Leveraging challenges of an algorithm-based symptom checker on user trust through explainable AI

YOUJIN HWANG, hci+d lab.,Seoul National University, Republic of Korea
TAEWAN KIM, DxD lab., Seoul National University, Republic of Korea
JUNHAN KIM, School of Information, University of Michigan, USA
JOONHWAN LEE, hci+d lab.,Seoul National University, Republic of Korea
HWAJUNG HONG, DxD lab., Seoul National University, Republic of Korea

An algorithm-based symptom checker is a service that predicts and informs the expected disease name based on the symptoms entered by users and informs the user of actions to be taken afterward. Few studies have been done on the perception and algorithm experience with the symptom checker at the best of our knowledge even though user-centered research on how users interact with algorithms and interpret the algorithmic results has recently emerged largely in the field of HCI. In this position paper, we share our results from an empirical study defining challenges that prevent user trust toward algorithm-based symptom checkers. Based on the identified challenges, we suggest design ideas for implementing explainable AI (XAI) to algorithm-based symptom checkers to enhance users' understanding and trust in existing symptom checkers.

CCS Concepts: • **Human-centered computing** → *Empirical studies in interaction design.*

Additional Key Words and Phrases: symptom checkers, algorithm experience, explainable AI, healthcare

ACM Reference Format:
Youjin Hwang, Taewan Kim, Junhan Kim, Joonhwan Lee, and Hwajung Hong. 2018. Leveraging challenges of an algorithm-based symptom checker on user trust through explainable AI. 1, 1 (April 2018), 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Due to the increased accessibility to health information on the Internet, many people perform self-diagnosis through the Internet before visiting the doctor [10]. In the United States, it has been reported that 1 in 3 adults has used the Internet or other tools for self-diagnosis [7]. Moreover, as the demand for remote diagnosis increases due to COVID-19, the importance of self-diagnosis services is further heightening. In particular, the popularization of algorithm-based symptom checkers, which have emerged with the recent development of artificial intelligence technology, has allowed people to reduce side effects caused by inaccurate information that can be encountered when searching for health information on the Internet.

Authors' addresses: Youjin Hwang, youjin.h@snu.ac.kr, hci+d lab.,Seoul National University, Republic of Korea; Taewan Kim, DxD lab., Seoul National University, Republic of Korea, taewankim@snu.ac.kr; Junhan Kim, School of Information, University of Michigan, USA; Joonhwan Lee, hci+d lab.,Seoul National University, Republic of Korea; Hwajung Hong, DxD lab., Seoul National University, Republic of Korea, hwajunghong@snu.ac.kr.

An algorithm-based symptom checker is a service that predicts and informs the expected disease name based on the symptoms entered by users and informs the user of actions to be taken afterward (e.g., Mayo Clinic symptom checker [4], Babylon Health [1], the Ada health app [2], and the K Health app [3]). Prior studies based on user surveys have been conducted to understand the symptom checker's usability and usage behavior. These studies deal with a variety of services, from web-based online symptom checkers to smartphone app-based symptom checkers. And mainly dealing with the performance of the symptom checker (that is, how accurately it predicts disease), and the motivation to use [12, 14].

However, there are few studies on the perception and experience of users related to the algorithm of the symptom checker. User-centered research on how users interact with algorithms and interpret the algorithmic results has recently emerged largely in the field of HCI [5, 6]. Users' perceptions and experiences of algorithms are important factors influencing their acceptance of algorithm-based outcomes and determining their future behavior [8]. Particularly in disease diagnosis, understanding how users accept algorithmic results (that is, diagnosis results) and use them in future actions can be of great importance and impact because they are directly associated with their health. This study attempts to understand this through exploratory research on users' algorithmic experience related to symptom checkers.

As one of the ways to increase the user trust of the AI system's algorithmic experience, research on artificial intelligence that can be explained is increasing. In particular, because of the black box structure of AI systems, it is difficult for lay users to obtain information on its mechanism of action. Therefore, designing an explainable artificial intelligence (XAI) has a positive effect on the user's algorithm experience [11]. According to a general framework for designing explainable AI, evaluation of explainability should be evaluated in terms of (1) fidelity, (2) completeness, and (3) robustness [13]. However, there is insufficient research on matters to be considered for a specific domain, such as self-diagnosis through symptom checkers. Our study attempts to understand the user's XAI needs for the algorithm through the use of symptom checkers.

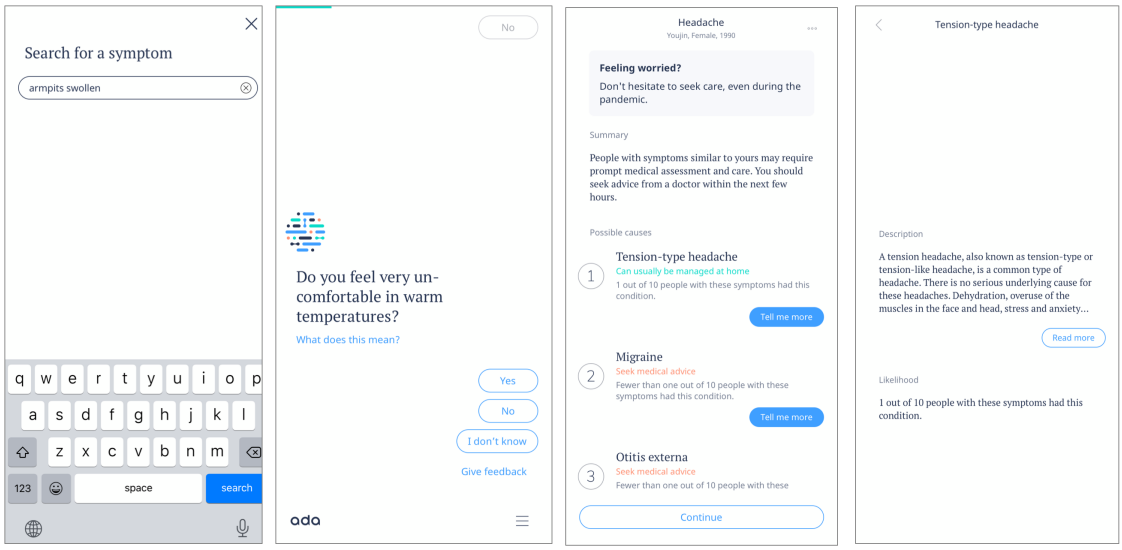
Based on this background, we set up the following research questions.

- What are the challenges users face with the algorithm experience when using the symptom checker?
- What are the opportunities of implementing explainable AI in increasing user trust toward algorithm-based symptom checker?

In this position paper, we discuss the design for improving users' trust in the symptom checker by implementing XAI. First, we discuss the challenges of using algorithm-based symptom checkers that affect user trust. Through this discussion, we identify design considerations for improving users' understanding and trust in the algorithm. We also briefly share our plans for the future study to discuss its needs and validity with the domain experts who are going to participate in the workshop.

2 EMPIRICAL STUDY

To define the challenges that prevent user trust toward symptom checkers, we conducted an empirical study with ADA application which is an algorithm-based symptom checker. Through ADA, users report their symptoms, and based on the reported symptoms, ADA matches them with symptoms of patients of similar age and gender and reports the statistical likelihood that the patient has a certain condition. We recruited 6 participants aged between 20 and 39 without any severe diseases (Female:4, Male:2) for the empirical study. Only one participant involved in the user study at the same time. We recruited participants who were capable of fluent communication through English since the ADA application only offers users English-based interaction. To empirically explore the user experience with ADA, we used



(a) Symptom searching (b) Answering questions (c) Result page _overall (d) Result page _ detail

Fig. 1. Screenshots of ADA. The interaction with ADA includes (1) symptom searching, (b) answering symptom-checking questions, (c) result page with disease prediction, and (d) result page with detailed information with a particular disease from (c). Through the empirical study, we have defined the challenges of algorithm-based symptom checkers that affect user trust. Then, we discuss the design for improving users' trust in the symptom checker by implementing explainable AI (XAI).

the think-aloud method to collect user data. Particularly, we asked them to keep focusing on the moment that affects their trust toward ADA. Also, we used clinical vignettes to assess ADA with participants without severe diseases to avoid ethical issues raised from involving patients in the study who are suffering from serious symptoms. Instead, two sample vignettes were identified from a previous study that explores the user behaviors of online, web-based self-diagnosing [15]. The vignettes depicted the symptoms of an acute health condition: mononucleosis or scarlet fever. These conditions were selected as we expected participants would not easily predict the diagnosis outcome.

Participants received one of two vignettes that depicted symptoms of illness. Participants talked out loud about their thoughts and actions while attempting to diagnose the symptoms. Participants were limited to 30 minutes of search time. All completed the task between 20 and 30 minutes. The participant was audio-recorded during the think-aloud to allow for later transcription and analysis. Transcribed data collected through think-aloud content was analyzed by authors through thematic analysis [9]

3 CHALLENGES THAT AFFECT USERS' TRUST ON ALGORITHM-BASED SYMPTOM CHECKER

In this section, we discuss challenges in using algorithm-based symptom checkers that affect user trust. The results below are based on the empirical user study we explained in the previous section.

3.1 Trust on algorithm-based question

The symptom checking scenario of the ADA is based on the set of questions that appear depending on the user's previous answers on detailed symptoms. This set of questions are necessary to distinguish the user's symptom from other similar symptoms. However, when it comes to user experience with ADA, participants all complained that the

presented questions seemed to be repeated several times and unnecessary. Participants described these experiences as "disturbing", "not understanding user's urgent situation", "not listening to users", and described the system as "not an expertise". In addition, most participants reported that questions seemed to be not relevant to the actual symptoms. Some participants questioned the system's intelligence when they confronted the irrelevant questions. For example, one of the participants reported that she answered the questions related to the fever-related symptoms, but the following question asked her about dental questions.

3.2 Trust on answering option

When answering questions that the ADA asked, participants showed hesitations when choosing answers among the answering options. Some participants said that answering options do not fully satisfy their situations. Also, some participants reported that the answering options are vague, and more detailed explanations are needed for them. For example, P2 was frustrated when she confronted the question which is "do you recognize significant weight loss in recent days?". She complained that she could not choose the answer among "Yes","no", and "I don't know". She said that she wanted to answer "3kg" because she was not sure that 3kg weight loss in 2 months is significant or not. She suggested the idea of giving users the explanation for significant weight loss or providing example cases to support accurate symptom explanation. Another example provided by P3 was the question of "do you feel very uncomfortable in warm temperature?". She was confused with the word "very uncomfortable" since as she said "I am not sure that how much uncomfortable is very uncomfortable." Additionally, there was reported difficulty in searching for a symptom through the searching bar as in Figure. P4 complained that "I searched swollen armpits through the searching bar, but nothing really came out as a result". This experience made me very confusing and lost some trust in the system. I am not sure whether it will really help me diagnose my status or not." Also, the need for natural language input was raised from some participants.

3.3 Trust on disease prediction

After the user completes symptom-checking questions, ADA presents the high related diseases based on the prediction model. It shows lists of possible causes (i.e. lists of possible diseases) of the symptoms. Also, ADA presents a description of presented diseases including risks, symptoms, diagnosis, treatment, prevention, and prognosis, and the likelihood of the general population who suffer from the disease. Most participants expressed uncertainty about the prediction results from the symptom checker. They insisted that the lists of final results seem not relevant to their reported symptoms. P5 said that he needs an additional explanation for why he got the final results and why provided lists of the diseases are relevant to his case. Moreover, some participants insisted that the system lacks personalization. P6 reported that even though her age is over 30, the results showed the disease that teenagers are most likely to suffer. Also, most participants questioned the types of data used to train the ADA's prediction model since the likelihood of particular diseases sometimes differs depending on race, sex, age etc.

4 DESIGN SUGGESTIONS FOR ENHANCING TRUST THROUGH EXPLAINABLE AI (XAI)

Based on the identified challenges that affect the user's trust in algorithm-based symptom checkers, we suggest implementing XAI to algorithm-based symptom checkers to enhancing users' understanding and trust in the algorithm. The goal of our design suggestions is to aid users to make informed judgments rather than accept information uncritically. We suggest three major designs for implementing XAI to an algorithm-based symptom checker.

4.1 XAI for question reasoning

To improve user experience with the reported challenges from participants including irrelevant and repetitive similar questions, we suggest implementing XAI for reasoning why particular questions appear during the symptom-checking interaction. For example, we suggest adding agent-based communication to implement XAI to the symptom checker. The agent does not need to proactively make a conversation but should be designed to be always available for the users. The role of the agent should be explained prior to the user to reduce confusion. By doing so, we expect the user can ask the agent any time they want to skip the question, or anytime they want to be explained why they are being asked such questions.

4.2 XAI for assisting symptom report

We found some pain points of the user in reporting their symptoms through ADA that are insufficient answering options and frequent cases of no option for particular symptoms that the user searched through the searching bar. Reporting symptoms is challenging not only in human-agent communication but also in human-human communication. Experts (e.g.doctors, clinical experts, etc) in the real world support patients' symptom reports by giving patients some of the possible symptom options for their cases while explaining why they suggest those options. Similarly, XAI could act as a recommender while users report their symptoms.

4.3 XAI for result counseling

We assume that result counseling would be the most needed function for the user who is using an algorithm-based symptom checker since all participants questioned why they happen to get the corresponding results. There should be explanations for how the algorithm predicts diseases based on the reported symptoms. Some participants even insisted that the final part of ADA interaction that shows disease prediction results is the part that most affected their trust toward ADA. Therefore, a transparent explanation for the result is needed to improve user understanding and trust in the algorithm of the symptom checkers.

5 CONCLUSION AND FUTURE WORK

We have seen how participants encountered certain obstacles in terms of explainable AI when using the algorithm-based symptom checker, ADA. We have also explained how participants had trouble trusting the system when they were unclear about how the results were processed. We are now planning to develop this idea into a design workshop. The purpose of the workshop is to understand how a symptom checker like ADA should be designed In the workshop, participants will have the opportunity to re-design ADA themselves in order to make the system more transparent and explainable. Especially because many of the themes identified through the exploratory study above were related to trust, we aim to target how users would want to redesign symptom checkers for them to be more trustworthy.

We would like to share with workshop participants about the possible social implications of our study. Symptom checkers are important in that they are capable of delivering a potential medical diagnosis to those with low access to healthcare such as rural communities. Moreover, the transparency and explainability of the system may provide patients a sense of self-awareness because they would be able to critically evaluate the information given by the symptom checkers, without blindly submitting to its suggestions. Through the insights gathered from the study, we hope to have a fruitful discussion with professionals working in the similar domain.

6 AKNOWLEGEMENT

This work was supported by AI Institute of Seoul National University (AIIS) through its AI Frontier Research Grant in 2020.

REFERENCES

[1] [n.d.]. Babylon Health UK The Online Doctor and Prescription Services. https://www.babylonhealth.com/
[2] [n.d.]. Health. Powered by Ada. https://ada.com/
[3] [n.d.]. K Health. https://khealth.ai/
[4] [n.d.]. Mayo Clinic - Mayo Clinic. https://www.mayoclinic.org/
[5] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–12.
[6] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, communication & society* 20, 1 (2017), 30–44.
[7] Michelle Castillo. 2013. More than one-third of U.S. adults use Internet to diagnose medical condition. https://www.cbsnews.com/news/more-than-one-third-of-us-adults-use-internet-to-diagnose-medical-condition/
[8] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. 2371–2382.
[9] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1968. The discovery of grounded theory; strategies for qualitative research. *Nursing research* 17, 4 (1968), 364.
[10] Lisa Neal Gualtieri. 2009. The doctor as the second opinion and the internet as the first. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. 2489–2498.
[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
[12] Stefanie Maria Jungmann, Timo Klan, Sebastian Kuhn, and Florian Jungmann. 2019. Accuracy of a Chatbot (ADA) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR formative research* 3, 4 (2019), e13863.
[13] 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.
[14] Tana M Luger, Thomas K Houston, and Jerry Suls. 2014. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. *Journal of medical Internet research* 16, 1 (2014), e16.
[15] Tana M Luger, Thomas K Houston, and Jerry Suls. 2014. Older Adult Experience of Online Diagnosis: Results From a Scenario-Based Think-Aloud Protocol. *J Med Internet Res* 16, 1 (16 Jan 2014), e16. https://doi.org/10.2196/jmir.2924