emotional support in the post-stage questionnaire. These constructs included likeability, emotional support, and human likeness. These constructs and corresponding questions were adapted from previous literature [40, 42, 89, 137] and could help us understand whether the participants perceived the proposed chatbot-based symptom checker as human-like and enjoyed the conversation with it. To capture how participants perceived the efficiency of each conversational style and explored the use of the toggle button, we constructed and adopted questions from previous works that measured the efficiency of chatbot [75, 166].

In total, 17 questions were developed on a seven-point Likert scale (from "strongly disagree" to "strongly agree") based on Study 1 findings and existing works, including the eight constructs of trust, satisfaction, efficiency, likeability, effectiveness, transparency, human likeness, and emotional support (See Appendix C for details).

Additionally, four NASA-TLX usability questions on a seven-point semantic differential scale were included in each questionnaire to measure each design's mental demand, performance, effort, and frustration level.

To analyze the questionnaire data, we first calculated Cronbach's α for each questionnaire construct to guarantee internal reliability. We also conducted a Wilcoxon signed-rank test on the questionnaire data in order to compare the user experience of the four types of conversational styles. We chose this statistical method because our data distribution was not normal, and

the Wilcoxon signed-rank test does not require a normal distribution [33]. Finally, we used the Benjamini-Hochberg procedure to adjust the p-values to avoid Type I errors. We also used the Kruskal–Wallis rank sum test to confirm whether the conversational types' order of presentation had an effect on the questionnaire responses. The results showed there was no statistical significance ($p > 0.05$). Therefore, no order bias was found.

To understand the use pattern of the toggle button, we conducted a log analysis by recording the timing of the participants' click actions and manually classifying them based on similar click patterns.

The qualitative data was captured through follow-up semi-structured interviews. We started the interview by asking which conversational style the participants preferred and why. Then, based on the participants' answers, we asked what they thought of each conversational style, how they felt about the caring language and/or explanations, when they thought the conversations were overloaded, and their timing preferences for receiving caring language and/or explanations. We also asked why the participant pressed or did not press the toggle button during their interactions with the different styles. Each of the interviews took between 20 to 40 minutes. Similar to how we conducted data analysis in Study 1, we then transcribed all the interviews and utilized an inductive thematic analysis to analyze the transcriptions.

6.3 Results

After analyzing questionnaire data, logs, and interview transcriptions, we identified two main themes: (1) how did emotional support and explanations affect user perceptions of the proposed CSC?; (2) when should the CSC provide emotional support and explanations? The detailed findings are explored in the following section. When reporting our qualitative findings, we used S1, S2, and so on to denote the different subjects.

6.3.1 How did Emotional Support and Explanations Affect the Users' Perceptions of the Proposed CSC?. For the questionnaire data, we first calculated Cronbach's α for each questionnaire construct to guarantee internal reliability. The results showed that most Cronbach's α values were between 0.8 and 0.9, which indicated a good internal consistency. Only three alpha values were between 0.6 and 0.7, which was also an acceptable reliability level [37]. The Cronbach's α and descriptive statistics for the questionnaire factors and variables are presented in Tables 2 and 3. Overall, the descriptive statistical results showed that the average rating of each factor was between four and five. Based on the analysis of the questionnaire data, the results (see Table 4) showed that only three factors out of eight (trust, transparency, and empathy) had statistically significant differences. In addition to the eight constructs, we also reported on the **NASA Task Load Index (NASA-TLX)** scores [56, 59] (see Table 5). We did not find statistically significant differences among these four factors. We further explored user feedback through a qualitative analysis of the interview transcriptions.

Trust. We examined whether caring language and/or explanations could improve user trust in CSCs. Based on a quantitative analysis of the questionnaire data, we found that the scores of the explainable style ($V = 53$, $p = 0.049$) were significantly higher than for the baseline style. Providing explanations increased trust scores by 0.25 points.

In line with the quantitative analysis, our findings reported that providing explanations could increase participants' trust in the CSC, since explanations could make the CSC appear more reliable from the participants' perspectives. For example, S34 said,

"*I think [an explanation] is nice, because it shows you how it came to that conclusion so I felt that it was more believable, and that this CSC really knew what it was talking about. So just [this] you know, made it more believable for me.*" (S34)

Table 2. Cronbach's α for the Questionnaires

		Cronbach's α			
Factor	Variable	Baseline	Emotional	Explainable	Combined
Trust	Q1, Q2	0.86	0.82	0.77	0.88
Satisfaction	Q3, Q4, Q5	0.91	0.88	0.95	0.9
Efficiency	Q6, Q7	0.63	0.82	0.78	0.87
Likability	Q8, Q9	0.9	0.85	0.95	0.82
Effectiveness	Q10, Q11	0.81	0.61	0.84	0.77
Transparency	Q12, Q13, Q14	0.88	0.85	0.8	0.84
Human likeness	Q15	NA	NA	NA	NA
Empathy	Q16, Q17	0.63	0.88	0.7	0.76

Table 3. Descriptive Analysis of Questionnaire Data

		Mean (SD)			
Factor	Variable	Baseline	Emotional	Explainable	Combined
Trust	Q1, Q2	5.49 (1.01)	5.66 (1.17)	5.74 (1.05)	5.68 (1.05)
Satisfaction	Q3, Q4, Q5	5.55 (1.30)	5.68 (1.33)	5.46 (1.34)	5.66 (1.29)
Efficiency	Q6, Q7	5.74 (1.12)	5.88 (0.93)	5.54 (1.27)	5.78 (0.99)
Likability	Q8, Q9	5.13 (1.12)	5.37 (1.35)	5.01 (1.63)	5.34 (1.39)
Usefulness	Q10, Q11	5.43 (1.67)	5.79 (1.04)	5.47 (1.18)	5.89 (0.96)
Transparency	Q12, Q13, Q14	4.68 (1.70)	5.57 (1.23)	5.72 (1.13)	5.78 (0.98)
Human likeness	Q15	4.88 (1.64)	5.32 (1.69)	4.65 (1.81)	5.21 (1.47)
Empathy	Q16, Q17	4.63 (1.66)	5.44 (1.41)	4.84 (1.74)	5.43 (1.41)

Table 4. Comparison Results of Four Conversation Styles

	Comparison results					
Factor	BS vs. EM	BS vs. EX	BS vs. CM	EM vs. EX	EM vs. CM	EX vs. CM
Trust	V = 52	V = 53*	V = 85.5	V = 86	V = 153.5	V = 137.5
Satisfaction	V = 151.5	V = 157	V = 128	V = 182	V = 123.5	V = 134
Efficiency	V = 75	V = 85.5	V = 142	V = 172.5	V = 119.5	V = 130
Likability	V = 85.5	V = 148.5	V = 99.5	V = 188.5	V = 135.5	V = 91.5
Usefulness	V = 90	V = 71	V = 52	V = 137	V = 91	V = 53.5
Transparency	V = 34**	V = 35**	V = 33**	V = 77	V = 92.5	V = 168.5
Human likeness	V = 61	V = 106.5	V = 42	V = 201	V = 109.5	V = 47.5
Empathy	V = 61**	V = 119.5	V = 68.5**	V = 222	V = 138.5	V = 86.5*

Statistical significance level: (*)p < 0.05. (**)p < 0.01. (***)p < 0.001.
BS: Baseline, EM: Emotional, EX: Explainable, CM: Combined.

In the above comment, S34 thought that by showing explanations, the CSC provided clear evidence of the relevant diagnostic knowledge. Therefore, she considered the CSC to be more reliable due to the provided explanations. Likewise, S1 commented,

"*I don't remember doctors ever explaining to me why he or she made such a decision, because I believe in the doctors' expertise, so they don't need to explain that to me, but when I use the symptom tracker, maybe I need some evidence to convince me that it is reliable. I need some explanations like checker three [explainable interface] or four [combined interface].*" (S1)

Table 5. NASA-TLX Usability Analysis

		Mean (SD)			
Factor	Variable	Baseline	Emotional	Explainable	Combined
Mental Demand	TLX1	3.18 (1.84)	2.91 (1.67)	3.09 (1.76)	3.38 (1.94)
Performance	TLX2	5.76 (1.64)	5.68 (1.39)	5.79 (1.02)	5.97 (1.01)
Effort	TLX3	2.91 (1.82)	2.74 (1.67)	2.91 (1.69)	3.06 (1.86)
Frustration	TLX4	2.65 (1.59)	2.53 (1.63)	3.1176 (1.73)	2.88 (1.80)

Here, S1 expressed that she did not need explanations from doctors due to their recognizable expertise; however, since the CSC was not an actual doctor, she needed explanations from the CSC to demonstrate its reliability.

In addition, though there is no statistically significant difference in the trust factor between the emotional style and the baseline style, our interview findings reported that providing emotional support to the participants also improved their trust in the app. As S22 told us: "*I feel like it tries to know my feelings, so I feel it is trustworthy.*" The greetings, in the beginning, made S22 view the CSC as credible. S30 shared a similar experience:

"*I feel [the emotional interface] is more comfortable and more believable. I feel like it is more trustworthy because it seems like a discussion, shows more caring, and is not too wordy to use. It had talks and responses, so it is like a human doctor. It understands what I have and feel.*" (S30)

In this example, S30 thought caring responses rendered the CSC human-like and similar to an actual doctor, as these responses delivered caring and made her feel the CSC understood her input. This perception improved S30's trust in the CSC.

Transparency. We wanted to know if caring language and/or explanations could enhance the perceived level of system transparency. From the quantitative analysis, we found that the scores of the emotional style ($V = 34$, $p = 0.002$), explainable style ($V = 35$, $p = 0.002$), and the combined style ($V = 33$, $p = 0.002$) were significantly higher than the baseline style. Providing explanations and/or caring language increased the transparency score by nearly, or over, 1 point, which had a large effect on a 7-point scale. The combined style outperformed the other two. Thus, we concluded that providing explanations and caring language could increase the CSCs' level of perceived transparency.

During the interviews, our participants' comments also supported this conclusion. For instance, S12 told us:

"*I feel I like more about the [explainable style], because compared with the [baseline style], this one gave more detailed information of specific diseases. I feel it gave me much more information about how it gave the results and the background of the diseases.*" (S12)

Here, compared to the baseline style, the explainable style better helped the user understand how the CSC generated the results and learn more detailed information about their potential illnesses.

Similarly, S15 enjoyed the combined style the most, since it was more transparent, offering information about its questions and symptoms. S15 commented,

"*I prefer the [combined style]. It has a sympathetic tone and it feels [like] there's a bit more of a connection. And it gives you explanations of why it's asking what it's asking. It gives the details on the questions and symptoms, while not feeling as mechanical.*" (S15)

Emotional support. Our quantitative analysis demonstrated that the scores of the emotional style ($V = 61$, $p = 0.006$) and combined style ($V = 68.5$, $p = 0.006$) were significantly higher than those of the baseline style. The scores of the combined style ($V = 86.5$, $p = 0.046$) were also significantly higher than those of the explainable style. This reflected the idea that providing caring

language or caring language and explanations could improve the perceived level of emotional support.

Our interview findings also found that most participants appreciated the emotional support messages and perceived the emotional style as empathetic. For example, S34 said,

"*I really liked the [emotional support]. It makes me feel like it actually cared about you, like it doesn't want you to be freaked out if you have appendicitis and you need surgery. It's just more reassuring and calming, and empathetic, like people who care about you.*" (S34)

In this case, S34 enjoyed the emotionally supportive language, as it delivered care and empathy to her, making her feel reassured, especially in cases of a serious condition.

Human likeness. We did not find a statistically significant difference between human likeness and conversational styles. The below introduced how our participants perceived the human likeness of emotional style and explainable style.

(1) Emotional style

For the emotional style, we did not find a statistically significant difference between it and other conversational styles in the construct of human likeness. One possible reason for this is that the emotionally supportive messages were offered after each of the user's responses. This way of providing emotional support was markedly different from a doctor in a medical consultation. S1 explained,

"*I don't need that after each time I answer a question; if I went to the doctor, I don't think the doctor would be that dramatic, comforting you that frequently.*" (S1)

Though there was no statistically significant difference, some participants thought that adding emotional support made the CSC seem like a doctor. As S8 said, "*It's like you're interacting with a normal doctor, the conversation's there. It's like you're talking to a physician.*" Similarly, S28 told us:

"*Greetings and asking about my name are good. It looks like a human-like conversation with a doctor, [a] good conversation. It was also a good start for following communication, so I believe it really works for me.*" (S28)

Here, S28 appreciated the greetings and the request for a name, because these sentences were doctor-like, creating a good start for further communication.

However, for some, the increased amount of emotional support rendered the CSC less human-like, as S2 explained:

"*because I have experience visiting the hospital, I don't think that doctors provide this kind of emotional responses. That's why I don't think I like it. It's hard to make it like a human.*" (S2)

In this case, S2 did not think the CSC was human-like because there was too much emotional feedback, which is unlike a clinical medical visit.

(2) Explainable style

When it came to the explainable style, most participants perceived that adding explanations made the conversation robotic and less human-like. For example, S30 stated, "*I just think [the explainable interface] is more robotic; you can't connect much with it. It doesn't look at how I really feel.*" Here, S30 didn't consider the explainable app to be human-like, since it did not probe his personal feelings. Likewise, S1 told us:

"*I think the [emotional interface] is more similar to a human, like human doctors, because I don't remember the doctors ever explain[ing] to me why he or she made such a decision.*" (S1)

In this case, S1 thought that the explainable style was unlike an actual doctor due to the explanations, as a doctor would not explain why he/she made such a diagnosis.

Information overload. Our participants complained about the information overload they encountered when interacting with emotional style and explainable style.

(1) Emotional style

In our interview study, we found that some participants looked at the caring language in terms of both quantity and length. For example, S23 complained: "*There were too many caring words, because when I'm sick, I like to go straight to the elements, to know the symptoms directly.*" Here, S23 desired a more concise conversational style instead of an excess of caring language when acquiring symptom information. Likewise, when being asked why he turned on the toggle button after answering two questions, S2 told us that this was because he thought there was too much caring language. He explained: "*I thought the emotional support was too much just after two questions. It's really unnecessary, so I pressed the toggle button to use the simple mode.*" (S2)

S32 shared a similar experience: "*I think it was too much when it was after every response, like it's nice at the beginning, or the end, but after every question, it was kind of a lot.*" She thought that the caring language was too much and should not be presented after each sentence.

In addition to the quantity, our participants also complained about the lengthiness of the caring language. For instance, S11 told us: "*I think I would mostly prefer for it to just stay at one sentence, with not so many lines.*" In this case, S11 wanted some caring language, but preferred that it was shorter.

(2) Explainable style

Our participants also complained about the numerous explanations provided. As S9 put it, "*the explanations after each sentence was very annoying and aggravating and makes you even not want to read it. It would just waste my time to go through it.*" She thought the explanations should not be presented after each question, as they lengthened the whole diagnostic process and made it annoying.

The lengthiness of explanations also confused some participants. For instance, S5 wanted shorter explanations that could be presented in just one or two sentences instead of a long paragraph. As she put it, "*It's overloaded with explanation, like it's a big long paragraph that you have to sit there and read. One or two sentences is good enough.*"

Privacy concerns about the emotional style. Though the security of the CSC was not measured in the post-study questionnaire, we found that in our interview study, some participants were concerned about the security of the CSC. These participants stated that the solicitation of users' names should be removed due to privacy and security concerns. For example, S12 commented,

"*I think asking my name is not so good because I don't want to tell the phone my name because I feel it's kind of private. I don't want to disclose my real name to a symptom checker. I'm not sure if the company will collect my information during my conversation with this conversational agent, so I kind of hesitate to disclose my name because that's personal information.*" (S12)

S12 was concerned that disclosing his name to the CSC might be harmful to his security, since he thought that the company behind the app might collect his personal information. S13 also emphasized that the CSC should not ask for the user's name since it was private information. He noted, "*The name part should be excluded. It is my privacy, so it shouldn't be there.*"

Efficiency, satisfaction, likeability, and effectiveness. From the questionnaire data analysis, we did not find a statistically significant difference in the perception of efficiency between different conversational styles. This might be because of the relatively short conversations we designed, as S1 told us: "*These conversations diagnosing my symptoms were not that long, so I think all styles are efficient for me, to get the diagnoses quickly.*"

Similarly, we did not find a statistically significant difference in the constructs of satisfaction, likability, and usefulness. One possible reason might be that the participants thought that all the conversational styles could meet most participants' self-diagnosis needs. For instance, though S14 liked the emotional style best, when being asked how he perceived the other conversational

styles, he said, "*I thought they are just all right.*" Similarly, S7 commented, "*I think both ways [emotional style and explainable style] work pretty well. Well, they both got to the point and told me what I had to do.*"

To summarize, our findings showed that providing emotional support and explanations affected the user perceptions of the proposed CSC in the following six ways: trust, transparency, emotional support, human likeness, information overload, and privacy. On the one hand, the caring language and/or explanations improved user trust in the CSC and enhance its perceived levels of transparency and emotional support; on the other hand, at times, they raised concerns regarding information overload and user privacy.

6.3.2 When should the CSC Provide Emotional Support and/or Explanations? By analyzing log data and interview transcriptions, we discovered when the CSC should provide emotional support and/or explanations. We report the detailed findings in the following sections.

The use of the toggle button. We observed four types of click patterns in the system log data. We then picked and plotted one representative case for each click pattern in Figure 5. In Figure 5, each stream represents the click pattern for each individual participant for one specific task. The x-axis indicates the time of one interaction. The y-axis indicates different actions. For different assigned tasks, some participants applied different click patterns.

For the interactions with the first click pattern (n = 10), the participants clicked on the toggle button after several seconds (45 seconds in the case shown in the diagram) from the start of the interaction. These participants turned the toggle button on and did not turn the button off again. They did this because they considered the information bothersome. For example, S1 told us: "*I just want the information. It's a little annoying.*"

For the interactions with the second click pattern (n = 9), the participants turned the toggle button on at 15s and turned it off at 67s. This set of participants turned the button on only during the middle stage. This was because they considered the extra information (caring language and explanations) during the middle of the conversation useless; as S11 put it, "*I was thinking the extra information in between was not really necessary.*"

For the interactions with the third click pattern (n = 127), the participants pressed the button as soon as the interaction was initiated. These participants said they did not want caring language or explanations at all because the most important thing was to get diagnoses as soon as possible. S9 said, "*because I like it to be short and quick. Without any of the extra comments from the bot. That's the only reason. I just wanted it to ask me what I have and quickly tell me what to do next.*"

For the interactions with the fourth click pattern (n = 160), the participants did not click on the toggle button at all during the interaction. This was because of their curiosity and their desire to receive emotional support and explanations, as S7 told us: "*I didn't press the button, because I just wanted to see how much detail it would be. I was curious about how much detail it would have.*" S12 told us that because symptom checking was health-related, it was important to receive sufficient information. She explained " *I didn't feel it was necessary [to use the toggle button]... it's very important and essential to my personal health, so I wanted to get as much information as possible.*"

The click pattern analysis revealed that our participants used a variety of strategies to decide when they wanted caring language or explanations. Our findings from the post-study interviews further revealed the participants' different requirements for emotional support and explanations in various contexts. We report on the detailed findings below.

When to provide emotional support? We found that the CSC might not provide the same amount of emotional support during the interaction at different diagnostic stages, in situations with different emergency levels, or for users with different personalities. That is, participants had distinct needs for emotional support in different contexts.

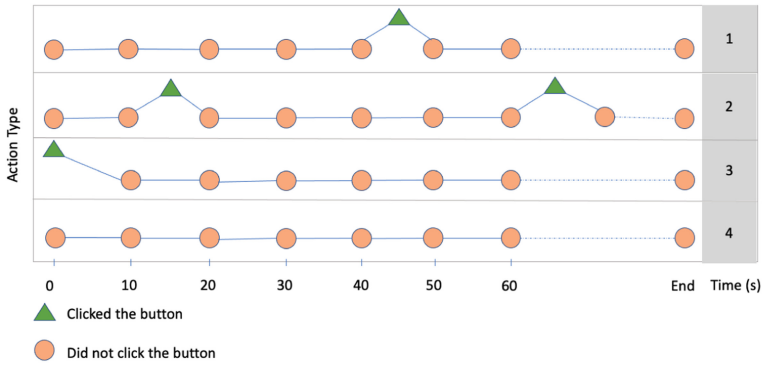

Fig. 5. Four types of click pattern.

(1) Offering emotional support at different stages

Our participants showed different preferences for emotional support during the different conversational stages. In the conversational openings, most participants (n = 28) enjoyed the emotional support, and at the end, most participants preferred emotional support (n = 32). In the middle, our participants' responses were differentiated: some participants appreciated the given emotional support (n = 21), while others disliked emotional support (n = 11), and one of our participants remained neutral.

In the openings, most of our participants liked the emotional support, such as the greetings. For example, S9 told us, "*It's like an opening of a conversation, and it makes me feel like I want to continue with the process*" For this participant, the greetings at the opening were like the beginning of a conversation, making her more willing to use the CSC. Similarly, S10 said,

"*I think the greetings are good. I think to ask someone how it's going is necessary because someone might not feel well and maybe would like to receive some greetings. It could be comforting.*" (S10)

Here, S10 thought the greetings were necessary, because they could comfort users who did not feel well. Only four participants expressed that they did not desire the addresses and greetings. For example, S34 stated,

"*I would maybe rather not have those features, the name-asking and greetings, because not having it would make it a little quicker and eliminate one step. It might make it easier for people.*" (S34)

S34 thought the greetings and name-asking were time-consuming and that removing these steps would make the apps easier to use.

In the endings, most participants liked the emotional support. For example, S14 told us:

"*I would say, after they tell you the diagnosis, then they could say something like "I hope you get better soon". I guess that's just kind of like a conversation in which someone says "well, I hope you get better'. Most people say that.*" (S14)

In this example, S14 appreciated the caring language at the end because he thought it was similar to the ending of a regular human-to-human conversation. Moreover, most participants thought the suggestions about what to do next were very useful. For example, S11 said that the suggestions provided toward the end could guide her to deal with her health problem. She explained,

"*I thought that the advice at the end was particularly useful, because it told you how to deal with the problem, like it explained that you might feel that your symptoms are relieved within a few days, or it might tell you that your symptoms would be relieved if you drank water or tea.*" (S11)

The suggestions were also perceived as useful by providing further treatment guidance when it was inconvenient for users to go to the doctor, such as during the COVID-19 pandemic. For instance, S23 considered the suggestions, such as drinking more liquids, helpful, as going go to doctor during the COVID-19 lockdown was difficult. S23 told us:

"*It's great, because you know, sometimes you don't want to [go out] during the coronavirus lockdown. Being told to drink liquids is helpful because now I don't need to go to the doctor to get diagnosed and told what I should do to treat myself so it's very okay.*" (S23)

The caring language in the middle of the conversation led to different participant opinions. Some participants enjoyed it, as S5 explained: "*I guess it offered a little bit more feeling, like "oh, I'm sorry to hear that", like you're actually talking to a human, and they understand you.*" In this instance, S5 desired the emotional support because it made the conversation feel more human-like, making her feel understood. However, the other set of participants did not want this emotional support, as S9 said, "*So in the beginning, just saying like how are you feeling. That's fine because it's just trying to know the status of my health. But in the middle, it is unnecessary to empathize with me. I just wanted it to be useful, I don't want it to be just a waste of time and a waste of me reading it.*" Here, S9 believed that compared with the greetings which can let her input her health status data, the emotional support responses shown in the middle were useless and time-wasting, since she did not need empathy from the SC at all.

S9 thought that the emotional responses should be offered after she received the final diagnostic result instead of in the middle of the conversation because it was a potential diagnosis, rather than the probing questions, that might make her upset.

(2) Providing different amounts of emotional support in different situations with different emergency levels

Our participants had different needs for emergency situations compared to non-emergency situations. In an emergency, our participants desired a more concise and direct conversation. For example, S24 preferred the baseline style in these situations so that he could diagnose himself quickly, as the given emotional support slowed down the conversation. He explained:

"*Okay, the human-like language in emergency situations will be a little bit slow. In an emergency, I would prefer the computerized language, the baseline checker. It goes straight to the point.*" (S24)

In non-emergency situations, participants appreciated the emotional support. For example, S34 desired the emotional support because it was more human-like and she had sufficient time to read the emotional support messages. She stated,

"*It [emotional style] seems more like you're talking to a human. Because if it's not an emergency, like if I have a common cold, a sore throat, I would actually take the time to sit down and read all of those [caring] messages. I'd rather have all the information at my fingertips and be able to really sit down and read it.*" (S34)

Likewise, S14 stated that because he would not see a doctor in a non-emergency, he needed care suggestions to know what he should do. He said, "*In a non-emergent situation where I wouldn't go to a health care professional or anything, it would probably be more helpful to have those small suggestions, like drink tea or water or something like that.*"

S14 also told us that he would not turn the toggle button on in a non-emergency, because he desired more details that could be acquired from emotional support messages, saying, "*if you weren't having an emergency maybe that's where you could turn it off [on emotional interface] and get a little bit more detail and a little bit more information.*"

When to provide explanations? Similar to emotional support, we found that participants had different needs for the number of explanations depending on the stage of the conversation and emergency level; we also found that users with different personalities had different preferences for the number of explanations.

(1) Offering explanations at different stages

Our participants showed different preferences for explanations at different stages of the conversation. At the end, most participants preferred explanations for the diagnostic results (n = 31). In the middle, our participants' responses were differentiated: some participants appreciated the

given explanations (n = 20), while others disliked the explanations (n = 12), and two of our participants remained neutral.

Most participants desired explanations at the end because they could help participants understand their potential conditions. For instance, S13 told us:

"*I turned off simple mode at the end, because I think it was helpful to have information. It also helped me know the condition I have, why the symptom checker has shown me that condition, why it has that condition for me. The more information you send me, the better I understand what is going on with me, what the symptoms are.*" (S13)

In this example, S13 turned the toggle button off to read the explanations toward the end. He believed that the explanations could help him understand which symptoms he had that had led to the final diagnosis. S15 also said that the final explanations made him realize how his symptoms signified a more serious disease, saying,

"*I actually really like the [explanations], and when I had to do the hyperthyroidism example, I would have never connected those symptoms together. So, it was really eye-opening that those were things that could actually be symptoms of a bigger issue.*" (S15)

For the middle of the conversation, our participants had differing attitudes. Some participants desired explanations because of their educational value. For example, S32 commented,

"*I have a lot of doctors in my family, so I've gotten used to the way they think, and I understood how a diagnosis can be somewhat subjective, so I appreciated the rationale behind [the questions] to help me make decisions on whether the diagnosis is reasonable.*" (S32)

S32 knew a diagnosis could be subjective, so she needed explanations of the rationale behind the questions to assist her to understand how the CSC generated the diagnosis and evaluate whether the diagnosis was rational. Similarly, S15 told us:

"*I think people that may not understand why these kinds of questions are important would find some educational value in it, a better understanding of why the app needed to know this medically for the symptoms.*" (S15)

Here, like S32, S15 thought that the explanations were valuable to helping users understand why the CSC asked the questions. On the contrary, others thought that the explanations were unnecessary and time-consuming. S20 disliked the explanations, as she thought their appearance increased the total time until she received a diagnosis. She said, "*it just adds too much time for me to get my diagnosis, so you don't want them in the middle at all.*" Moreover, S9 stated that the explanations were unnecessary since they were common sense, as she stated,

"*The information that's presented here in the red box [in the middle] is completely unnecessary. Everybody knows what it is trying to do, that they're trying to understand my symptoms and find out what I have possibly.*" (S9)

In line with the above comments, S4 believed that the CSC should not provide explanations if it was common sense why the question was being asked; at the same time, she thought that explanations should be offered for questions with a less clear rationale behind them. She stated,

"*For some, [it is] common knowledge, common sense; they don't have to provide this information. For example, if you go to see the doctor, and you said that you have an abdominal pain, they would ask you where it is, but they would not tell you why, that's redundant, that is too much. But it would be helpful to tell me why it asked about the vomiting, because I don't know whether abdominal pain is related to vomiting.*" (S4)

According to S4, for the question on pain location, offering explanations was unnecessary as the reason for it was common sense; for the question about vomiting, an explanation was more necessary, since she did not know about the relationship between abdominal pain and vomiting.

(2) Providing a different number of explanations in different situations with different emergency levels

Our participants had different needs for emergency situations compared to non-emergency situations. In an emergency, our participants desired a more concise and direct conversation. For example, S19 told us:

"*If you are in an emergency situation, like if you have a severe abdominal pain, then you want it to be direct, because you're not feeling good, you don't want to have to read the whole explanation, you just want to know your disease and get better.*"

As S19 said, she disliked the explanations in emergency situations, as she thought reading the explanations was too time-consuming and thought receiving a diagnostic result as soon as possible was more important. S34 also did not want explanations in these situations since the explanations were overloaded; rather, she preferred a fast and easy diagnostic process. She explained,

"*I feel that the [baseline style] is just quicker and maybe a little bit easier, so, you can get the result faster, and it's also a little less stressful or overwhelming, because there's not so many extra messages in between, like all the explanations, on why they asked that question.*" (S34)

In non-emergency situations, most participants desired the explanations. Unlike a situation requiring a quick response, a non-emergency situation allowed participants to take the time to read the explanations and acquire enough information and details. As S7 stated,

"*If it's more of a serious emergency like with the appendix, then you might need to get to it really quickly, so you might not want as detailed of an explanation. In non-emergency situations, I'd probably want more talking. A cold and things like that are more chronic, so I'd probably want to know why I'm sick like this for a month or so, and might want to hear more details.*" (S7)

(3) Personalizing explanations for people with different personalities

Our findings also reported that people with different personalities had different explanation needs. For instance, S27 explained,

"*I think this is related to my personality. I guess every time I have to go to the doctor, if I have an issue where I don't know what's going on with me, I appreciate detailed responses very much and I like the explanations.*" (S27)

S27 enjoyed detailed medical information even when consulting with a doctor. This was one of the reasons why she desired explanations.

For the other participants who appreciated the baseline interface, their personalities, and lifestyle also played a role. As S20 put it

"*I [like] just the most direct questions and answers, so I can do it fast, when I'm busy... If there's a lot of words, then I will certainly get annoyed. I'm always busy. I'm always rushing, so I don't need to have a conversation with a bot. I don't need these explanations.*" (S20)

S20 preferred the baseline style because it was a quicker interaction. Her life was always busy and fast-paced, so she appreciated doing things quickly. Therefore, the explanations were useless and unnecessary to her.

In conclusion, our findings in this section indicated that the participants' needs for emotional support and explanations varied in different contexts. Generally speaking, their needs were related to different conversational stages, different emergency levels, and different personalities. These requirements also demonstrated the importance of the toggle button.

7 DISCUSSION

In this work, we reported that users desired emotional support, explanations, and efficiency when interacting with CSCs. We also directly built onto our prior work by proposing four types of conversational styles and testing their effects on the user experience. The results showed that user trust can be improved by providing explanations; the perceived level of transparency increased by offering either emotional support or explanations; and the perceived empathy of the technology

raised by offering emotional support. We did not find statistically significant differences in the constructs of satisfaction, likability, and usefulness across these four proposed conversational styles. Furthermore, we found that users desired different amounts of emotional support and explanation in different contexts to be provided. In this section, we discuss these findings and present design implications for the future conversational design of CSCs.

7.1 Beyond the Doctor-like Design

Our findings reflected our participants' desire not only for doctor-like professionalism but also for other features in their use of CSCs. The existing research mainly centered around the probing questions and the input functions of the conversational design [164, 166]. Few had explored user needs, such as how CSCs should provide doctor-like features and explanations. Based on our findings, we state that CSC conversational design should not only be doctor-like, but also consider the features that go beyond this likeness.

Our study first clarified what kind of human-like features CSCs should offer users. A stream of studies in the healthcare domain has underlined the importance of presenting a human-like, empathetic conversational style [94, 128]. These prior studies mainly focused on chatbots designed for the mental health domain [94, 128] and stressed the need for showing empathy [124]. Our findings complement this understanding as we found that users desire doctor-like professionalism and emotional support, which can improve the transparency and empathy of the system. To be more specific, most users appreciated the greetings, addresses, caring language, and treatment options.

In addition to the doctor-like features, we found explanations were desired by most users, which made the conversations robotic and less human-like from our participants' perspective. While providing explanations were found to positively affect the users' trust and the system's transparency, aligning with previous literature [150, 159], it differs from a doctor's consultation, as our participants reported. Prior research has found that people usually personify chatbots as human-like actors [91, 109], such as friends or family members [51]. Similarly, the CSCs were usually regarded as a virtual doctor. For example, Zuoshou Doctor and Ada claim that they are virtual doctors on their websites [7, 8]. Our results, however, revealed that instead of being completely designed as a virtual doctor, CSCs should also have non-doctor-like features. This is because different from actual doctors who have established authority, CSCs need to provide more evidence (e.g., explanations) to prove their reliability, as stated by our participants.

Providing both doctor-like emotional support and non-doctor-like explanations can thus render the CSC more transparent and enhance users' trust. It is noteworthy that the goal of providing emotional support and explanations is not to enhance the user's blind trust, but informed trust. Blind trust can lead to abuse of power and pose risk to the wellbeing of healthcare consumers [19], whereas informed trust, armed with proper information, can help people make more rational decisions. In clinical settings, patients desire shared decision-making, which means they want to become a part in making treatment decisions [147]. By providing information regarding outcomes, uncertainties, and education to healthcare consumers and involving them in the decision-making process, healthcare consumers could develop informed trust in medical professionals and improve their decision-making quality [19, 32, 34, 77]. Shared decision-making is therefore an ethical imperative when healthcare consumers make health-related decisions. Aligning with this line of research, our work reflected how CSCs could assist users in shared decision-making and enhance their informed trust by offering emotional support and explanations.

However, besides providing emotional support and explanations to enhance the transparency of CSCs, other features could also be considered when it comes to the ethical challenges that might

arise for health consumers. For example, our participants showed privacy concerns when using the designed CSC. Previous work has also reflected on potential risks of using CSCs, such as poor specificity [119]. Different from human doctors, who own medical expertise [83], lay users usually lack such knowledge to discern the accuracy and effectiveness of CSCs, which may lead to unintended harm to users, such as mortality [95]. However, few existing works have proposed practical solutions to solve ethical concerns or enacted new policies (e.g., data protection policy [63]). Our work thus highlighted that, in addition to emotional support and explanations, CSCs might also provide information regarding their data source [74], data privacy policies, the accuracy or uncertainty of their diagnoses, and potential risks. This would enable healthcare consumers to obtain better information to reduce the chance of blindly trusting CSCs. Future research on healthcare conversational agents should therefore emphasize the development of various policies to address transparency, safety, and data privacy concerns.

7.2 Trade-offs between Efficiency and the Offerings of Emotional Support/Explanations

Our findings stressed the user requirements for efficiency in the conversational design of CSCs. Efficiency is a critical measurement of the user experience with chatbots [90]. Previous literature has repeatedly demonstrated the value of a quick and efficient interaction when communicating with a chatbot [12, 26, 166]. Aligning with this line of research, our findings revealed that some users wished for an efficient and fast interaction with the CSC to acquire potential medical recommendations as soon as possible. The emotional support and explanations, while preferred by some users, may lead to long-winded conversations, and thus sacrifice efficiency. A few studies have examined issues of efficiency in CSCs. However, these studies mainly focused on the efficiency of the probing questions, such as whether too many pieces of information were presented at once [117] and if the probing questions were redundant [164]. It didn't consider the issues of efficiency generated from the provision of human-like responses or explanations. Our findings, nevertheless, illuminated that human-like features and transparency can come at the expense of efficiency. Therefore, we call for more attention to the trade-offs among user needs.

The first trade-off is between a human-like design and efficiency. Researchers have emphasized that human-like features should not sacrifice the system's efficiency [48, 49, 54]. Instead, a human-like design should augment the efficiency by asking intelligent questions and increasing the diversity of chatbot responses [70]. Our interview studies reported that some participants considered the caring language to be too much. While participants preferred having caring language, they thought its quantity and length should be reduced. For example, most participants complained about the emotional support offered after their every response. Although providing human-like features is widely recommended [24, 52, 155], in the case of CSCs, efficiency should not be sacrificed when providing emotional support, especially considering the health conditions the users may have can be emergent.

The second trade-off is figuring out how to balance the transparency and efficiency of the chatbot, as improving the system's transparency often damages its efficiency [68]. Little prior research has considered the trade-off between efficiency and transparency. Our findings reported that this kind of trade-off is critical to the user experience. For example, most of our participants complained that the provided explanations were overloaded in terms of quantity and length. Some participants also turned the toggle button on to avoid the explanations in the middle of the conversations.

The existing studies did not propose specific ways to improve the efficiency of healthcare chatbots given the importance of both human-like features and transparency, and we designed a toggle button on our chatbot's interface to bridge this research gap. We found that by using a toggle button, users could decide whether or not to receive emotional support responses and/or explanations.

When the number of emotional support responses and/or explanations was perceived as being too much, the user could simply turn on the toggle button. In this way, users could weigh multiple factors and reach a balance by themselves. Our study sheds light on how to ensure efficiency to improve the conversational design of healthcare chatbots, and we call for more research on the trade-offs between efficiency and human-like design/transparency.

7.3 User Control for Different Contexts

Our findings reported that not all CSC users like emotional support and explanations either throughout the conversation or in all situations; rather, they want to control how they receive both. While most prior studies call for providing explanations for algorithms [46, 66, 150] and empathy [36], the use of explanations [18] and empathetic language may not always be useful or desired for all users. Few existing studies have scrutinized the cost of explanations/emotional support responses or considered how to grant control to users. An existing study on symptom checkers highlighted the potential of providing customized explanations, arguing that a user may desire explanations with different levels of detail in different situations [150]. However, this study did not deeply explore what kind of explanations should be offered in what situation and how the CSC can be customized to grant users control. Providing users with control has usually been an important feature in the chatbot design [12, 163] and recommendation systems [99]. For example, users desired flexibility and freedom to customize the responses provided by the chatbot [163] and the level of detail explained. Building on prior discussions, our study stressed the need for user control and customization in the healthcare chatbot design. In the case of CSCs, we suggest a toggle button be designed to provide users with the control and freedom of receiving emotional support responses and explanations, especially with the following three kinds of concerns.

First, we found that in different conversational stages, our participants had different preferences for emotional support responses and explanations. At the conversation's opening, most participants preferred the emotional support responses. This echoes with most research on healthcare chatbots, which found that chatbots should show empathy and emotion to promote the interaction and mitigate users' moods [24, 38, 89]. At the end of the conversation, most users desired emotional support responses and explanations. In the middle of the conversation (the probing stage), however, we found that some users disliked the caring language (11 out of 34) and explanations (12 out of 34). Different from the past researchers who viewed a user's interaction with a chatbot as a whole [39, 97], we found users had nuanced needs for each interaction stage. Previous studies regarded the involvement of human-like characteristics as a binary, either-or choice [39, 97]. Our study uncovers that users have nuanced requirements for the different stages of the interaction, instead of an overall binary decision. Our study emphasizes that the users' needs for emotional support and explanations varied at the different interaction stages. Hence, our study stresses the importance of timing and user control when offering emotional support and explanations.

Second, we argued that we should pay attention to different contexts (emergent vs. non-emergent) when designing the conversation of a CSC. Matching with a prior study that stressed the importance of context and situation in chatbot conversational design [25], our study emphasizes that different conversational designs, rather than a universal design standard, should be considered in different contexts. In an emergency (e.g., a severe pain), most participants desired a more concise and direct conversation, since, in these moments, they usually valued the efficiency of the CSC; in non-emergency situations (e.g., a common cold), our participants placed a higher value on the detailed information. We argue that future research should keep potential contexts in mind when it comes to conversational design.

Third, our findings also revealed that emotional support and explanations are not beneficial for everyone. Previous literature emphasized that we need to tailor explanations to different users

[9, 65, 131], including the level of detail [57] and the complexity [131] of explanations. Resonating with these studies, we found that our participants held different attitudes toward the explanations: people who preferred detailed information desired the explanations, while people with a fast-paced lifestyle didn't. One-size-fits-all explanations may be suitable for certain users, but can be seen as information overload for other uses. Therefore, our study stresses that explanations should be tailored to different users with different personal characteristics and needs.

7.4 Design Implications

Drawing from these discussions, we propose the following design implications. Our findings indicated that, on the one hand, human-like features and explanations were desired by most participants. More specifically, human-like features and explanations positively affected the CSCs' perceived levels of trust, transparency, and empathy. With this line of thinking, CSC conversations mirror a clinical inquiry, have human-like features, and provide explanations. First, to resemble a doctor, the CSCs' probing questions can be designed with a flexible sentence structure and a progressive, one-by-one sequence. Second, the CSC may greet and address users at the beginning and deliver detailed treatment information at the end. Third, our study stressed the importance of the provision of sufficient support, i.e., emotional support, explanations, data source [74], data privacy policies, the accuracy or uncertainty of diagnoses, and potential risks to enhance the transparency of CSCs and allay ethical concerns. Broadly speaking, healthcare chatbots are usually opaque without disclosing how they are programmed, which may make the produced results hard to understand and cause low transparency [162]. Without sufficient information, health consumers may blindly trust the produced results, leading to potential ethical concerns. Ethical concerns are a key factor that can significantly influence users' attitudes toward healthcare chatbots [28]. Moving forward, healthcare chatbots might put in more effort to enhance their transparency by offering emotional support, explanations about algorithm mechanisms, privacy policies, accuracy, and warnings of risks. By doing so, healthcare chatbots can empower healthcare consumers to make informed decisions and reduce the potential risks of blindly trusting.

On the other hand, human-like features and explanations should be tailored in terms of different conversational stages, emergency levels, and individual personal preferences. A toggle button can be designed for users to control and decide whether/when they want human-like features and/or explanations. When users turn the toggle button on, they will not receive the human-like responses and explanations; when they turn the button off, the human-like responses and explanations continue to show up. In the future, more research on granting user control should be done. For example, users may want to customize the treatment information and other explanations' level of detail.

8 LIMITATIONS

Our study has several limitations. First, the sample size of the experimental study was relatively small. This may have caused a potential selection bias. Larger-scale studies can be conducted and BUS-11 (a Bot Usability Scale [15]) may be applied in the future to extensively verify user experiences of interaction with CSCs. In addition, the amount of material delivered by the emotional support and explanations was different. The user experience might have been influenced by the differing amounts of information. Second, our participants had a relatively high level of knowledge regarding symptom checkers and chatbots. This may limit our results' generalizability to users who are familiar with symptom checkers and chatbots. Third, our developed CSC applied simple rule-based algorithms. Further studies may improve on the design by using deep-learning and machine-learning models. Fourth, it would be meaningful to see how the user experience with the different conversational styles evolves in practice. Lastly, as our study is exploratory, we did

not test the use of the toggle button individually. The timing of the use of the toggle button can be tested in the future through further, more focused experimental design.

9 CONCLUSION

Our study investigated the user experiences of CSCs' conversational styles and explored potential design solutions using an experimental study. Through an interview study, we identified user needs (i.e., emotional support and doctor-like probing mode, explanations, and efficiency) for CSC app interactions. Following the interview findings, we designed and tested four types of conversational styles. The results shed light on two regards: (1) how do emotional support and explanations affect the users' perceptions of and experiences with the proposed CSC? and (2) when should the CSC provide emotional support and/or explanations? Our findings demonstrated the benefits of affording user control. Our study informs the future conversational design of healthcare chatbots with the consideration of improving user control for distinct contexts.

APPENDICES

A DEMOGRAPHIC INFORMATION

The demographic information of our participants for study 1 and study 2 are shown in Tables A.1 and A.2 separately.

Table A.1. Study 1 Demographic Information Table

Number	Age	Gender	CSC apps	Profession
U1	24	F	Ada, K Health	Master student
U2	29	F	Ada, K Health, Your.MD	PhD student
U3	27	M	K Health	PhD student
U4	24	M	K health, your.MD	Master student
U5	28	F	Ada, K Health	PhD student
U6	25	F	Ada, K Health	Master student
U7	26	F	K Health	PhD student
U8	26	F	Ada, K Health	Software engineer
U9	21	F	Ada, K Health, Your.MD	Master student
U10	30	M	K Health	PhD student
U11	54	F	K health	University staff
U12	24	F	Ada, K Health	Landscape Designer
U13	24	M	Ada	Undergraduate
U14	26	F	K health, Your.MD	Master student
U15	30	F	Ada, K health	Freelancer
U16	25	F	K health	PhD student
U17	29	F	Ada, K health	PhD student
U18	25	M	Ada, K Health, Your.MD	PhD student
U19	23	M	Ada, K Health	Undergraduate
U20	20	F	Ada	Undergraduate
U21	22	F	Ada, K Health	Undergraduate
U22	26	F	Babylon, K health	Master student
U23	26	M	Babylon, K health	Master student
U24	19	F	K health	Undergraduate
U25	20	M	K health	Undergraduate

Table A.2. Study 2 Demographic Information Table

Number	Age	Gender	Profession	Education
1	30	F	Student	Master's degree or above
2	27	F	Statistician	Master's degree or above
3	24	M	Software engineer	Master's degree or above
4	25	M	Student	Master's degree or above
5	41	F	Homemaker	Master's degree or above
6	52	F	Advocate worker	Bachelor's degree
7	72	M	Optometrist	Master's degree or above
8	47	F	Homemaker	High school
9	37	F	Preschool teacher	Master's degree or above
10	27	M	Nursing	Bachelor's degree
11	21	F	Student	Undergrad
12	25	M	Student	Master's degree or above
13	20	M	Student	Undergrad
14	40	M	Package handler	High school
15	30	M	Disabled	College Certificate
16	33	F	Self-employed	High school
17	71	F	Retired	Associates
18	30	M	Human Resources	Master's degree or above
19	40	F	Self-employed	High school
20	55	F	SPED Instructional Assistant	Bachelor's degree
21	19	F	Student	Bachelor's degree
22	31	M	Financial advisor	Bachelor's degree
23	50	M	Lawyer	Bachelor's degree
24	47	M	Plumber	Bachelor's degree
25	41	M	Physician	Bachelor's degree
26	42	M	Junior manager	Bachelor's degree
27	36	F	Homemaker	Master's degree or above
28	40	M	Dietatian	Master's degree or above
29	21	F	Student	Bachelor's degree
30	35	M	Unemployed	College Certificate
31	52	M	Foreman	Master's degree or above
32	21	F	Student	Bachelor's degree
33	45	Non binary	Baker	Bachelor's degree
34	26	F	Student	Master's degree or above

B CLINICAL DECISION TREES

The clinical decision trees we used to develop our CSC apps are shown in Tables B.1, B.2, B.3, and B.4 separately.

Table B.1. Decision Tree for the Training Session

	Symptom questions			
Rule	Headache	Cough	Fever	Decision
1	Y	Y	N	Acute bronchitis
2	Y	Y	Y	Flu
3	Y	N	Y	Flu
4	Y	N	N	Common cold

Table B.2. Decision Tree for Scenario 1

	Symptom questions				
Rule	Start time for runny nose	Sore throat	Cough	Fever	Decision
1	Less than one week	N	N	N	Viral sinusitus
2	One week to the last month	N	N	N	Hay fever
3	One month and more than one month	Y	N	N	Hay fever
3	One month and more than one month	N	N	N	Hay fever
4	Less than one week	Y	N	N	Common cold
5	One week to the last month	Y	N	N	Common cold
6	–	Y	Y	N	Common cold
7	Less than one week	Y	Y	Y	Common cold
8	One week to the last month	Y	Y	Y	Common cold
9	One month and more than one month	Y	Y	Y	Acute bronchitis
10	Less than one week	N	Y	Y/N	Common cold
11	One week to the last month	N	Y	N	Common cold
12	One month and more than one month	N	Y	Y/N	Acute bronchitis
13	One week to the last month	N	Y	Y	Common cold
14	Less than one week	N	N	Y	Viral sinusitus
15	One week to the last month	N	N	Y	Viral sinusitus
16	One month and more than one month	N	N	Y	Hay fever
17	Less than one week	Y	N	Y	Common cold
18	One week to the last month	Y	N	Y	Common cold
19	One month and more than one month	Y	N	Y	Hay fever

Table B.3. Decision Tree for Scenario 2

	Symptom questions					
Rule	Start time	Constipation	Dry skin	Weight gain	Feeling cold	Decision
1	Less than one week	N	N	N	N	Covid-19
2	One week to the last month	N	N	N	Y/N	Iron deficiency anaemia
3	One month and more than one month	N	N	N	N	Iron deficiency anaemia
4	–	Y	N	N	N	Constipation
5	–	Y	Y	Y/N	N	Hypothyroidism
6	–	Y	N	Y	Y/N	Hypothyroidism
7	Less than one week	Y	N	N	Y	Covid-19
8	One week to the last month	Y	N	N	Y	Hypothyroidism
9	One month and more than one month	Y/N	N	N	Y	Iron deficiency anaemia
10	–	Y	Y	Y/N	Y	Hypothyroidism
11	–	N	Y	Y/N	N	Hypothyroidism
12	–	N	Y	Y/N	Y	Hypothyroidism
13	Less than one week	N	N	Y	N	Covid-19
14	One week to the last month	N	N	Y	Y/N	Hypothyroidism
15	One month and more than one month	N	N	Y	Y/N	Hypothyroidism
16	Less than one week	N	N	Y/N	Y	Covid-19

Table B.4. Decision Tree for Scenario 3

Questions	Conditions	Choice	Score
Q1, Q2, Q3, Q4, Q5	Location for abdominal pain	Lower right	8
	Location for abdominal pain	Upper left, middle, upper right, lower left	0
	If have pain migration	Yes	3.2
	If have pain migration	No	0
	Did you have similar pain previously	Yes	0
	Did you have similar pain previously	No	1.5
	If psoas sign positive	Yes	2.38
	If psoas sign positive	No	0
	If have pain when pressure is removed abruptly	Yes	3.7
	If have pain when pressure is removed abruptly	No	0
Q6, Q7	If have vomiting	Yes	0.92
	If have vomiting	No	0
	If have pain before vomiting	Yes	2.76
	If have pain before vomiting	No	0
Q8	If lack or loss of appetite for food	Yes	1.27
	If lack or loss of appetite for food	No	0
Q9	If have nausea	Yes	0.9
	If have nausea	No	0
Q10	If have fever	Yes	1.9
	If have fever	No	0

If the total score is >= 10, the user will be diagnosed with acute appendicitis with a high probability.
If the total score is <= 5, the user will be diagnosed with acute appendicitis with a low probability.
If the total score is between 5 and 10, the user will be diagnosed with acute appendicitis with a moderate probability.

C POST-STAGE QUESTIONNAIRE

1. Trust: To measure if the users trust in the system.
   — *Q1: The medical diagnoses recommended by the symptom checker is reliable.*
   — *Q2: The symptom checker' responses in the interaction process are believable.*
2. Satisfaction: To measure if users were satisfied with the system.
   — *Q3: Overall, I am satisfied with the symptom checker.*
   — *Q4: I will recommend this symptom checker to my friends.*
   — *Q5: I will use this symptom checker in the future.*

3. Efficiency: To measure if users can acquire diagnoses efficiently.
   — *Q6: I was able to acquire medical suggestions quickly using this symptom checker.*
   — *Q7: The symptom checker helped me to make medical decisions faster.*
4. Likeability: To measure if users enjoy the use of the system.
   — *Q8: I like having conversations with this symptom checker.*
   — *Q9: Having conversations with this symptom checker is a pleasant experience.*
5. Effectiveness: To measure if the system can help users successfully.
   — *Q10: The symptom checker helps me to make better medical choices.*
   — *Q11: I find useful medical recommendations using this symptom checker.*
6. Transparency: To measure if users understand how the system works.
   — *Q12: The symptom checker provided sufficient information for me to make a good decision.*
   — *Q13:The symptom checker explained why medical diagnoses were recommended to me.*
   — *Q14: I understand why the medical diagnoses were recommended to me.*
7. Human likeness: To measure if the users consider the system as human-like.
   — *Q15: I feel the interaction with the symptom checker is similar to a human-like conversation.*
8. Emotional support: To measure if the system can provide emotional support to users.
   — *Q16: I feel the symptom checker provided sufficient emotional support.*
   — *Q17: When interacting with the symptom checker, I feel it seemed to understand me.*
9. NASA TLX
   — *Q18: Mental demand: how mentally demanding was the task?*
   — *Q19: Performance: how How successful were you in accomplishing what you were asked to do?*
   — *Q20: Effort: how hard did you have to work to accomplish your level of performance?*
   — *Q21: Frustration: after the task, how insecure, discouraged, irritated, stressed, and annoyed were you?*

D STUDY 1 INTERVIEW QUESTIONS

1. What kinds of symptom checkers have you used? What are they?
2. Why did you start to use them?
3. Did you achieve your goals? What kind of goals?
4. Which checker do you prefer?
5. Why do you prefer this one to others?
6. What kinds of functions does it provide??
7. Do you feel these functions are enough for you to understand and diagnose your conditions?
8. What functions do you think are most useful? Why?
9. What functions would you like to have? Why?
10. How do you feel about having conversations with these symptom checkers?
11. How do you perceive the relation between questions and diagnostic results?
12. How do you feel about the order of questions?
13. What do you think of the speed of conversations?
14. What forms of conversations you prefer when talking with the chatbot (e.g., text, voice, avatar, human-like?)
15. What benefits do you perceive in symptom checkers?
16. What negative consequences are associated with using symptom checkers?
17. What challenges do you encounter when using symptom checkers?
18. What was the best thing about your experience of using the symptom checker?

E STUDY 2 INTERVIEW QUESTIONS

1. Which conversational style do you prefer?
2. Why do you prefer this one to others?
3. Compared to conversations with a real human, what do you think of the conversations with the symptom checker? Any difference?
4. How do you feel about the conversation of these four symptom checkers? What features you prefer or not?
5. I noticed that you pressed the button (at XX time) could you please explain the reason?
6. Why did/didn't you press the button (at that time) during the interaction?
7. At what time/stage of the conversation do you want extra information (caring words or explanations)? Why?
8. When do you think these extra information caring words or explanation are too overloaded or suitable?)
9. Do you like the interface with caring words/explanation? Why?
10. Do you think the interface with caring words/explanations make you feel uncomfortable? Why?
11. How do you feel about the toggle button?
12. What challenges did you encounter when having conversation with the chatbots?

ACKNOWLEDGMENTS

We thank all the participants for their time and for sharing their experiences. We also thank the reviewers for their insightful suggestions.

REFERENCES

[1] 2020. *8 things to know about online symptom Checker Applications.* Retrieved from https://www.beckershospitalreview.com/healthcare-information-technology/8-things-to-know-about-online-symptom-checker-applications.html. Accessed 10-1-2020.
[2] 2021. What is conversation design? Retrieved from https://developers.google.com/assistant/conversation-design/what-is-conversation-design. Accessed 8-1-2021.
[3] 2022. *Better Information. Better Health.* Retrieved from https://www.webmd.com/. Accessed 1-30-2022.
[4] 2022. *Buoy Health.* Retrieved from https://www.buoyhealth.com/. Accessed 1-30-2022.
[5] 2022. *The Chatbot Will SeeYou Now - on any platform.* Retrieved from https://assets.ctfassets.net/iqo3fk8od6t9/08pX2xBSPIjHltvLkv6YG/c3f139de4d6777d4af934c53b8dae204/Healthily-AI-whitepaper_The-chatbot-will-see-you-now.pdf. Accessed 1-30-2022.
[6] 2022. *Flow XO: Premier AI Online Chatbot Software.* Retrieved from https://flowxo.com/. Accessed 1-30-2022.
[7] 2022. *Health. powered by Ada.* Retrieved from https://ada.com. Accessed 1-30-2022.
[8] 2022. *Zuoshou Doctor in the Apple App Store.* Retrieved from https://apps.apple.com/cn/app. Accessed 1-30-2022.
[9] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2021. Reason explanation for encouraging behaviour change intention. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems.* 68–77.
[10] Alfredo Alvarado. 1986. A practical score for the early diagnosis of acute appendicitis. *Annals of Emergency Medicine* 15, 5 (1986), 557–564.
[11] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189.
[12] 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.
[13] W. F. Bail. 2014. The complete guide to communication skills in clinical practice. *Program.–2014.–34 p* (2014).
[14] Eileen Bendig, Benjamin Erb, Lea Schulze-Thuesing, and Harald Baumeister. 2022. The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. *Verhaltenstherapie* 32, 1 (2022), 64–76.

[15] Simone Borsci, Martin Schmettow, Alessio Malizia, Alan Chamberlain, and Frank Van Der Velde. 2022. A confirmatory factorial analysis of the Chatbot Usability Scale: A multilanguage validation. *Personal and Ubiquitous Computing* (2022), 1–14.
[16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.
[17] Susan E. Brennan. 1990. Conversation as direct manipulation: An iconoclastic view. *The Art of Human-computer Interface Design* (1990), 393–404.
[18] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*. 169–178.
[19] Michael Calnan and Rosemary Rowe. 2007. Trust and health care. *Sociology Compass* 1, 1 (2007), 283–308.
[20] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a chatbot for digital counselling. In *Proceedings of the 31st International BCS Human Computer Interaction Conference*. 1–7.
[21] Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-adapted Interaction* 13, 1 (2003), 89–132.
[22] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. Can a humorous conversational agent enhance learning experience and outcomes?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[23] Alan Chamberlain, Andy Crabtree, Tom Rodden, Matt Jones, and Yvonne Rogers. 2012. Research in the wild: Understanding'in the wild'approaches to design and development. In *Proceedings of the Designing Interactive Systems Conference*. 795–796.
[24] Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 102–109.
[25] Meira Chefitz, Jesse Austin-Breneman, and Nigel Melville. 2018. Designing conversational interfaces to reduce dissonance. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. 219–223.
[26] Zhifa Chen, Yichen Lu, Mika P. Nieminen, and Andrés Lucero. 2020. Creating a chatbot for and with migrants: Chatbot personality drives co-design activities. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 219–230.
[27] Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2020. Towards explainable conversational recommendation. In *Proceedings of the IJCAI*. 2994–3000.
[28] Yang Cheng, Chenxing Xie, Yanding Wang, and Hua Jiang. 2023. Chatbots and health: Mental health. *The International Encyclopedia of Health Communication*. John Wiley & Sons.
[29] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[30] Hyojin Chin and Mun Yong Yi. 2019. Should an agent be ignoring it? A study of verbal abuse types and conversational agents' response styles. In *Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
[31] Janghee Cho and Emilee Rader. 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
[32] William C. Cockerham. 2016. *The New Blackwell Companion to Medical Sociology*. John Wiley & Sons.
[33] William Jay Conover. 1999. *Practical Nonparametric Statistics*. John Wiley & Sons.
[34] Angela Coulter and Alf Collins. 2011. Making shared decision-making a reality. *London: King's Fund* 621 (2011).
[35] Sebastian Cross, Ahmed Mourad, Guido Zuccon, and Bevan Koopman. 2021. Search engines vs. symptom checkers: A comparison of their effectiveness for online health advice. In *Proceedings of the Web Conference 2021*. 206–216.
[36] Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic chatbot response for medical assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–3.
[37] Khairul Azhar Mat Daud, Nik Zulkarnaen Khidzir, Ahmad Rasdan Ismail, and Fadhilahanim Aryani Abdullah. 2018. Validity and reliability of instrument to measure social media skills among small and medium entrepreneurs at Pengkalan Datu River. *International Journal of Development and sustainability* 7, 3 (2018), 1026–1037.
[38] Mauro De Gennaro, Eva G. Krumhuber, and Gale Lucas. 2020. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology* 10 (2020), 3061.
[39] Virginie Demeure, Radosław Niewiadomski, and Catherine Pelachaud. 2011. How is believability of a virtual agent related to warmth, competence, personification, and embodiment? *Presence* 20, 5 (2011), 431–448.

[40] Stephan Diederich, Tim-Benjamin Lembcke, Alfred Benedikt Brendel, and Lutz M. Kolbe. 2021. Understanding the impact that response failure has on how users perceive anthropomorphic conversational service agents: Insights from an online experiment. *AIS Transactions on Human-Computer Interaction* 13, 1 (2021), 82–103.

[41] Stephan Diederich, Sascha Lichtenberg, Alfred Benedikt Brendel, and Simon Trang. 2019. Promoting sustainable mobility beliefs with persuasive and anthropomorphic design: Insights from an experiment with a conversational agent. (2019).

[42] Laury Donkelaar. 2018. *How Human Should a Chatbot be?: The Influence of Avatar Appearance and Anthropomorphic Characteristics in the Conversational Tone Regarding Chatbots in Customer Service Field.* Master's thesis. University of Twente.

[43] Ahmed Fadhil and Silvia Gabrielli. 2017. Addressing challenges in promoting healthy lifestyles: The al-chatbot approach. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare.* 261–265.

[44] Ahmed Fadhil and Gianluca Schiavo. 2019. Designing for health chatbots. arXiv:1902.09022. Retrieved from https://arxiv.org/abs/1902.09022.

[45] Ahmed Fadhil, Yunlong Wang, and Harald Reiterer. 2019. Assistive conversational agent for health coaching: A validation study. *Methods of Information in Medicine* 58, 01 (2019), 009–023.

[46] Xiangmin Fan, Daren Chao, Zhan Zhang, Dakuo Wang, Xiaohua Li, and Feng Tian. 2021. Utilization of self-diagnosis health chatbots in real-world settings: Case study. *Journal of Medical Internet Research* 23, 1 (2021), e19928.

[47] Stefano Federici, Maria Laura de Filippis, Maria Laura Mele, Simone Borsci, Marco Bracalenti, Giancarlo Gaudino, Antonello Cocco, Massimo Amendola, and Emilio Simonetti. 2020. Inside pandora's box: A systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs. *Disability and Rehabilitation: Assistive Technology* 15, 7 (2020), 832–837.

[48] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? An exploratory interview study. In *Proceedings of the International Conference on Internet Science.* Springer, 194–208.

[49] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: User experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces.* 1–9.

[50] Hamish Fraser, Enrico Coiera, and David Wong. 2018. Safety of patient-facing digital symptom checkers. *The Lancet* 392, 10161 (2018), 2263–2264.

[51] Yang Gao, Zhengyu Pan, Honghao Wang, and Guanling Chen. 2018. Alexa, my love: Analyzing reviews of Amazon echo. In *Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation.* IEEE, 372–380.

[52] Radhika Garg and Subhasree Sengupta. 2020. Conversational technologies for in-home learning: Using co-design to understand children's and parents' perspectives. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–13.

[53] Stella George. 2019. From sex and therapy bots to virtual assistants and tutors: How emotional should artificially intelligent agents be?. In *Proceedings of the 1st International Conference on Conversational User Interfaces.* 1–3.

[54] Ulrich Gnewuch, Stefan Morana, Marc T. P. Adam, and Alexander Maedche. 2018. Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. In *Proceedings of the 26th European Conference on Information Systems: Beyond Digitization-Facets of Socio-Technical Change.* 143975.

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

[56] Julio Guerra-Hollstein, Jordan Barria-Pineda, Christian D. Schunn, Susan Bull, and Peter Brusilovsky. 2017. Fine-grained open learner models: Complexity versus support. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization.* ACM, 41–49.

[57] Mouadh Guesmi, Mohamed Amine Chatti, Laura Vorgerd, Shoeb Joarder, Shadi Zumor, Yiqi Sun, Fangzheng Ji, and Arham Muslim. 2021. On-demand personalized explanation for transparent recommendation. In *Proceedings of the Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* 246–252.

[58] D. Mike Hardin Jr. 1999. Acute appendicitis: Review and update. *American Family Physician* 60, 7 (1999), 2027.

[59] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Proceedings of the Advances in Psychology.* 139–183.

[60] Sandra Hauser-Ulrich, Hansjörg Künzli, Danielle Meier-Peterhans, and Tobias Kowatsch. 2020. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): Pilot randomized controlled trial. *JMIR mHealth and uHealth* 8, 4 (2020), e15806.

[61] Simon Hoermann, Kathryn L. McCabe, David N. Milne, and Rafael A. Calvo. 2017. Application of synchronous text-based dialogue systems in mental health interventions: Systematic review. *Journal of Medical Internet Research* 19, 8 (2017), e7023.

[62] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923. Retrieved from https://arxiv.org/abs/1712.09923.
[63] Thi Vuong Hua and Yanhong Hou. 2020. Factors that influence the intention to use self-diagnosis Apps in Vietnam. *Journal of Health, Medicine and Nursing* 72 (2020), 47–56.
[64] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In *Proceedings of the Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 946–956.
[65] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: Investigating social facilitation in human-machine team creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[66] Youjin Hwang, Taewan Kim, Junhan Kim, Joonhwan Lee, and Hwajung Hong. 2018. Leveraging challenges of an algorithm-based symptom checker on user trust through explainable AI. (2018).
[67] Youjin Hwang, Donghoon Shin, Sion Baek, Bongwon Suh, and Joonhwan Lee. 2021. Applying the Persona of user's family member and the doctor to the conversational agents for healthcare. In *CHI 2020 Workshop on Conversational Agents for Health and Wellbeing*.
[68] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1, 11 (2019), 517–521.
[69] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the use of a context view for chatbots. In *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*. 1–6.
[70] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 895–906.
[71] Mlađan Jovanović, Marcos Baez, and Fabio Casati. 2020. Chatbots as conversational healthcare services. *IEEE Internet Computing* 25, 3 (2020), 44–51.
[72] Anjali Khurana, Parsa Alamzadeh, and Parmit K. Chilana. 2021. ChatrEx: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. In *Proceedings of the 2021 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 1–11.
[73] Junhan Kim, Yoojung Kim, Byungjoon Kim, Sukyung Yun, Minjoon Kim, and Joongseek Lee. 2018. Can a machine tend to teenagers' emotional needs? A study with conversational agents. In *Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[74] Junhan Kim, Jana Muhic, Lionel Peter Robert, and Sun Young Park. 2022. Designing chatbots with black americans with chronic conditions: Overcoming challenges against COVID-19. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[75] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[76] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[77] Tiffany Hyun-Jin Kim. 2014. Challenges of establishing trust in online entities and beyond. In *Proceedings of the 4th International Workshop on Trustworthy Embedded Devices*. 49–49.
[78] A Baki Kocaballi, Juan C. Quiroz, Liliana Laranjo, Dana Rezazadegan, Rafal Kocielnik, Leigh Clark, Q. Vera Liao, Sun Young Park, Robert J. Moore, and Adam Miner. 2020. Conversational agents for health and wellbeing. In *Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
[79] Ahmet Baki Kocaballi, Juan C. Quiroz, Dana Rezazadegan, Shlomo Berkovsky, Farah Magrabi, Enrico Coiera, and Liliana Laranjo. 2020. Responses of conversational agents to health and lifestyle prompts: Investigation of appropriateness and presentation structures. *Journal of Medical Internet Research* 22, 2 (2020), e15823.
[80] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User preferences for hybrid explanations. In *Proceedings of the 11th ACM Conference on Recommender Systems*. 84–88.
[81] Sari Kujala, Iiris Hörhammer, Riitta Hänninen-Ervasti, Tarja Heponiemi. 2020. Health professionals' experiences of the benefits and challenges of online symptom checkers. In *Proceedings of the MIE*. 966–970.
[82] Raina Langevin, Ross J. Lordon, Thi Avrahami, Benjamin R. Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[83] Stephen R. Latham. 2002. Medical professionalism. *Mt Sinai J Med* 69, 2002 (2002), 363–9.
[84] 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.

[85] David M. Levine and Ateev Mehrotra. 2021. Assessment of diagnosis and triage in validated case vignettes among nonphysicians before and after internet search. *JAMA Network Open* 4, 3 (2021), e213287–e213287.
[86] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[87] Q. Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable ai user experiences. arXiv:2104.03483. Retrieved from https://arxiv.org/abs/2104.03483.
[88] Yuting Liao and Jiangen He. 2020. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 430–442.
[89] Bingjie Liu and S. Shyam Sundar. 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (2018), 625–636.
[90] Mingming Liu, Qicheng Ding, Yu Zhang, Guoguang Zhao, Changjian Hu, Jiangtao Gong, Penghui Xu, Yu Zhang, Liuxin Zhang, and Qianying Wang. 2020. Cold comfort matters-how channel-wise emotional strategies help in a customer service chatbot. In *Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.
[91] Irene Lopatovska and Harriet Williams. 2018. Personification of the Amazon Alexa: BFF or a mindless companion. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*. 265–268.
[92] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA" The Gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5286–5297.
[93] Deborah Lupton and Annemarie Jutel. 2015. "It's like having a physician in your pocket!" A critical analysis of self-diagnosis smartphone apps. *Social Science and Medicine* 133 (2015), 128–135.
[94] Syaheerah Lebai Lutfi, Fernando Fernández-Martínez, Jaime Lorenzo-Trueba, Roberto Barra-Chicote, and Juan Manuel Montero. 2013. I feel you: The design and evaluation of a domotic affect-sensitive spoken conversational agent. *Sensors* 13, 8 (2013), 10519–10538.
[95] Carl Macrae. 2019. Governing the safety of artificial intelligence in healthcare. *BMJ Quality and Safety* 28, 6 (2019), 495–498.
[96] Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. 2019. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*. 111–119.
[97] Daniel McDuff and Mary Czerwinski. 2018. Designing emotionally sentient agents. *Communications of the ACM* 61, 12 (2018), 74–83.
[98] Raphael Meyer von Wolff, Sebastian Hobert, and Matthias Schumann. 2019. How may I help you?–State-of-the-art and open research questions for chatbots at the digital workplace. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
[99] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 397–407.
[100] Michael L. Millenson, Jessica L. Baldwin, Lorri Zipperer, and Hardeep Singh. 2018. Beyond Dr. Google: The evidence on consumer-facing digital tools for diagnosis. *Diagnosis* 5, 3 (2018), 95–105.
[101] Stephen Miller, Stephen Gilbert, Vishaal Virani, and Paul Wicks. 2020. Patients' utilization and perception of an artificial intelligence–based symptom assessment and advice technology in a British primary care waiting room: Exploratory pilot study. *JMIR Human Factors* 7, 3 (2020), e19713.
[102] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[103] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems With Applications* 129 (2019), 56–67. DOI: https://doi.org/10.1016/j.eswa.2019.03.054
[104] Nika Mozafari, Welf H. Weiger, and Maik Hammerschmidt. 2020. The chatbot disclosure dilemma: Desirable and undesirable effects of disclosing the non-human identity of chatbots. In *Proceedings of the ICIS*.
[105] Andreea Muresan and Henning Pohl. 2019. Chats with bots: Balancing imitation and engagement. In *Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
[106] Fedelucio Narducci, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2018. Improving the user experience with a conversational recommender system. In *Proceedings of the International Conference of the Italian Association for Artificial Intelligence*. Springer, 528–538.
[107] Hien Nguyen and Judith Masthoff. 2009. Designing empathic computers: The effect of multimodal empathic feedback using animated agent. In *Proceedings of the 4th International Conference on Persuasive Technology*. 1–9.
[108] Svetlana Nikitina, Sara Callaioli, and Marcos Baez. 2018. Smart conversational agents for reminiscence. In *Proceedings of the 2018 IEEE/ACM 1st International Workshop on Software Engineering for Cognitive Services*. IEEE, 52–57.

[109] Young Hoon Oh, Kyungjin Chung, Da Young Ju. 2020. Differences in interactions with a conversational agent. *International Journal of Environmental Research and Public Health* 17, 9 (2020), 3189.
[110] Stefan Olafsson, Teresa K. O'Leary, and Timothy W. Bickmore. 2020. Motivating health behavior change with humorous virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.
[111] Ulrika Östlund, Lisa Kidd, Yvonne Wengström, and Neneh Rowa-Dewar. 2011. Combining qualitative and quantitative research within mixed method research designs: A methodological review. *International Journal of Nursing Studies* 48, 3 (2011), 369–383.
[112] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017), 1–40.
[113] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery* 24, 3 (2012), 555–583.
[114] Hyanghee Park and Joonhwan Lee. 2020. Can a conversational agent lower sexual violence victims' burden of self-disclosure?. In *Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
[115] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of Medical Internet Research* 21, 4 (2019), e12231.
[116] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A model of social explanations for a conversational movie recommendation system. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 135–143.
[117] Sanjana Ponnada. 2020. Reimagining the COVID-19 digital experience: The value of user empowerment and accessibility in risk communication. In *Proceedings of the 38th ACM International Conference on Design of Communication*. 1–3.
[118] Ashish Viswanath Prakash and Saini Das. 2020. Intelligent conversational agents in mental healthcare services: A thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems* 12, 2 (2020), 1.
[119] Rebecca Anhang Price, Daniel Fagbuyi, Racine Harris, Dan Hanfling, Frederick Place, Todd B. Taylor, and Arthur L. Kellermann. 2013. Feasibility of web-based self-triage by parents of children with influenza-like illness: A cautionary tale. *JAMA Pediatrics* 167, 2 (2013), 112–118.
[120] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[121] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[122] Piyush Ranjan, Archana Kumari, and Avinash Chakrawarty. 2015. How can doctors improve their communication skills? *Journal of Clinical and Diagnostic Research: JCDR* 9, 3 (2015), JE01.
[123] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI.. In *Proceedings of the IUI Workshops*, Vol. 2327. 38.
[124] James Ross and Chris Watling. 2017. Use of empathy in psychiatric practice: Constructivist grounded theory study. *BJPsych Open* 3, 1 (2017), 26–33.
[125] Hyeyoung Ryu, Soyeon Kim, Dain Kim, Sooan Han, Keeheon Lee, and Younah Kang. 2020. Simple and steady interactions win the healthy mentality: Designing a chatbot service for the elderly. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.
[126] Jane S. Saczynski, Jorge Yarzebski, Darleen Lessard, Frederick A. Spencer, Jerry H. Gurwitz, Joel M. Gore, and Robert J. Goldberg. 2008. Trends in prehospital delay in patients with acute myocardial infarction (from the Worcester Heart Attack Study). *The American Journal of Cardiology* 102, 12 (2008), 1589–1594.
[127] Marcel Salathé, Thomas Wiegand, and Markus Wenzel. 2018. Focus group on artificial intelligence for health. arXiv:1809.04797. Retrieved from https://arxiv.org/abs/1809.04797.
[128] Samiha Samrose, Kavya Anbarasu, Ajjen Joshi, and Taniya Mishra. 2020. Mitigating Boredom Using An Empathetic Conversational Agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.
[129] Briane Paul V. Samson and Yasuyuki Sumi. 2020. Are two heads better than one? Exploring two-party conversations for car navigation voice guidance. In *Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.
[130] Shruti Sannon, Brett Stoll, Dominic DiFranzo, Malte Jung, and Natalya N. Bazarova. 2018. How personification and interactivity influence stress-related disclosures to conversational agents. In *Proceedings of the Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 285–288.
[131] Johanes Schneider and Joshua Handali. 2019. Personalized explanation in machine learning: A conceptualization. arXiv:1901.00770. Retrieved from https://arxiv.org/abs/1901.00770.

[132] Jessica Schroeder, Chelsey Wilkes, Kael Rowan, Arturo Toledo, Ann Paradiso, Mary Czerwinski, Gloria Mark, and Marsha M. Linehan. 2018. Pocket skills: A conversational mobile web app to support dialectical behavioral therapy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–15.

[133] Kristen M. Scott, Simone Ashby, and Julian Hanna. 2020. "Human, all too human": NOAA weather radio and the emotional impact of synthetic voices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[134] Anna-Maria Seeger, Jella Pfeiffer, and Armin Heinzl. 2021. Texting with human-like conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems* 22, 4 (2021), 8.

[135] Hannah L. Semigran, Jeffrey A. Linder, Courtney Gidengil, and Ateev Mehrotra. 2015. Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ* 351 (2015).

[136] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. Face value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[137] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[138] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.

[139] Kacper Sokol and Peter A. Flach. 2018. Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *Proceedings of the IJCAI*. 5868–5870.

[140] Yuan Sun and S. Shyam Sundar. 2022. Exploring the effects of interactive dialogue in improving user control for explainable online symptom checkers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[141] Nina Svenningsson and Montathar Faraon. 2019. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*. 151–161.

[142] Ella Tallyn, Hector Fried, Rory Gianni, Amy Isard, and Chris Speed. 2018. The ethnobot: Gathering ethnographies in the age of IoT. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[143] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*. 42–51.

[144] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4–5 (2012), 399–439.

[145] Diana-Cezara Toader, Grațiela Boca, Rita Toader, Mara Măcelaru, Cezar Toader, Diana Ighian, and Adrian T. Rădulescu. 2020. The effect of social presence and chatbot errors on trust. *Sustainability* 12, 1 (2020), 256.

[146] David Traum. 2017. Computational approaches to dialogue. *The Routledge Handbook of Language and Dialogue. Taylor and Francis* (2017), 143–161.

[147] Juhi Tripathi, Shalabh Rastogi, and Ashok Jadon. 2019. Changing doctor patient relationship in India: A big concern. *Int. J. Commun. Med. Public Health* 6 (2019), 3160–3164.

[148] Chun-Hua Tsai and Peter Brusilovsky. 2021. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* 31, 2021 (2021), 591–627.

[149] Chun-Hua Tsai and Peter Brusilovsky. 2021. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* 31, 3 (2021), 591–627.

[150] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[151] James M. Wagner, W. Paul McKinney, and John L. Carpenter. 1996. Does this patient have appendicitis? *Jama* 276, 19 (1996), 1589–1594.

[152] H. Kenneth Walker, W. Dallas Hall, and J. Willis Hurst. 1990. Clinical methods: The history, physical, and laboratory examinations. 3rd ed., Butterworths.

[153] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[154] Weiyu Wang and Keng Siau. 2018. Living with artificial intelligence–developing a theory on trust in health chatbots. In *Proceedings of the 16th Annual Pre-ICIS Workshop on HCI Research in MIS*. Association for Information Systems San Francisco, CA.

[155] Justin D. Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: Teaching strategies for successful human-agent interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 448–459.

[156] Anna-Jasmin Wetzel, Roland Koch, Christine Preiser, Regina Müller, Malte Klemmt, Robert Ranisch, Hans-Jörg Ehni, Urban Wiesing, Monika A. Rieger, Tanja Henking. 2022. Ethical, legal, and social implications of symptom checker apps in primary health care (CHECK. APP): Protocol for an interdisciplinary mixed methods study. *JMIR Research Protocols* 11, 5 (2022), e34026.

[157] Daricia Wilkinson, Öznur Alkan, Q. Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P. Knijnenburg, and Elizabeth Daly. 2021. Why or why not? The effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems* 39, 4 (2021), 1–21.

[158] Jennifer Wisdom and John W. Creswell. 2013. Mixed methods: Integrating quantitative and qualitative data collection and analysis while studying patient-centered medical home models. *Rockville: Agency for Healthcare Research and Quality* (2013).

[159] Claire Woodcock, Brent Mittelstadt, Dan Busbridge, Grant Blank. 2021. The impact of explanations on layperson trust in artificial intelligence–driven symptom checker apps: Experimental study. *Journal of Medical Internet Research* 23, 11 (2021), e29386.

[160] Ziang Xiao, Michelle X. Zhou, and Wat-Tat Fu. 2019. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 437–447.

[161] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction* 27, 3 (2020), 1–37.

[162] Lu Xu, Leslie Sanders, Kay Li, and James C. L. Chow. 2021. Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR Cancer* 7, 4 (2021), e27850.

[163] Xi Yang and Marco Aurisicchio. 2021. Designing conversational agents: A self-determination theory approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[164] Yue You and Xinning Gui. 2020. Self-diagnosis through AI-enabled chatbot-based symptom checkers: User experiences and design considerations. In *Proceedings of the AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association, 1354.

[165] Yue You, Yubo Kou, Xianghua Ding, and Xinning Gui. 2021. The medical authority of AI: A study of AI-enabled consumer-facing health technology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[166] Jennifer Zamora. 2017. I'm sorry, dave, I'm afraid I can't do that: Chatbot perception and expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 253–260.

Received 16 August 2022; revised 15 January 2023; accepted 20 February 2023