5-7-2021

Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers

Chun-Hua Tsai

Yue You

Xinning Gui

Yubo Kou

John M. Carroll

Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers

Chun-Hua Tsai , Yue You, Xinning Gui, Yubo Kou, John M. Carroll
College of Information Sciences and Technology, Pennsylvania State University

ABSTRACT

Online symptom checkers (OSC) are widely used intelligent systems in health contexts such as primary care, remote healthcare, and epidemic control. OSCs use algorithms such as machine learning to facilitate self-diagnosis and triage based on symptoms input by healthcare consumers. However, intelligent systems' lack of transparency and comprehensibility could lead to unintended con- sequences such as misleading users, especially in high-stakes areas such as healthcare. In this paper, we attempt to enhance diagnostic transparency by augmenting OSCs with explanations. We first conducted an interview study (N=25) to specify user needs for explanations from users of existing OSCs. Then, we designed a COVID-19 OSC that was enhanced with three types of explanations. Our lab-controlled user study (N=20) found that explanations can significantly improve user experience in multiple aspects. We discuss how explanations are interwoven into conversation flow and present implications for future OSC designs.

CCS CONCEPTS

• **Human-centered computing** → **Interaction paradigms**; **Interaction design process and methods**.

KEYWORDS

Symptom Checker; COVID-19; Explanation; Health; Transparency

ACM Reference Format:

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 *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3411764.3445101

INTRODUCTION

Online symptom checkers (OSCs) are intelligent systems that apply algorithms and machine learning approaches (e.g., clinical decision tree) to help patients with self-diagnosis or self-triage [62]. As healthcare consumers increasingly rely on the Internet for health information [9], OSCs are widely used in various health contexts. For example, patients may use OSCs to check their early symptoms as well as gain more knowledge about their body conditions before a doctor's visit [53]. OSCs help healthcare consumers identify the appropriate care level and services and whether they need medical attention from healthcare providers [81]. A 2019 survey shows that 22% of US health consumers have used a symptom checker app in the past 12 months [22]. In major app stores, most popular symptom checkers (e.g., Babylon, Ada, Your.MD, K health, WebMD) have been downloaded tens of millions of times [46]. Developers of AI-powered symptom checkers promise various benefits such as providing quality diagnosis [10] and reducing unnecessary visits and tests [34, 62]. Certain healthcare providers like the Sutter Health network even encourage patients to self-diagnose with OSCs, be- cause they can support patients *"at the first onset of symptoms"* [25]. OSCs are not designed to replace professional diagnosis. Instead, it can engage the patient early in the healthcare cycle to strengthen the healthcare quality and lower the cost and ease the burden for the professional [22]. The usage of OSCs can further benefit other health contexts like remote health and epidemic control [48]. During the COVID-19 pandemic in 2020, the government, industry, and large OSC developers such as Ada and Babylon released dedicated Coronavirus Self-Checker [7, 24, 32].

Despite these promises, OSCs also have limits. Unlike real health- care professionals, most OSCs do not *explain why* the OSCs provide such diagnosis or *why* a patient falls into a disease classification. OSCs' data and clinical decision model are usually neither trans- parent nor comprehensible to lay users, which may lead to less user trust in the system [54]. Transparency has been discussed in the scholarship of *explainable AI (XAI)* [21], which focuses on explaining and justifying the outcomes of AI-driven decisions or recommendations [13, 36, 40, 51, 72]. Explainable interfaces have been shown to be able to improve user experience, i.e., trust, under- standability, and satisfactions [18, 70]. For example, *explanations* can help users to understand reasoning process of recommendation

models [27], and how their actions can impact and control the system [31], which in turn contributes to system *inspectability* [36] or *transparency* of the recommendation process [39, 40, 70]. However, in the high-stakes domain of healthcare, little attention has been paid to the transparency issue of OSCs.

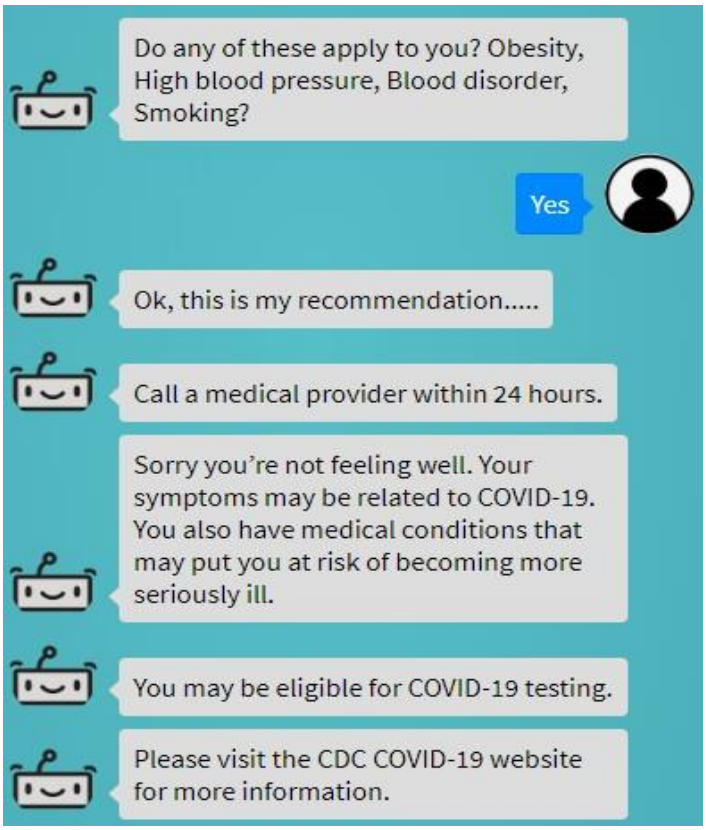

Figure 1: The example of diagnostic and triage recommendation of the COVID-19 OSC.

In this paper, we explore how explanations could be used to promote diagnostic transparency of online symptom checkers. To answer this question, we conducted two studies: First, we conducted an interview study to explore *what explanation needs exist in the existing use of OSCs*. Second, informed by the first study's results, we used a design study to investigate *how explanations affect the user perception and user experience with OSCs.* We designed an OSC in the context of COVID-19 and tested it with three styles of explanations. Combining results from the two studies, we found that explanations can significantly improve overall user experiences in multiple aspects, such as trust, transparency perception, and learning. Besides, we showed that by interweaving explanations into

conversation flow, OSC could facilitate users' comprehension of the diagnostics in a dynamic and timely fashion.

Our contributions to HCI are three-fold: First, we contributed empirical insights into user experiences with explanations in health- care domain. Second, we derived conceptual insights into OSC transparency. Third, we proposed design implications for improving transparency in healthcare technologies, and especially explanation design in conversational agents.

RELATED WORK

Symptom Checker

Many private companies and public organizations (e.g., the National Health Service; the American Academy Pediatrics) have launched OSCs. A 2019 survey shows that 22% of the U.S. health consumers had used an OSC app in the past 12 months [22]. OSCs are usually built upon expert systems that consist of a medical knowledge base and an inference engine [19, 34]. The medical knowledge base depicts the probabilistic relationship between symptoms and diseases, and the inference engine collects patient information, formulate symptom inquiries, and perform diagnosis based on the user information and the medical knowledge base [34]. The outputs are possible diseases that the patient may have and triage advice regarding whether the patient needs to use a medical visit.

OSCs have been claimed to have potential benefits. For instance, OSCs may reduce the numbers of medical visits and thus save physicians' and patients' time and financial cost [16, 62]. How- ever, OSCs may cause unintended consequences [16, 62]. Semigran et al. [62] listed potential, troubling scenarios: if patients with a life-threatening problem being misdiagnosed and not told to seek medical care, the morbidity and mortality will increase; or, if patients with minor illnesses being told to seek emergency care, the times and costs for patients, physicians, and society will increase and patients and caregivers' anxiety will escalate.

OSC research is limited, but has already pointed to several urgent challenges in this area. First, in terms of methodology, the majority of OSC research seeks to evaluate OSCs' diagnostic accuracy by asking multiple medical experts to review OSCs' features and performance (e.g., comparing a checker's suggested diagnosis to clinical vignettes). OSC users, albeit an important stakeholder group, are largely neglected. Very little research has been undertaken to

investigate how real-world users actually use these symptom checkers, as identified by recent review articles [1, 8, 46], with one exception [59] testing the Technology Trust Model on users of one OSC and another [85] examining the user experience of OSCs. Critical questions such as how patients and caregivers interpret and use the diagnosis and advice provided by OSCs, how valid they assume the diagnosis and advice are, what impact OSCs have on their care-seeking and health outcome, how to design explainable OSCs that can empower users to evaluate the diagnosis and advice, and what constitutes a satisfactory explanation, remain to be answered [44, 62].

Second, expert evaluations of OSCs (e.g. [3, 4, 11, 53, 62, 65]) have shown that diagnosis and triage advice provided by symptom checkers are often inaccurate. Yet, those OSCs fail to describe the evidence base underpinning and offer no explanation for how the OSCs come to the results, which impose risks on the users and put the users into the challenging position to evaluate and decide how to deal with the diagnosis and triage advice [33, 62]. Thus, it is essential that users are able to understand and explain how the diagnosis and triage advice is made [79], so that they can make more informed decisions.

Due to OSCs' enormous popularity in the high-stakes context of healthcare, much research is needed to address how explanations could be supported in health consumers' interaction with OSCs. This research starts to address this research gap through the combination of an interview study and a design study.

Explainable UI and Styles

Enhancing transparency in intelligent systems has been discussed in the research of *explainable AI (XAI)* [21], which focuses on ex- plaining and justifying the outcomes of AI-driven decisions or recommendations [13, 36, 40, 51, 72]. Explainable intelligent systems can achieve different *explanatory goals* by single-style or hybrid explanations [39, 64]. Providing explanations has been studied in improving user satisfaction, user perception, and user experience as well as system transparency [39, 50, 56, 69]. The personalized explanation model in hybrid recommender systems [40] has inspired the presentation of explanations, i.e., how to formulate the text to the users and choosing explanation style based on the recommendation model [17]. Previous research has examined different explanation styles including rule-based explanations [39], feature-based explanations [45], algorithmic

explanations [14], post-hoc explanation [74], and rationale-based explanation [15], generating by text or visualizations to offer local or global explanations of the recommendation models [47]. Existing explanation styles are limited in addressing the multifaceted user experience in high-stake contexts such as healthcare, which requires a further understanding of the user's mental model and the explanation goal when generating and presenting explanations to the users [2].

The user's mental model represents the knowledge of information systems generated and evolved through the interaction with the system [47]. Miller argued that explainable AI researchers should consider the multi-discipline knowledge to improve the user experience, not just from the researcher's intuition to provide global explanations (explaining the decision model) [41, 47]. The *user-centered XAI* needs to consider the cooperative principles and the conversations between researchers, domain experts, and the users to identify the goals of explanations (e.g., How, Why, Why not, What if, etc.) [41, 57]. The philosophy of science and epistemology has discussed the theories of contrastive explanation, and counterfactual causality [35, 58, 78]. These theories could help us to define the user-centered explanation through three principles. 1) *human explanations are contrastive* : perceiving abnormality plays an important role in seeking an explanation, i.e., the users would be more like to figure out an unexpected recommendation [47]. 2) *human explanations are selective*: the users may not seek a *"complete cause"* of an event; instead, the users tend to seek useful information in the given context. The selective explanations could reduce long causal chains' effort and the cognitive load of processing countless modern AI models' parameters. 3) *human explanations are social*: the process of seeking an explanation should be interactive, such as a conversation. The explainer and explainee can engage in in- formation transfer through dialogue or other means [28]. These principle informed the design of *rationale-based*, *feature-based* and *example-based* explanations in this research.

Our work is deeply rooted in the conversational explanation in XAI [60, 67], i.e., the known mechanism of how the user requests an explanation from a chatbot-based AI application. Our design space is situated within a rich body of studies on conversational AI or chatbots applications, e.g., AI-driven personal assistant [42, 84].

The design of conversational agents offer several advantages over traditional WIMP (Windows, Icons, Menus, and Pointers) interfaces [6, 43]: 1) a natural and familiar way for

users to tell the system about themselves, which in turn improves the usability of a system as well as updates the user's mental model to the system. 2) the design is flexible (like a dialogue [71]) and can accommodate diverse user requests without requiring users to follow a fixed path (e.g., the controllable interfaces [75]). 3) the design could augment by a personified persona, in which the anthropomorphic features could help attract user attention and gain user trust [68]. The HCI community has long been interested in the interaction benefits offered by conversational interfaces [84]. We referenced to multiple state-of-the-art conversational agent techniques in designing our OSC, e.g., group decision supports [63], psychotherapy [61], in-depth interview for healthcare [12] and educational [83].

We found little attention has been paid to the transparency and explanation issues in high-stake domains such as healthcare and OSCs, despite their importance and enormous popularity. Our works aim to bridge the explanation design in *explainable AI (XAI)* and the underexplored medical domain [29]. We adopted a user- centric evaluation framework in measuring the user experience of an explainable OSC [37, 38, 72, 73]. The framework contains explicit (questionnaire) and qualitative user feedback through semi- structured interviews. Our experiment adopted the established assessment tools, e.g., NASA-TLX [23] and ResQue [54], to measure the subjective user feedback, e.g., satisfaction, trust, and workload.

RESEARCH DESIGN

To answer the research question of how explanations could be used to promote diagnostic transparency of online symptom checkers, we designed two consecutive studies. First, we conducted an inter- view study to explore existing OSC users' experiences, with a focus on explanation. The interview study was premised on 1) the general recognition of using explanations to enhance the transparency of intelligent systems, and 2) the lack of empirical evidence on whether and how users need explanations in their interactions with OSCs, and specifically how explanations could help. We deemed an interview study appropriate for identifying and clarifying the role of explanations in users' interactions with OSCs. Building upon the specified user needs for explanations, we used Study 2 to de- sign and test explanation design on an OSC. Study 2 tested and extracted concrete design knowledge about explanations for OSCs. We obtained the IRB approvals for both studies before they took place.

STUDY 1: DO USERS NEED EXPLANATIONS?

This interview study solicited user needs for information in their OSC experiences. The study was exploratory by nature, but the goal was to generate a list of concrete interaction scenarios where causal information specific OSC design could be beneficial if provided.

Method

From September 2019 to June 2020, we conducted 25 semi- structured interviews with users of OSC apps. We recruited participants through both social media and the web-based participant recruitment tool provided by our institution, which is open to the public in local communities. Qualified participants were selected through a screening survey with the following eligibility criteria: (1) the participant was over the age of 18; (2) the participant' purpose of using OSC app(s) was to seek potential diagnoses; and (3) the last time the participant used the OSC app(s) was less than one year. We finally selected 34 eligible participants. We first conducted a pilot study to test and revise our interview protocol with 9 participants. The pilot study data is not included in our data analysis. After finalizing our interview protocol based on the pilot study, we officially started the data collection process.

Our final set of 25 participants aged from 19 to 54. They have diverse educational backgrounds and occupations (e.g., student, landscape designer, university staff, software engineer, professor). The OSC apps they used include Ada, Ask NHS, K Health, Babylon, Your.MD, and a system checker provided by a university. The majority of them (21 out of 25) used more than one OSC apps. They used OSC apps to consult for various symptoms, such as headache, allergy, and irregular menstruation. We conducted individual inter- views with our participants. Each interview took from 40 minutes to about 1 hour. We started with general questions such as what OSC apps they have used, when and why they used the OSC app(s). We then asked more specific questions focused on different aspects of their experiences with and perception of the OSC app(s), such as the conversation experiences, the diagnosis process of the app(s), the results they received and the following actions they took. Five interviews took place offline prior to the COVID-19 outbreak. 20 were conducted online through Skype, Zoom, and phone call, due to the distance between the participants and us or the COVID-19 outbreak after February 2020. Each participant (including partici- pants of our pilot study) was compensated with a $20 USD Amazon gift card for their time and effort. The demographic data

are shown in Appendix A.

We used thematic analysis to analyze our data [5]. Driven by our analytic interest in understanding whether users need explanations and if so, what kinds of explanations they would like to have, we adopted an inductive approach [5]. Three authors first familiarize themselves with the data by repeated reading, during which each of us marked an initial list of ideas that are related to our analytic interest [5]. We then held meetings to compare and combine our initial lists of code through extensive discussions. With the consolidated list of codes, we then sorted codes into potential themes and sub-themes. Next, we review the set of candidate themes and sub-themes together. Through rounds of discussion, we defined and further refined the themes and sub-themes and achieved a satisfactory thematic map of the data [5]. Our final thematic map consists of three themes pertaining to the types of explanations that our participants would like to have, which we present in our findings. When reporting our participants' quotes, to protect their privacy, we use S1, S2, etc., to denote different participants.

Findings

Confused by the questions. Our participants felt confused about the questions asked by the OSCs in terms of the sequence, quantity, and relevance. They found that the sequence of questions oftentimes seemed to be random. For example, S10 found the questions asked by Your.MD and K Health jumped back and forth without a clear order, *I think these checkers ask questions without a clear order. For instance, after asking about diet, Your.MD suddenly asked 'have you ever traveled in the last three months', 'if it is something related to the environment'... it asked about my symptoms at first, and then it asked about the environment, and then asked about my symptoms again. I felt it was annoying, to be honest. I feel it's unprofessional...*

S14 also found that sometimes the order of questions was *strange* to her, and would like to know why a certain question was asked after the previous ones,

Ada asked me some very strange questions. It first asked me if I had a headache, then it asked me about my breathing muscles, and then cough. I think it should ask about my cough first, and then about my breathing muscles. Because you must have certain symptoms before a physical reaction, that is, it should ask about my

cough first, then ask about my breathing muscles. Also, I don't know what the relation between respiratory muscles and cough is. It would be better if there's some information. For example, it may have a dialog box that pops up. For example, it can show in the box, 'I ask you this, because you show a certain symptom that may be related to this...

On the quantity of the questions, our participants reportedly experienced frustration and boredom if being asked too many questions. For example, S6 told us that she lost patience in the end when she was using Ada and K Health, *"I feel they have so many questions, you need several minutes to finish it. Because the questions are too many, I don't have such patience to answer them all. So I kind of lost patience at the end.*

Participants sought to explain why those OSCs had to ask so many questions. For example, S5 compared the experiences of using OSCs versus the experiences of consulting human doctors. She acknowledged that the OSCs asked much more questions than human doctors maybe because checkers could not see her and could not assess her conditions based on how she looked,

I think they asked me too many questions. They tried to cover every single possible aspect. But I remember when I was seeing a doctor, the conversation was much shorter, maybe because doctors can exclude some possible diagnosis just by looking at me.

Most of our participants also questioned the relevance of questions asked by the OSCs. For instance, S14 was confused about why the OSC asked her questions related to her eyes while she had nose allergies,

But Ada asked me some questions that I think (these questions) were irrelevant. For example, it asked me whether my eyes were red or swollen. I think it's just allergies, which may not have much to do with my eyes. Then I was asked if my eyes had uncontrolled movements. I felt that it had nothing to do with my nose allergies. It may have, but I don't know.

S2 wished there could be some explanations provided regarding how the questions were related to the consulted symptoms, saying,

When I use Your.MD and when it asks a question i think unrelated to my symptom, I would doubt what is the meaning, why does it ask that? So, maybe I would trust it if it explains a little bit why the question is related. For example, when it asks whether I am sexually active, it can provide some information like what is the percentage of female

who are sexually active have this symptom or something like that. Just explain why it asks this question; otherwise, I will be concerned about my privacy. And I will doubt the accuracy of this checker.

Seeking explanations for the result. Our participants were curious to know how OSCs generated the diagnosis and suggestions. They would like to know which symptoms and information they input actually lead to the suggested diagnosis and advice. For instance, S11 who used Ada felt that there seemed to be not much relation between the questions he answered and the results he received,

Actually, not much relation (between the questions he answered and the results). For the possible causes listed to me, Ada doesn't tell [me] why my symptoms have a match. It just says something in a statistical way, like how many people might have this cause. I think the APP should show the relations, like explain why it thinks this might be a possible cause, which question it asked and which answers I gave have led me to this diagnosis...

The lack of such explanations made S11 felt the OSC not so helpful,

The questions Ada asked me, were too broad and the results it showed to me were like a ton of possible causes of my diseases, which did not help and I didn't get a clear vision, a clear explanation about why it thought I was having this disease and what the APP did was just like telling me to go to the doctor immediately and that did not help at all..

Another telling case is that S13, who was a university staff used the OSC provided by the university, would like to know what symptoms she inputs actually led to the diagnosis,

Like I said, if they were to be able to, at least for the (University name redacted) one,...if in the end, the results could say, "we think you have anxiety, and the reason we think you have anxiety is because of this and this". If they could just give us a little bit of explanation as to why they think the results are what they are. I think that would be a good feature.

S13 further described in detail what kinds of explanation she

would like to have, *I kind of wish that after they would give me the results, they would say, and this is how we arrived at the results... like whenever you put all your*

Information in, then this little pop-up screen comes up and it says it (the health condition) could be this, could be that. It would be nice that at the end of this, it could say something like this is the process by which our symptom checker app used in order to determine that it could be one of these three or four or five reasons. I just wish the symptom checker, after they give the results, would tell me "this is our formula for giving that result."

S16 who used K Health and Your.MD shared the same opinion as S13. He emphasized that providing some explanation regarding which symptoms lead to the diagnosis could help him know what are the main problems that he need to pay extra attention to, *..if it can explain which symptoms lead to the result...First, it's easy to do, and second, from the user's point of view, we will better know what are the main problems that have contributed to such a result, so that we can pay attention to it.*

S21 tried to guess what input led to the diagnosis, and thought maybe the OSCs assigned different weights to different symptoms that she had input,

When I want to check what I got, usually I type different symptoms...like the irregular periods, hair loss, or maybe stomachache...I think the apps count everything in terms of the probability, or they assign my irregular periods more weight to get the final results... For example, I have irregular periods, this is a main symptom for me recently...but I also have other symptoms, like stomachache, right?... Do these apps counting all of these symptoms equally? They may give my irregular material more weight?

Seek explanations for the knowledge base. Our participants were interested in learning about the data source and also the possible similar cases (e.g., someone has/have the same symptoms as oneself) that could help them to compare against their own conditions. They believed that such explanations could help them understand and interpret the recommended diagnosis and medical advice better. For instance, S17 would like to know more information regarding the data source and whether the OSCs were tied to medical institutions, which could impact her trust towards the OSCs,

I think a lot of people are skeptical specifically towards the data because you can make up the numbers, say whatever you want. Yeah, People want to know that it's reputable and then it's kind of been credential. So in many cases, reputations are

important. And if it's tied to a hospital system or if it's nonprofit, people may like view that as more authentic or valuable than it's just for profit business who is doing it without any medical guidance

S22 who used K Health would like to know more about the knowledge base and how the system actually uses her data and compares her to others, *I would like to know who built them (the*

checkers). I mean, what kinds of expert inputs are given?... I noticed that they said they might compare me to a couple of hundreds cases. Well, they did it based off of my age and my gender. Was there other things that might be involved also? So whenever they're there taking this data, just maybe a little more information about how exactly they're doing these comparisons and things like that.

S20 told us that she could not assess the results she received without knowing how the OSC collected the data, although the OSC told her people with similar symptoms had the possible diseases that she probably had.

Part of me is like that (the list of possibilities of each potential diagnosis provided by the checker) makes sense like people with those symptoms had those things. But I was like, did people just fill out this symptom checker and then they're assuming that people indeed had a migraine. I don't know, I don't know how they collected the data, So I'm not totally sure.

Summary

Based on the user study findings, we proposed three possible ex- planation styles in enhancing the transparency and explainability of OSCs.

First, the OSC users were confused by the *questions* proposed by the system, e.g., the sequence, quantity, and relevance. Therefore, it could be helpful to provide explanations *during* the conversation and interaction, i.e., after each prompted question. Hence, we generated our first explanation style by providing the *rationale* of each question prompted by the symptom checkers. That is, for each promoted question, the system should provide additional information about 1) *why* does the system ask these questions? 2) *how* many questions would be asked or needed? 3) *what* information does the system need from me?

Second, the OSC users desired a better *understanding* of the medical recommendations

from the system, i.e., what are the *reasons* that the system used to make the decision? This indicates that a user can benefit from seeing explanation *after* the conversation and interaction, i.e., after the user received the medical recommendations. Based on this finding, we designed our second explanation style to provide post-hoc explanations based on the features that are used to make the medical recommendations. In this case, the user need is to explain the prediction from the system, i.e., *how* does features of the instance contribute to the prediction [41]? *why* do I receive this medical recommendation?

Third, the OSC users wanted to know the data source and to compare the similar cases to help them to understand the medical recommendation better. That is, to provide an *example* identical to the instance and with the same record as the prediction, i.e., the example-based explanations [41]. The example would be provided after the user received the recommendation. Thus, we generated our third explanation style by providing post-hoc explanations based on the *data source* and an *example* with similar or the same symptoms, which to fulfill the user needs of *why* do I receive this medical recommendation? and *what* is the data source used to decide for me?

STUDY 2: HOW TO EXPLAIN?

To explore the effects of providing explanations in OSCs, we chose to replicate the Centers for Disease Control and Prevention (CDC)'s coronavirus self-checker [7]. The COVID-19 OSC is an interactive clinical assessment tool for users "*on deciding when to seek testing or medical care if they suspect they or someone they know has contracted COVID-19 or has come into close contact with someone who has COVID-19* "[7]. We chose to design the OSC for COVID-19 for four reasons: First, COVID-19 is a crucial issue in the global pandemic. A usable OSC may be used to help contain the current and future epidemics. In fact, in our Study 1, two participants (S17 and S24), whom we interviewed when the large-scale outbreak of COVID-19 had started in the U.S., told us they had started using symptom checkers to check whether they were infected or not frequently. Second, many companies (e.g., Apple, Google) and medical and public health institutions (e.g., Mayo Clinic, CDC) have developed COVID-19 OSCs in response to the COVID-19 crisis, which has been widely adopted. For instance, reopened universities during the COVID-19 pandemic are at high risk. Many universities have implemented or adopted COVID-19 symptom checkers to help students

and employees conduct self-screening after reopening the campus, e.g., Florida State University [76], Pennsylvania State University [66], Quinnipiac University [77], the University of Iowa

[49] and more. Developing a better COVID-19 symptom checker can help individual users and institutions manage the risks. Third, focusing on one specific disease helps us control the variables in the experimental platform better to compare the effects of the proposed explanations styles. Fourth, it helps lower the learning curve for our study participants since they were aware of the on-going pandemic and have a certain knowledge about COVID-19.

System Design and Explanation Styles

The COVID-19 OSC is a chatbot-based self-assessment tool that via text (shown in Figure 1). We chose to design a chatbot user interface due to the following reasons: 1) All the OSC apps adopted by our participants in study 1 and the COVID Self-Checker offered by CDC are chatbot-like designs, so we can assume the prospective users and health consumers are already familiar with the interface.

2) According to our literature review, conversational interfaces like chatbot have advantages over traditional WIMP interfaces, such as offering natural user interactions (e.g., conversation), al- lowing diverse user requests (e.g., allowing various symptoms), and personifying features to gain user trust [84]. 3) The conversational explanation is the widely adopted mechanism of how the user requests an explanation from robots or AI-driven applications [60, 67].

The COVID-19 OSC was built based on *Botman*, an agnostic PHP library that is designed to simplify the task of developing chatbot and conversational interface [1]. We implemented the app using the Laravel framework and PHP scripting language. We chose the framework to minimize the costs for our research team and easy- to-use user interfaces for most participants to lower the learning curve. The COVID-19 OSC has three components (shown in Figure 2a): 1) an input box to receive user commands. 2) input and output functions to present messages from the chatbot and users; messages are distinguishable by the personas and box colors. 3) *quick response buttons* that provide user clickable buttons to accelerate the message input.

Figure 2a presents an example of how a user began the conversation by typing a

command, which would trigger a predefined conversation method. The method would promote the welcome messages to the user and asked if the user would like to proceed or

[1] https://botman.io/

abort the conversation. Each question came with two clickable buttons (Yes/No), allowing the user to respond by simply clicking the desired answer. Figure 2b presents an example of how the *COVID-19 OSC* prompted questions based on the clinical decision tree (see Appendix B), i.e., the "**Feeling Ill**", "**2-Week Contact**" and "**Showing Symptom**" questions.

The original CDC's COVID-19 OSC generates diagnostic and triage recommendations through clinical decision tree algorithm, while the process and rationale of diagnostic recommendations are not transparent to the users. In our replica, we add three style of explanations to improve the transparency of the diagnostic and triage recommendations, including *rationale-based*, *feature-based* and *example-based* explanations (shown in Figure 3). The proposed explanation styles were inspired by the empirical findings in Study 1 and the conversational explanation in XAI [60, 67].

Style 1: Rationale-based Explanation this style focuses on the immediate feedback on the *rationale* of the proposed questions. That is, the style would provide an explanation after *each question* the system promoted to the user. For example, the "**2-Week Con- tact**" question promoted the user "*In the last two weeks, did you care for or have close contact with someone diagnosed with COVID- 19? (Yes/No)*". The rationale of this question is to know if the user has close contact history that may imply a high-risk coronavirus exposure, so the tool would offer the explanation as "*I ask your contact history of COVID-19 to determine your risk of virus exposure.*" Another example shown in Figure 3a, the "**Medical Condition**" helps people assess their symptoms and determine if they may need to be tested for COVID-19, through chatting with the chatbot question asked the user's existing condition, e.g., "*Do any of these apply to you? Obesity, High blood pressure, Blood disorder, Smoking?*". The rationale of this question is that the underlying medical conditions are at increased risk of severe illness from COVID-19. The tool would provide explanations as "*I ask two above questions due to people who have severe underlying medical conditions like heart or lung disease or diabetes seem to be at higher risk for developing more serious complications from COVID-19 illness.*" All rationale-based explanations (and the

wordings) are based on the CDC guideline [7]. The rationale-based explanation is an intentional post-answer explanation to avoid possible bias to the users.

Style 2: Feature-based Explanation this style offers a *personalized summary* based on user input *answers*, which are the features used in the clinical decision tree model. The explanation provides an itemized summary to the users so that they can understand and review *why* they received a particular recommendation from the system. The *answers* are explained by the structure of *binary (Yes/No)* or *countable* questions. Figure 3b presents a sample feature- based explanation containing six features regarding the clinical recommendation derived from the user' input answers. Feature 1, 2, 4, and 5 are binary questions that will be explained in affirmative or negative sentences, e.g., for the question of "*Feeling Ill*"; the ex- planation would be "*you are feeling ill.*" for a positive response and "*you are not feeling ill.*" for the negative response. Feature 3 and 6 would be countable questions that will be explaining in a string array. The explanation would be an *array* to list all the items in the iteration. For instance, for the question of "*Showing Symptom*", the explanation would be "*you have two symptoms (breathing, muscle) related to COVID-19*" if the user submitted positive answers to the following symptoms: [moderate difficulty breathing] and [muscle aches or body aches].

Style 3: Example-based Explanation this style offers an *identical example* of a patient who receiving the same clinical recommendation with the same user input *answers*. This explanation would only provide the key factors related to the clinical recommendation. The logic to determine the key factor is to prioritize two *latter* and *positive* answers from the user inputs. If there are no positive answers, then the last two negative answers would be used. Figure 3c presents a sample of the example-based explanation, the user' input answers triggered the *rule 9* in Table 5. The explanation would focus on the two positive and latter answers of the question "*Showing Symptom*" and "*Medical Condition*" since the two questions are the key questions to decide the final clinical decision. For *rule 2*, the explanation would be focusing the two latter negative answers, i.e., "*(Explanation) Based on CDC COVID-19 guidance, a person who has no symptom and no COVID-19 contact history would get the same recommendation above.*". The explanation promoted the user about the information/data source, i.e., *Based on CDC COVID-19 guidance*.

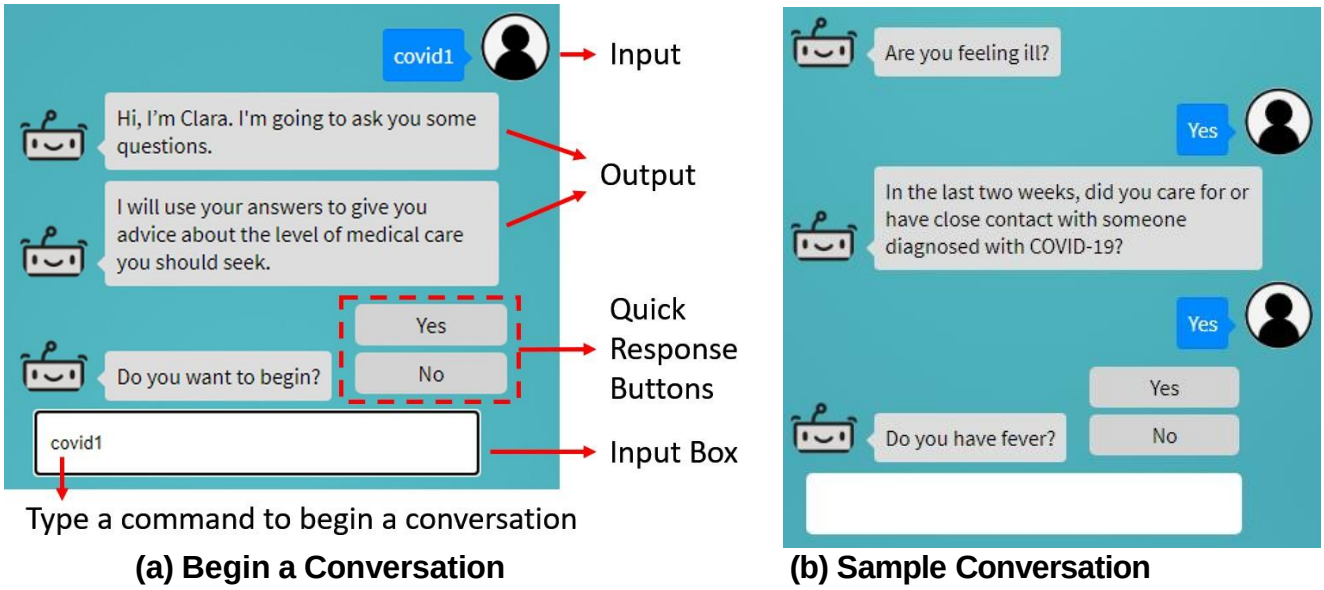

Figure 2: The system design of the *COVID-19 OSC*: (a) An example of how a user began a conversation to get diagnostic about COVID-19. (b) A sample conversation of how a user interacted with the tool.

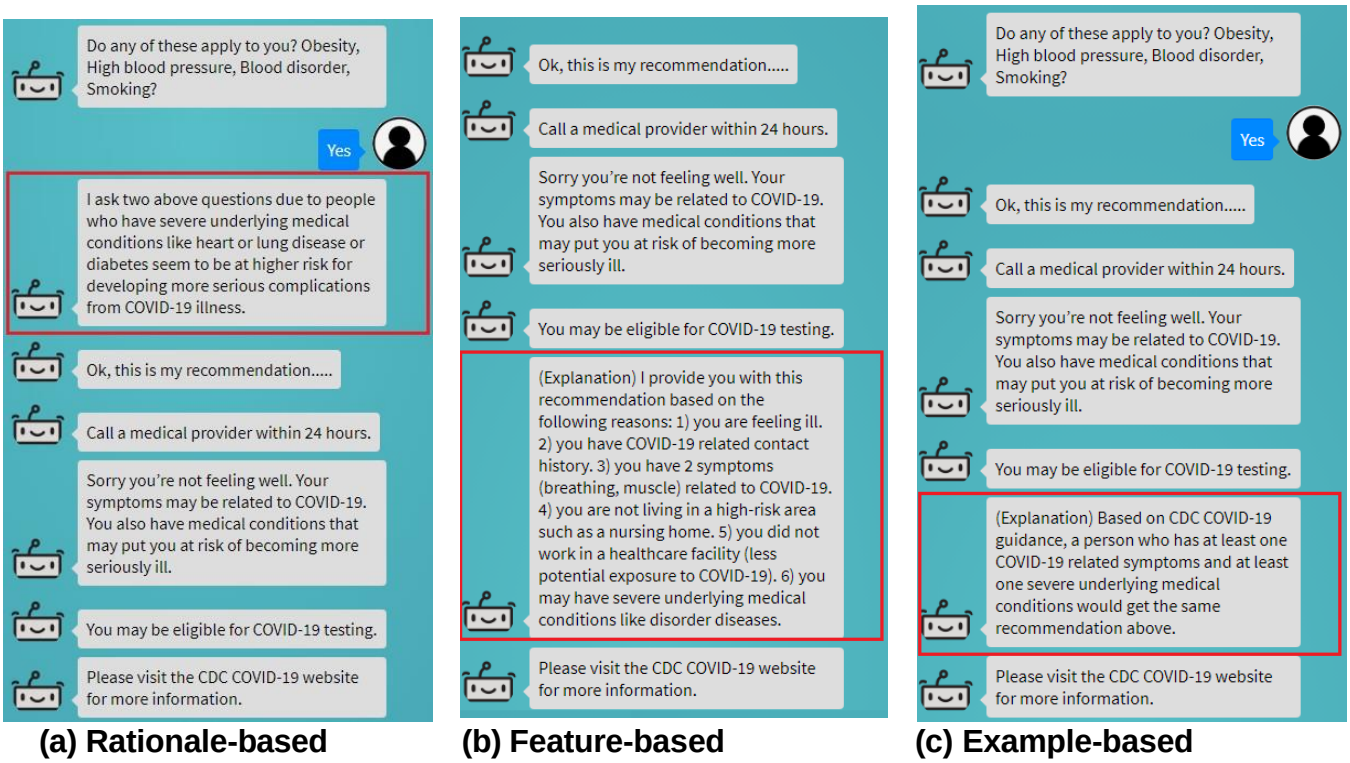

Figure 3: The diagnostic and triage recommendations of the *COVID-19 OSC* and the explanations (highlighted by the red box). (a) Rationale-based: this design provides the rationale of each asked questions. (b) Feature-based: this design offers explanations by summarizing the user answers. (c) Example-based: this design explains the recommendation by giving an identical example as well as the data source (e.g., based on CDC guidance).

Evaluation

Participants. We recruited participants from our university's participant recruitment website, social media, and word-of-mouth. To be qualified in this study, every participants should 1) be an adult (18+ years); 2) be a registered undergraduate or graduate student in the university; 3) has a computer or laptop with the internet to participate in the study remotely. We chose to recruit university students for this study because the coronavirus checkers were widely adopted in many universities for screening the Fall 2020 returning students [49, 66, 76, 77]. Their feedback could better help to refine the COVID-19 OSC for the users in the educational context.

A total of 20 students were recruited for this user study, from August 17 to August 28, 2020. The subjects included fourteen females and six males whose ages ranged from 19 to 37 years old (M=25.65, SD=4.27; M=Mean, SD=Standard Deviation). There were five undergraduate students and fifteen graduate students from different majors and diverse cultural backgrounds. The demographic data are shown in Appendix A.

To control for any prior experience with the COVID-19 checkers and conversational agent techniques, we included two questions in the pre-study questionnaire. The average score of COVID-19 checkers familiarity "*I feel I am familiar with COVID-19 checker or screen app*" was low (M=2.70, SD=1.83) on a seven-point scale. The average score of conversational agent familiarity "*I feel I am familiar with any chatbot agent or app*" was relatively higher (M=4.75, SD=1.91) on a seven-point scale. The feedback indicated the majority of participants were not familiar with any existing COVID-19 checkers but had sufficient knowledge of conversational agents to use the proposed design in this study. Each participant received $20 grocery gift card as compensation.

Study Design. Given the extraordinary circumstances of the COVID-19 pandemic and the requirement from the IRB office, all user studies were conducted virtually through Zoom. We shared the experimental system link and asked the participant to share his/her screen in the Zoom meeting. All user study sessions were video and audio recorded. We obtained written consent through an online form at the beginning of each user study session.

We deployed an experimental design with four conditions: 1) A baseline interface, a standard checker with all explanation styles disabled (Figure 1). 2) a rationale-based

interface, offering explanation after each prompted questions to the user (Figure 3a). 3) a feature-based interface, offering post-hoc explanations based on the user input answers (Figure 3b). and 4) an example-based interface, offering post-hoc explanation based on a similar case (Figure 3c). We followed a within-subject design, asking all participants to use each interface for one training and three study tasks. All of them were required to complete the first training session to ensure their familiarity with the interfaces. The participants were assigned to fill out a post-stage questionnaire at the end of each condition. At the end of the study, participants were invited to a semi-structured interview. To minimize the learning effect and ordering bias, we followed a Latin square design to balance the conditions given to each participant. The question order in the post-stage questionnaire was randomized to prevent the order bias.

In the training session, we urged the user study participants to use the first interface they were assigned, so they had a chance to familiarize themselves with the system. To minimize individual bias, the subjects were told to act like a college student based on three pre- defined scenarios. They were requested to use the COVID-19 OSC to find out if the student needed immediate medical attention and COVID-19 testing, based on the diagnostic results. Participants were given the same information-seeking scenarios for each interface.

The assigned tasks were designed to be realistic tasks that could be naturally performed by college students. The scenarios were designed to cover three specific rules in the decision tree model, to ensure the participant could experience sufficient use cases of checking COVID-19, i.e., scenario 1, 2, and 3 represents the specific *rule 4, 7, and 9* in Appendix B.

Scenario 1: Sage (female, 26 years old) is a junior year college student who wants to self-screen for potential COVID-19 symptoms. She does not feel ill, but she has close contact with someone diagnosed with COVID-19 in her dormitory. No other high-risk contact history. Please use the *COVID-19 checker* to find out if she needs immediate medical attention and COVID-19 testing.

Scenario 2: Ben (male, 20) is a first-year college student who wants to self-screen for potential COVID-19 symptoms. He served as a volunteer in a local hospital last week. He starts to feel ill and has symptoms like fever, sore throat and loss of taste today. He does not know if he had contact with someone diagnosed with COVID-19. He did not visit nursing homes or other high-risk areas in the past two weeks. Please use the *COVID-19 checker* to find out if she

needs immediate medical attention and COVID-19 testing.

Scenario 3: Ron (male, 22) is a senior year college student who wants to self-screen for potential COVID-19 symptoms. He did not visit nursing homes or medical facilities in the past two weeks. Today, he feels ill and has symptoms like difficulty breathing (not life-threatening) and muscle aches. He was diagnosed with high blood pressure disease last year and is un- der treatment. Please use the *COVID-19 checker* to find out if he needs immediate medical attention and COVID-19 testing.

Participants took between 49 and 72 minutes (M=60.37, SD=7.02) to complete the study. They spent around five minutes in each interface: Baseline (M=4.59, SD=1.41), Rationale (M=5.12, SD=1.16), Feature (M=4.51, SD=1.48), and Example (M=5.06, SD=1.48). The Rationale interface took a longer time to complete, but we did not find a significant difference between spent time across the four interfaces. The post-study interviews took 15 to 30 minutes (M=25.40, SD=2.66) to complete.

Measurements and Data Analysis. We adopted a mixed- method approach to address the research questions, sequentially combining statistical measures, quantitative survey data, and inductive thematic analysis, in which we used the qualitative data to explore quantitative findings [82]. The quantitative data was collected based on subjective metrics to measure the effectiveness of the explainable *COVID-19 checker*. The subjective measures were captured by the post-stage questionnaire (a 7-point Strongly dis-agree to Strongly agree Likert scale). We adopted a user-centric evaluation framework to measure the user experience of the pro- posed explainable OSCs. We asked identical four NASA-TLX usability questions (the physical construct was not included due to the four interfaces not having much difference in the aspect of physical demands) and 18 user perception questions in the questionnaires. The survey questions were selected and modified from the applicable existing constructs (factors) from the works of [36, 55, 56, 75], including eight constructs of quality, control, effectiveness, trust, transparency, satisfaction, awareness, and learning. (See Appendix C for details). The choice of constructs and questions were inspired by the works of [37, 38, 72, 73] to evaluate the explainable user interface in multiple contexts. The qualitative data was collected by conducting post-experiment semi-structured interviews. We prepared 10 interview questions about the pros and cons of using the COVID-19 OSC (See Appendix D for details.). We do not re- port the emergent themes generated from

our thematic analysis of the interview data because they do not contain nuanced insights, since participants mainly talked about how they appreciated the explanations, etc..

Results

To determine and compare the effects of the three explanation styles in the COVID-19 OSC, we conducted a paired Wilcoxon signed- rank test for each factor. Table 1 presents the subjective score of user feedback on the baseline interface and the three interfaces augmenting explanations. Based on the quantitative analysis of survey data, we found explanations could significantly increase user perception in multiple aspects. We further adopted inductive thematic analysis to explore user feedback from post-study interviews.

Post-hoc Explanations Could Improve Perception of Diagnostic Quality. Based on the post-survey analysis, We found all scores from the manipulation groups were higher than the baseline interface. Still, only the feature-based explanation (V = 10, p = 0.014) and the example-based explanation (V = 15, p = 0.011) were significantly higher than the baseline. The explanations increased the quality score by nearly 1 point, which was a fairly large effect on a 7-point scale. Thus, we confirm that explanations can enhance the user perception of diagnostic quality.

We found participants preferred to read the explanations *after* they had received the medical recommendation. P3 shared about how the explanations helped her accept the medical recommendation from the COVID-19 OSC: "*It would like to have the [diagnostic] decision first.. and combined with my experience before, like what information and knowledge I got before, or from a website. So if I feel [the explanations] make sense to me. I will follow the recommendation.*" This acceptance of the medical recommendation signified that P3 perceived a high-quality diagnosis from the COVID-19 OSC.

The rationale-based explanation seemed to require significant cognitive loads to remember all the explanations and process them in the end, thus hurting its user experience. P4 commented that "*the app is like giving me feedback on each question or each answer I responded on... this making rationale like a recommendation at the end. And also, throughout the progress.*"

Explanations Could Facilitate Medical Decision-Making. The COVID-19 OSC could assist the patients in making medical decisions such as seeking appropriate medical

resources or professionals. The scores of rationale-based explanation (V = 11, p = 0.030), feature-based explanation (V = 5, p < 0.01) and the example-based explanation (V = 19.5, p < 0.01) were significantly higher than the baseline. Although the feature-based one was outperformed the other two manipulations, all explanation styles increase the effectiveness scores by nearly or over 1 point, which is a fairly large effect on a 7-point scale. We thus confirm that providing explanations can help the user make better medical decisions.

Table 1: User Perception of Explanation Styles

Factor	Variable	Alpha	Baseline Mean (SD)	Rationale Mean (SD)	Feature Mean (SD)	Example Mean (SD)
Quality	Q1, Q2	0.94	4.87 (1.62)	5.50 (1.06)	**5.80 (0.89)** *	**5.70 (1.27)** *
Control	Q3, Q4	0.08	5.82 (0.84)	6.12 (0.68)	6.25 (0.69)	6.07 (0.61)
Effectiveness	Q5, Q6	0.83	4.82 (1.29)	**5.65 (1.00)** *	**5.90 (0.96)** **	**5.47 (1.01)** **
Trust	Q7, Q8, Q9	0.91	4.68 (1.47)	**5.65 (0.94)** *	**5.71 (0.94)** **	**5.58 (1.09)** **
Transparency	Q10, Q11	0.90	3.25 (1.43)	**5.87 (1.04)** ***	**6.45 (0.64)** ***	**5.85 (1.32)** ***
Satisfaction	Q12, Q13, Q14	0.94	4.81 (1.71)	**5.65 (1.11)** *	**5.88 (1.15)** **	**5.45 (1.29)** *
Awareness	Q15, Q16	0.82	4.72 (1.46)	**5.55 (0.94)** *	**5.92 (0.87)** **	**5.45 (1.26)** **
Learning	Q17, Q18	0.85	4.35 (1.47)	**5.67 (1.01)** **	**5.35 (1.38)** *	**5.07 (1.35)** *

Statistical significance level: (*)p <0.05. (**)p <0.01. (***)p<0.001

Table 2: NASA-TLX Usability Analysis

Factor	Variable	Baseline Mean (SD)	Rationale Mean (SD)	Feature Mean (SD)	Example Mean (SD)
Mental Demand	TLX1	1.95 (1.43)	2.10 (1.25)	1.80 (1.15)	1.75 (1.32)
Performance	TLX2	6.10 (0.78)	6.40 (0.68) *	6.45 (0.68) *	6.30 (0.65)
Effort	TLX3	1.95 (1.39)	1.85 (1.34)	2.05 (1.43)	1.95 91.35)
Frustration	TLX4	1.85 (1.30)	1.65 (1.03)	1.55 (0.94)	1.60 (1.27)

During the post-study interviews, participants commended the post-hoc explanations (in particular, the feature-based and example- based explanations) as a personalized health condition summary. For instance, P2 stated "*[the system] gives a pretty detailed explanation as to why they recommend what they do. So as a result, I'll be able to make a more subjective decision based on what I've been told through the chatbot rather than just like telling me to do this ... so you feel a bit more confident in yourself as to, like, whether or not I should get tested*". In this example, the explanation provided sufficient information to assist the user to

proactively and confidently make the medical decision. On the other hand, the explaining can help the user to comprehend the medical conditions and information. P4 said "*I think the summarized information is useful when I'm talking to medical professional*". The participant mentioned the explanations could help her better communicate with the medical professional, which can also contribute to a better medical decision for the user.

Explanations Could Improve User Trust in Diagnosis. We focused on whether and how explanations could improve user trust in the COVID-19 OSC. We found the scores of rationale-based ex- planation ($V = 9$, $p = 0.011$), feature-based explanation ($V = 11.5$, $p < 0.01$), and example-based explanation ($V = 2.5$, $p < 0.01$) were significantly higher than the baseline system. All explanation styles increased trust scores by nearly or over 1 point, which was a fairly large effect on a 7-point scale. The feature-based one was outperformed by the other two manipulations. Providing explanations can increase the user trust in the OSC.

The post-study interviews confirmed this point. For instance, P3 shared "*No matter what kind of explanations... they provide you a different format of explanations... that provide me some rationale behind [the COVID-19 OSC], So it made me more trusting*". The example helps explain how rationale-based explanations can lead to better user trust, e.g., rationale behind asking certain questions, or rationale behind certain diagnostic decisions. We notice the *authority information sources* can increase the user trust, like P1 mentioned "*As a user, I would probably trust the app more if I know it has some affiliation or connection with CDC or a source that people just trust*". That is, in the example-based explanation, the COVID-19 OSC dis- closed the data source from an official authority, which made the participant trust the medical recommendation. P14 shared similar feedback, "*... and because the CDC is like just this governmental generalized resource, it makes it more trustworthy and reliable*".

Explanations Could Enhance Diagnostic Transparency. One of the primary goals of providing the explanations was to make the COVID-19 OSC's diagnostic recommendation more transparent. We would like to know if the users knew *why* the medical recommendations were provided to them. We found the scores of rationale-based explanation ($V = 4.5$, $p < 0.001$), feature-based explanation ($V = 0$, $p < 0.001$) and example-based explanation ($V = 5.5$, $p < 0.001$) were significantly higher than the baseline system. All

explanations condition increased the transparency score by nearly or over 2 to 3 points, which was an enormous effect on a 7-point scale. The feature-based was outperformed by the other two manipulations. We confirm that providing explanations can increase the transparency of the OSC.

The qualitative user feedback also supported this conclusion. For example, P20 said "*the [feature-based explanation] gave me why such a medical decision was made. Yeah, I know about what are my symptoms. And yeah, I know why does the reason that it gave give such recommendations.*" The participant would be able to understand the *reasons* of why the OSC made the decision. It is an important aspect to measure the level of system transparency. P10 further mentioned how she perceived transparency thanks to the explanations, "*I have an apparent idea of why I get this kind of summary, and I know the reasons, and this is more convincing to me. If I don't know the reasons, I will still be a little bit worried about whether [the medical recommendations] are true or not*". Explaining helped eliminate the uncertainty and confusion for receiving the recommendations. Transparency made the medical decision more convincing.

Explanations Could Help Increase Health Awareness. Inter- acting with the COVID-19 OSC can help users gain awareness of their body conditions or symptoms, which is critical in the COVID- 19 pandemic. We found the scores of rationale-based explanation (V = 29.5, p = 0.025), feature-based explanation (V = 8, p < 0.01) and example-based explanation (V = 19, p < 0.01) were significantly higher than the baseline system. All explanations condition increased the awareness score by nearly or over 1 point, which was a fairly large effect on a 7-point scale. The feature-based was outperformed by the other two manipulations. We thus conclude that providing explanations can increase the awareness of body conditions or diseases.

The qualitative user feedback suggested that participants agreed the explanations would function like a *reminder* of their body condition and COVID-19 information. For instance, the feature-based and example-based explanations would be a personalized summary that the users can use as a *script*. For instance, P12 mentioned "*If you want to treat [the explanation] like a script. So you can have a conversation with another person.*" The rationale-based explanation would be useful for understanding COVID-19-related knowledge as well as communicating with others, like the contact history, high- risk area, associated symptoms, etc.

For instance, P18 said "*Like an explanation for something... if you're trying to tell someone else that they need to get [COVID-19] tested too.*" P3 said "*[the explanations] gives me more sense of focus on my body condition*".

Explanations Could Facilitate Learning. It would be beneficial if the COVID-19 OSC could help users learn more about COVID-19. We found the scores of rationale-based explanation ($V = 24.5$, $p < 0.01$), feature-based explanation ($V = 36.5$, $p = 0.033$), and example-based explanation ($V = 36.5$, $p = 0.033$) were significantly higher than the baseline system. All explanation styles increased the learning scores by nearly or over 1 point, which was a fairly large effect on a 7-point scale. The rationale-based one was outperformed by the other two manipulations. We confirm that providing explanations can help users learn new knowledge from the OSC.

In addition, we found explanations could have an extended effect beyond the specific use scenario. For instance, P12 said "*yeah, like in both settings like, you know, educating them so that they can talk to the doctor, but all the words that they can use to tell their family and friends that this might be happening to them.*" Thus, explanations could facilitate disease-related conversations or remind family members and friends after the use of an OSC. It is interesting to see the explanations on questions would better help the users to incidentally learn new knowledge from the COVID-19 OSC.

Explanations Could Contribute to User Experience. User experience is multifaceted. Here, we discuss the aspects of satisfaction and usability. We found the scores of rationale-based explanation ($V = 34$, $p = 0.046$), feature-based explanation ($V = 22$, $p < 0.01$), and example-based explanation ($V = 4.5$, $p < 0.001$) were significantly higher than the baseline system. All explanation styles increased the satisfaction score by nearly or over 1 point, which was a fairly large effect on a 7-point scale. The feature-based was outperformed by the other two manipulations. We confirm that providing explanations can increase the satisfaction of the OSCs.

All participants in the study agreed that the four systems were easy to learn and use. However, explaining every single question required more cognitive loads to read and process the information, and thus more time to finish the tasks. A similar pattern can be found in the NASA-TLX test (shown in Table 2). Both rationale- based ($V = 0$, $p = 0.019$) and feature-based explanations ($V = 0$, $p = 0.026$) helped significantly improve the information-seeking

performance (TLX2). We did not find any significant difference in mental demand (TLX1), effort (TLX3), and frustration (TLX4), which indicated all interfaces used in this user study were comparably easy-to-use. However, providing too much information and explanations may cause an extra cognitive burden for users.

DISCUSSION

In this paper, we presented our work to promote diagnostic transparency by augmenting online symptom checkers with explanations. Through the first interview study, we found that users desired to see explanations about questions, data sources, and results, which existing OSCs fail to accommodate. Our follow-up design study proposed and tested three explanation styles in a COVID-19 OSC. Our research echoes previous work on how explanations could enhance the transparency of intelligent systems [39, 70]. We showed that a more transparent OSC helps reduce uncertainties and con- fusion for receiving the medical recommendation. In addition, our research shows that explanations improve the overall user experience of OSCs, because 1) users perceive better diagnostic quality in the post-hoc explanations; 2) users can better understand symptoms and clinical decision model; 3) the explanations can help users organize medical information so they can make better decisions, such as whether to see a medical professional; and 4) an understandable or comprehensible medical recommendation supported by authoritative sources can enhance user trust.

Our research also demonstrate the contributions to explainable AI (XAI). First, our work provides empirical evidence to extend the understanding of how to present explanations in a conversational manner [57, 60, 67]. The human-AI conversation is the widely adopted mechanism of how the user requests an explanation from a conversational AI-driven application. We designed, embedded and evaluated three explanation styles in a chatbot and discussed the related user experiences in a healthcare context. Second, our work advances the effort of improving transparency and trust in high-stakes domains requiring reliability and safety such as XAI in healthcare. Prior work mostly focuses on the decision model explainability [52], our works pay more attention to the end-user side experience and usability.

Third, we discuss the possible usage of XAI in a pandemic context. Our work contributes to the urgent challenge of the COVID-19 pandemic in terms of designing an easy-to-use, understandable, trustworthy symptom checker [30]. Fourth, our research highlights the broader

social implications of providing explanations to the Online Symptom Checkers (OSCs). Our finding indicates providing explanations helps users to learn and internalize medical knowledge about their symptoms, strikingly different from the non-explanation scenario where users receive suggestions but do not know why. Users could then spread such knowledge to other people who have not used the OSCs. In this way, explanations at the interface could have a broader impact and reach a wider population.

Our works provide empirical evidence on how users may use OSCs to check their early symptoms and gain more knowledge about their conditions. The OSC's recommendations could help users identify the appropriate care level and services before seeking medical advice [53, 81]. The users could become active participants in the medical decision-making process, which can be seen as a form of patient empowerment [80]. Our findings could advance the traditional health recommender systems in terms of personal health persuasion, empowerment, and trust.

Design Implications: Our research indicates the need to enhance some form of transparency and explanation in designing OSCs. We suggest the following design implications: First, explaining the information source/authority could affect user trust. The users would like to know if the medical recommendation and information were endorsed by the authority or healthcare professions. The findings shed light on the new opportunities to engage the professional in the loop of explanations. Second, users may prefer different explanations in medical information seeking. The explanations should not be static or "one-size-fits-all". Users may like some form of control and customization in receiving the explanations. For instance, a user may like to see *light* explanations in the beginning and *full* explanations after receiving the medical recommendations. The users may be more willing to see the explanations when the recommendation is *abnormal* (i.e., out of their expectations). The findings imply that the explanations should be adaptable and controllable based on the users' real-time information needs. Third, we observed that system transparency has an impact on users' health awareness. Users tend to know the *self-interest* information instead of the system's decision model. The finding indicates the importance of providing *relevant* explanations to the users, e.g., the background information or reasons for the prompted medical recommendations. The explanation is then could be used as a way to help the user further learning and gaining health awareness about themselves.

LIMITATIONS

We are aware of some limitations of our studies. First, we narrowed the disease scope to a specific disease (i.e., the COVID-19) to better control the variables in our user study. However, many real-world OSCs allow users to explore numerous symptoms and diseases. Interaction and conversation would be more complicated. External validity would require further exploration and studies. Second, we had a relatively small sample size, which may limit the statistical power. Our sample is also limited to the characteristic of our sampling pool and may not represent the general OSC user population. We are aware that the information needed for a symptom checkers is general across different age groups. There is a strong need to test the design in different age groups in future works. Third, we can attribute differences to which explanation participants inter- acted, but we cannot draw causal conclusions about which parts of the texts caused which effects. We have considered adding control and combined explanations in our future works. Due to the constraint of within-subject study, we aimed to minimize the bias across conditions, so the combined condition was not included. All these limitations need to be further explored in future works.

CONCLUSION AND FUTURE WORKS

This paper reports on a research project about explanations in user's interactions with online symptom checkers. We showed that existing online symptom checkers are limited in the explanations they offer. The appropriate provision of explanation interwoven with the conversation flow could improve user experiences with online symptom checkers in multiple ways. Our research points to much research in the future regarding what information needs healthcare consumers have and how to provide explanations for consumer-facing healthcare systems.

Our findings highlight how explanations can be interwoven with conversation flow, yielding implications for future OSC designs. First, *Put User in Control*: our findings depict the diverse user needs that the OSCs users may need the explanations in a different situation. For instance, based on the post-experiment interview, users with severe or acute symptoms may not like to view the explanations before receiving the diagnosis. Curious learners may want to know more descriptions of the prompted questions. A returning user may want to disable the

explanations to speed-up the query completely. We would consider and evaluate a controllable explanation setting that users can request or browse the explanation when needed in our future design [72].

Second, *Multifaceted Explanations*: In this paper, we introduced three explanation styles for the question and the diagnostic pre- diction. However, more transparency could be provided in diverse ways. For example, a patient may be more willing to provide true symptoms or health information or re-use the OSC if the system explained how their data would be protected. The users may trust the OSC more if they are told the diagnostic predictions are confirmed or provided by reliable sources. Our future research, system design, and evaluation would consider diverse and various explanation styles as well as combinations of them.

Third, *Context-Aware Explanations*: our findings already demonstrated how OSCs could help promote health awareness and learning and remind users to pay attention to their condition or under- lying symptoms. Participants also appreciated the authoritative sources embedded in the explanations. This suggests that future explanations could be more proactive and context-aware. For example, regarding preventive measures and resources of epidemic diseases such as COVID-19, the OSCs could include actionable information about testing eligibility and the locations to get a test.

Fourth, *Be cautious of the Potential Downsides*: Explaining may come with negative aspects. For instance, too many explanations may cause information overload that reversely impairs user experience. Facing the overloaded explanations, users may *ignore* the prompted messages or suggestions from the systems. On the other hand, the healthcare system design tends to hide its decision process to the users [62]. Providing explanations at an inappropriate level of transparency may reveal the sensitive details so the user could *hack* the system or *intrude* the patient privacy, e.g., to know the flaw of the algorithm or to detect sensitive data sources, etc. For a high-stake domain like healthcare, all these negative aspects may cause health, liability, and legal risks. More studies would be needed to tackle all these challenges.

ACKNOWLEDGMENTS

Many thanks to our study participants. We are grateful for the anonymous reviewers' insightful and constructive feedback. This work was supported by Penn State College of IST's seed grant award (150000004308 INTR).

REFERENCES

[1] Stephanie Aboueid, Rebecca H Liu, Binyam Negussie Desta, Ashok Chaurasia, and Shanil Ebrahim. 2019. The use of artificially intelligent Self-Diagnosing digital platforms by the general public: Scoping review. *JMIR medical informatics* 7, 2 (2019), e13445.

[2] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental Models of Mere Mortals with Explanations of Reinforcement Learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–37.

[3] AC Berry, BD Cash, B Wang, MS Mulekar, AB Van Haneghan, K Yuquimpo, A Swaney, MC Marshall, and WK Green. 2019. Online symptom checker diagnostic and triage accuracy for HIV and hepatitis C. *Epidemiology & Infection* 147 (2019).

[4] Andrew C Berry, Brooks D Cash, Madhuri S Mulekar, Bin Wang, Anne Melvin, and Bruce B Berry. 2017. Symptom checkers vs. doctors, the ultimate test: a prospective study of patients presenting with abdominal pain. *Gastroenterology* 152, 5 (2017), S852–S853.

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Susan E Brennan. 1990. Conversation as direct manipulation: An iconoclastic view. *The art of human-computer interface design* (1990), 393–404.

[7] CDC. 2020. *Coronavirus Self-Checker*. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/coronavirus-self-checker.html

[8] Duncan Chambers, Anna J Cantrell, Maxine Johnson, Louise Preston, Susan K Baxter, Andrew Booth, and Janette Turner. 2019. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ open* 9, 8 (2019), e027743.

[9] Christina Cheng and Matthew Dunn. 2015. Health literacy and the Internet: a study on the readability of Australian online health information. *Australian and New Zealand journal of public health* 39, 4 (2015), 309–314.

[10] J Copestake. 2018. Babylon claims its chatbot beats GPs at medical exam. BBC News 2018 Jun 27. https://www.bbc.com/news/technology-44635134

[11] Benjamin Marshall Davies, Colin Fraser Munro, and Mark RN Kotter. 2019. A novel insight into the challenges of diagnosing degenerative cervical myelopathy using web-based symptom Checkers. *Journal of medical Internet research* 21, 1 (2019), e10868.

[12] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068.

[13] Cecilia di Sciascio, Peter Brusilovsky, and Eduardo Veas. 2018. A study on user-controllable social exploratory search. In *23rd International Conference on Intelligent User Interfaces*. ACM, 353–364.

[14] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 408–416.

[15] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

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

[17] Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. *AI Magazine* 32, 3 (2011), 90–98.

[18] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *Interna- tional Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[19] Joe Greek. 2017. *Artificial Intelligence: Clever Computers and Smart Machines*. The Rosen Publishing Group, Inc.

[20] 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.

[21] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[22] Kristen Hanich, Yilan Jiang, and Pooja Kamble. 2019. Virtual Care and Remote Monitoring: Connected Health at Home. https://www.bbc.com/news/ technology-44635134

[23] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183..

[24] Ada Health. 2020. *About the Ada COVID-19 assessment and screener https:///*ada.com/covid-19-screener/

[25] Sutter Health. 2019. *Sutter Health Teams Up With Ada Health to Improve Patient Care with On-Demand Healthcare Guidance*. http://www.mobilehealthtimes.com/sutter-health-teams-up-with-ada-health-to-improve-patient-care-by-delivering-on-demand-healthcare-guidance/

[26] Mathijs P Hendriks, Xander AAM Verbeek, Thijs van Vegchel, Maurice JC van der Sangen, Luc JA Strobbe, Jos WS Merkus, Harmien M Zonderland, Carolien H Smorenburg, Agnes Jager, and Sabine Siesling. 2019. Transformation of the National Breast Cancer guideline into data-driven clinical decision trees. *JCO clinical cancer informatics* 3 (2019), 1–14.

[27] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[28] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[29] 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? (2017). arXiv:1712.09923

[30] M Shamim Hossain, Ghulam Muhammad, and Nadra Guizani. 2020. Explainable AI and mass surveillance system-based healthcare framework to combat COVID- I9 like pandemics. *IEEE Network* 34, 4 (2020), 126–132.

[31] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2016. User control in recommender systems: Overview and interaction challenges. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 21–33.

[32] Timothy J Judson, Anobel Y Odisho, Aaron B Neinstein, Jessica Chao, Aimee Williams, Christopher Miller, Tim Moriarty, Nathaniel Gleason, Gina Intinarelli, and Ralph Gonzales. 2020. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. *Journal of the American Medical Informatics Association* 27, 6 (2020), 860–866.

[33] Annemarie Jutel and Deborah Lupton. 2015. Digitizing diagnosis: a review of mobile applications in the diagnostic process. *Diagnosis* 2, 2 (2015), 89–96.

[34] Cheng-Kai Kao and David M Liebovitz. 2017. Consumer mobile health apps: current state, barriers, and future directions. *PM&R* 9, 5 (2017), S106–S115.

[35] Boris Kment. 2006. Counterfactuals and explanation. *Mind* 115, 458 (2006),261–310.

[36] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 43–50.

[37] Bart P Knijnenburg and Martijn C Willemsen. 2015. Evaluating recommender systems with user experiments. In *Recommender Systems Handbook*. Springer, 309–352.

[38] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[39] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User preferences for hybrid explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 84–88.

[40] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 379–390.

[41] 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.

[42] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. 2018. All Work and No Play?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[43] 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.

[44] 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 & Medicine* 133 (2015), 128–135.

[45] 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 2019 Conference on Intelligent User Interface*. ACM, 1–12.

[46] 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.

[47] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[48] Biswajit Mohanty, Abrar Chughtai, and Fethi Rabhi. 2019. Use of Mobile Apps for epidemic surveillance and response–availability and gaps. *Global Biosecurity* 1, 2 (2019).

[49] University of Iowa Hospitals & Clinics. 2020. *Check Your COVID-19 Risk Online*. https://uihc.org/check-your-covid-19-risk-online

[50] 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.

[51] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.

[52] Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly. 2020. Explainable ai in healthcare. In *2020 International Conference on Cyber Situational Awareness, Data*

Analytics and Assessment (CyberSA). IEEE, 1–2.

[53] Lucy Powley, Graham McIlroy, Gwenda Simons, and Karim Raza. 2016. Are online symptoms checkers useful for patients with inflammatory arthritis? *BMC musculoskeletal disorders* 17, 1 (2016), 362.

[54] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556.

[55] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 157–164.

[56] 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.

[57] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI.. In *IUI Workshops*.

[58] David-Hillel Ruben. 2015. *Explaining explanation*. Routledge.

[59] Bahae Samhan. 2019. Self-Diagnosis Mobile Applications A Technology Trust Perspective. (2019).

[60] Briane Paul V Samson and Yasuyuki Sumi. 2020. Are Two Heads Better than One? Exploring Two-Party Conversations for Car Navigation Voice Guidance. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[61] 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.

[62] 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), h3480.

[63] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE 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.

[64] Amit Sharma and Dan Cosley. 2013. Do social explanations work?: studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1133– 1144.

[65] Carl Shen, Michael Nguyen, Alexander Gregor, Gloria Isaza, and Anne Beattie. 2019. Accuracy of a popular online symptom checker for ophthalmic diagnoses. *JAMA ophthalmology* 137, 6 (2019), 690–692.

[66] Vilma Shu. 2020. *Penn State Go: COVID-19 symptom checker part of new faculty/staff experience*. https://news.psu.edu/story/627659/2020/08/05/administration/penn-state-go-covid-19-symptom-checker-part-new-facultystaff

[67] Kacper Sokol and Peter A Flach. 2018. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant.. In *IJCAI*. 5868–5870.

[68] Lee Sproull, Mani Subramani, Sara Kiesler, Janet H Walker, and Keith Waters. 1996. When the interface is a face. *Human-computer interaction* 11, 2 (1996), 97–124.

[69] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (1 Oct. 2012), 399–439.

[70] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

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

[72] Chun-Hua Tsai. 2020. *Controllability and explainability in a hybrid social recom- mender system*. Ph.D. Dissertation. University of Pittsburgh.

[73] Chun-Hua Tsai and Peter Brusilovsky. 2018. Beyond the Ranked List: User-Driven Exploration and Diversification of Social Recommendation. In *23rd International Conference on Intelligent User Interfaces*. ACM, 239–250.

[74] Chun-Hua Tsai and Peter Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 391–396.

[75] Chun-Hua Tsai and Peter Brusilovsky. 2020. The Effects of Controllability and

Explainability in a Social Recommender System. *User Modeling and User-Adapted Interaction* (2020).

[76] Florida State University. 2020. *Stay Healthy Training & Daily Wellness Check Application*. https://news.fsu.edu/news/2020/08/12/stay-healthy-training-daily- wellness-check-application/

[77] Quinnipiac University. 2020. *University introduces QU Symptom Checker app to keep campus community safe*. https://www.qu.edu/today/qu-symptom-checker.html

[78] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI confer- ence on human factors in computing systems*. 1–15.

[79] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: beyond the black box. *Bmj* 364 (2019).

[80] Martin Wiesner and Daniel Pfeifer. 2014. Health recommender systems: concepts, requirements, technical basics and challenges. *International journal of environmental research and public health* 11, 3 (2014), 2580–2607.

[81] Aaron N Winn, Melek Somai, Nicole Fergestrom, and Bradley H Crotty. 2019. Association of Use of Online Symptom Checkers With Patients' Plans for Seeking Care. *JAMA Network Open* 2, 12 (2019), e1918561–e1918561.

[82] 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).

[83] 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.

[84] 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 (TOCHI)* 27, 3 (2020), 1–37.

[85] Yue You and Xinning Gui. 2020. Self-Diagnosis through AI-enabled Chatbot-

based Symptom Checkers: User Experiences and Design Considerations. In *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association.

A DEMOGRAPHIC INFORMATION

The demographic information are shown in Table 3 and 4.

B CLINICAL DECISION TREE (CDT)

Symptom checkers are widely adopted in the medical domain to provide diagnostic and triage to patients or the general public [62]. These medical-related recommendations could be generating by a data-driven clinical decision tree (CDT) [26]. We create a CDT for COVID-19 based on the CDC recommendations [7] (shown in Table 5). We manually tested all possible symptom combinations and stored the CDC recommendations for building the CDT model. In the CDT, we defined 10 rules based on symptoms or contact history question combinations. The rules are determined by three COVID-19 related symptom questions and four COVID-19 contact history questions. The detailed description is listed below.

- COVID-19 Symptom Questions

1. "**Feeling Ill**": This question asked if the user feel sick, e.g., "*are you feeling ill? (Yes/No)*"
2. "**Showing Symptom**": This question asked if the user have any symptoms related to the coronavirus, e.g., "*Do you have any of the following? [fever], [cough], [moderate difficulty breathing], [sore throat], [muscle aches or body aches], [vomiting or diarrhea] or [new loss of taste/smell] (Yes/No).*" This question would repeat seven times to ask the symptom independently.
3. "**Medical Condition**": This question asked if the user has any underlying medical conditions that required fur- ther attention, e.g., "*do any of these apply to you? [chronic lung disease, diabetes], [high blood pressure] (Yes/No)*" This question would repeat two times to ask the user for two different sets of conditions.

- COVID-19 Contact History Questions

1. "**2-Week Contact**": This question asked if the user has close contact with COVID-19 confirmed cases, e.g., "*In the last two weeks, did you care for or have close contact with someone diagnosed with COVID-19? (Yes/No)* "

2. "**Nursing Home**": This question asked if the user belongs to the population who are at high risk, e.g., "*do you live in a long-term care facility or nursing home?*"

3. "**Healthcare Facility**": This question asked if the user worked in high-risk workplaces, e.g., "*In the last two weeks, have you worked or volunteered in a healthcare facility or as*

a first responder? Facilities include a hospital, other medical settings (including dental care setting), or long-term care facilities. (Yes/No)"

4. "**PPE**": This question asked if the user has proper personal protective equipment while working in the high-risk workplace, e.g., "*Did you wear all recommended personal protective equipment while you were in close contact with someone diagnosed with COVID-19? (Yes/No)*"

Our online COVID-19 symptom checker was designed to pro- vide the COVID-19 diagnostic based on the CDT table. For example, if there is a healthcare worker (**Healthcare Facility**=Y) who has close contact with COVID-19 confirmed case (**2-Week Contact**=Y), but he/she is not feeling ill (**Feeling Ill**=N) and not live in a nursing home (**Nursing Home**=N) and wearing shield and mask (**PPE**=Y) in the workplace. The user input data would fit *Rule 3* that the diagnostic would be he/she does not require immediate medical attention (**Medical Attention**=N) but eligible for COVID-19 testing (**COVID-19 Testing**=Y). To be noted, the *dash symbol (-)* means the questions are not needed for the rules. For instance, if a user is not felling ill (**Feeling Ill**=N) and has no COVID-19 contact history (**2-Week Contact**=N), the user input fulfill *rule 2* that need neither immediate medical attention nor COVID-19 testing, no more information is required from the user.

Table 3: Study 1 Demographic Information

Participation ID	Profession	Used Checkers	Gender	Age
S1	Graduate	Ada; K health; your.MD	F	24
S2	Graduate	Ada; K health; your.MD; ask NHS	F	29
S3	University Staff	mayoclinic; webmed; K health	F	24
S4	Graduate	K health; your.MD	M	24
S5	Graduate	K health; Ada	F	28
S6	Graduate	K health; Ada	F	25
S7	Graduate	K health	F	26
S8	Software Engineer	K health; Ada	F	26
S9	Graduate	K health	M	27
S10	Faculty	K health; Ada; your.MD	F	40
S11	Undergraduate	Ada	M	24
S12	Graduate	K health; your.MD	F	26
S13	University Staff	K health; WebMD; Health on demand	F	54
S14	Landscape Designer	K health; Ada	F	25
S15	Graduate	K health; Ada; WebMD	F	30
S16	Graduate	K health; your.MD	M	25
S17	University Staff	WebMD; Ada; K health	F	30
S18	Graduate	Ada; K health	M	26
S19	Undergraduate	WebMD; Ada; K health	M	23
S20	Undergraduate	WebMD; Ada; K health	F	22
S21	Graduate	Babylon; K health	F	26
S22	Graduate	K health	M	26
S23	Undergraduate	WebMD; K health	F	19
S24	Undergraduate	K health; WebMD	M	20
S25	Graduate	K health; WebMD	M	23

Table 4: Study 2 Demographic Information

Participation ID	Program	Major	Gender	Age
P1	Graduate	Information Science and Technology	F	27
P2	Undergraduate	Applied Data Science	M	21
P3	Graduate	HCI	F	29
P4	Graduate	Information Science and Technology	F	24
P5	Graduate	Information Science and Technology	F	23
P6	Graduate	Information Science and Technology	M	30
P7	Graduate	Information Science and Technology	F	24
P8	Graduate	Information Science and Technology	M	29
P9	Graduate	Forest Resources	M	26
P10	Undergraduate	Mechanical Engineering	F	21
P11	Graduate	Learning Design and Technology	F	27
P12	Graduate	Finance	F	27
P13	Graduate	Information Science and Technology	F	37
P14	Undergraduate	Human-Centered Design & Development	M	22
P15	Graduate	Biology	F	29
P16	Graduate	Psychology	F	24
P17	Undergraduate	Material Science And Engineering	M	20
P18	Graduate	Applied Statistics	F	25
P19	Undergraduate	Computer Science	F	19
P20	Graduate	Industrial Engineering	F	29

Table 5: Clinical Decision Tree (CDT) for COVID-19 OSC

Rule	Symptom / Contact History Questions							Decisions	
	Feeling Ill	2-Week Contact	Showing Symptom	Nursing Home	Healthcare Facility	PPE	Medical Condition	Medical Attention	COVID-19 Testing
1	N	Y	-	Y	-	-	-	N	N
2	N	N	-	-	-	-	-	N	N
3	N	Y	-	N	Y	Y	-	N	Y
4	N	Y	-	N	N	-	-	N	N
5	N	Y	-	N	Y	N	-	Y	Y
6	Y	Y/N	>1	Y	-	-	-	Y	Y
7	Y	Y/N	>1	N	Y	-	-	Y	Y
8	Y	Y/N	>1	N	N	-	0	N	Y
9	Y	Y/N	>1	N	N	-	>1	Y	Y
10	Y	Y/N	0	-	-	-	-	N	N

C POST-STAGE QUESTIONNAIRE

1. Quality: To measure if the users perceive the good medical recommendations.

Q1: The app provides good medical recommendations for me.

- *Q2: I like the health recommendations provided by the app.*

2. Control: To measure if the system is easy to use.

Q3: I become familiar with the app very quickly.

- *Q4: The app helped me to make medical decisions faster.*

3. Effectiveness: To measure if the system helps the users make good decisions.

Q5: The app helps me to make better medical choices.

Q6: I find useful medical recommendations using the app.

4. Trust: To measure the user confidence in the system.

Q7: I am convinced by the medical suggestions that the app recommended to me.

Q8: The app can be trusted.

Q9: The app provides sufficient medical recommendations for me to make a good medical decision.

5. Transparency: To measure if the system explain how the system works.

Q10: The app explains why medical recommendations were recommended to me.

- *Q11: I understand why the medical recommendations were recommended to me.*

6. Satisfaction: To measure if the system is ease of use or enjoyment.

Q12: Overall, I am satisfied with the app Q13: I will use this app again.

- *Q14: I would like to share this app with my friends or col- leagues.*

7. Awareness: To measure if the users aware more about COVID-19 after use the system.

Q15: The app makes me more aware of my body conditions regarding COVID-19e.

- *Q16: The app provides medical recommendations as I expected.*

8. Learning: To measure the learning effects after use the system.

Q17: The app helps me to better communicate COVID-19 related information (e.g., preventive measures, testing policy, etc.) with others.

- *Q18: The app helps me to learn COVID-19 related information (e.g., preventive measures, testing policy, etc.*

Except the *control* construct, all statistics summarized here sup- ports good internal consistency (Cronbach's α), and the descriptive statistics for the composite variables are available in Table 1. We also reported the NASA Task Load Index (NASA-TLX) scores [20, 23] in Table 2.

D POST-EXPERIMENT INTERVIEW QUESTIONS

1. Which COVID-19 OSC do you prefer?
2. Why do you prefer this COVID-19 OSC to others?
3. Which COVID-19 OSC is enough for you to understand and diagnose your conditions?
4. Which COVID-19 OSC do you trust?
5. Do you know how or why does the COVID-19 OSC generates the results?
6. How do you assess COVID-19 OSC's diagnoses?
7. What benefits do you perceive in the COVID-19 OSCs?
8. What negative consequences are associated with using the COVID-19 OSCs?
9. What challenges do you encounter when using the COVID- 19 OSCs?
10. What functions would you like to have (if any)?