Heuristic Evaluation of COVID-19 Chatbots

Sviatlana Höhn[1,2,3][0000−0003−0646−3738] and Kerstin Bongard-Blanchy[1,4][0000−0001−9139−1622]

[1] University of Luxembourg, Esch-sur-Alzette, Luxembourg
[2] sviatlana.hoehn@uni.lu
https://chatbots.uni.lu
[3] Supported by the Luxembourg National Research Fund (FNR), SLANT, 13320890
[4] kerstin.bongard-blanchy@uni.lu

Abstract. Chatbots have been adopted in the health domain and their number grew during the current COVID-19 pandemic. As a new kind of interface, chatbots combine visual elements with natural conversation. While conversational capabilities of chatbots improve, little attention has been given to the evaluation of the user experience and chatbot usability. This paper presents the results of a heuristic review of 24 COVID-19 chatbots on different channels (webchat vs messengers), for diverse topics (symptom-checker vs FAQ) and with varying interaction styles (visual-centric vs content-centric vs conversation-centric). It proposes a generic evaluation framework with 12 heuristics based on Nielsen's ten heuristics and adapted to the conversational interface context. The results point at the strengths (immediate feedback, familiar language, consistent wording and visual design) as well as shortcomings (little user control and freedom, missing permanent menu and help options, lack of context understanding and interaction management capabilities) of COVID-19 chatbots. The paper furthermore gives recommendations for chatbot design in similar contexts.

Keywords: Conversational UX Analysis · Chatbots · Healthcare.

1 Introduction

Chatbots specialised in COVID-19 matters have been developed to help people cope with the pandemic. Authorities like the WHO, CDC, Ministries of Health of different countries, Red Cross, hospitals and insurance companies provide free of charge chatbots that talk about Coronavirus. Among them are bots for symptom checking [16], for information about emergencies in the region and world, for psychological distress monitoring [5], and artificial business advisors [15]. Tech companies provide the required infrastructure and templates [22].

Although research on dialogue systems, including robots, chatbots and voice assistants, has advanced in many aspects, such conversational interfaces still pose significant challenges to researchers and designers in the human-computer interaction domain [3]. Consideration for chatbot user experience (UX) has gained momentum, starting with effort to adapt classical UX evaluation methods to the

chatbot context [9, 21], stretching to user interviews to analyse user needs and expectations [12]. However, basic principles of UX design are not yet commonly applied in the chatbot domain. Current appreciations of chatbots range hence from a *poor relative of an intelligent assistant that performs only one well-defined domain-specific task* [4] to a *fully-capable conversational software that maintains long-term interaction with its user via text messages* [6]. Moore and Arar (2019) [14] argue that chatbots today are similar to the Internet in 1997: made by laypeople based on a set of quickly self-acquired skills.

In this regard, deficits in the accuracy of medical symptom checkers have been found, together with strong risk-averse responses [18]. COVID-19 chatbots for symptom-checking show significant differences in their sensitivity and specificity [16]. However, inaccuracy and unsound conversational design in medical applications can be life-threatening [19].

This paper, therefore, seeks to evaluate the usability of 24 COVID-19 chatbots to answer the **research questions**:

1. What types of COVID-19 chatbots exist?
   (a) On which channels are they available?
   (b) Which service, content or topic within the COVID-19 area do they offer?
   (c) Which interaction styles do they use?
2. How *usable* are COVID-19 chatbots?

Following an overview of related work, Section 3 presents 39 evaluation aspects grouped under 12 heuristics and explains the evaluation procedure. Section 4 provides insights in content, topics, channels and conversation styles of COVID-19 chatbots and presents the heuristic evaluation results. Section 5 discusses the strengths and weaknesses of the tested bots, Finally, Section 6 formulates recommendations for satisfying conversational UX, especially in e-health domain.

The paper contributes conversational UX analysis with a new framework for the evaluation of conversational interfaces that, in contrast with most recent scholar work [21], covers chatbots of *all* interaction styles. The new framework helps to formulate design recommendations for conversational interfaces.

2 Related Work

Multiple objective and subjective metrics for evaluation of conversational interfaces have been developed within the last two decades by major international initiatives; see, for instance, McTear et al. (2016)[13, Chap. 17]. Objective methodologies cover UX aspects in the best case by the notion of "user satisfaction". The most prominent objective methodology for spoken dialogue system evaluation, PARADISE, dates from the late nineties [25]. Messenger APIs and widgets for interaction management by bots in messengers (e.g. carousel) were not existent by that time. The PARADISE framework has also been used for the prediction of user satisfaction. User satisfaction is expected to be high if the task success is maximised while the dialogue costs are minimised. Methods for

subjective evaluation include the Subjective Assessment of Speech System Interfaces (SASSI) questionnaire [10]. It builds on 34 criteria such as system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed.

More recent scholar initiatives suggest to study chatbots from the perspective of *conversational UX design*, see for instance contributions at CHI 2017 conversational UX Design Workshop[5]. Researchers seek to formulate principles and guidelines for conversational UX Design as a distinct discipline.

Moore and Arar (2019) recommend using conversation analysis to improve conversational design. They classify natural language interfaces by their *interaction styles*: system-centric (e.g. voice control, web search; require valid, technical input), content-centric (e.g. FAQ; document-like responses), visual-centric (e.g. desktop or mobile interfaces; use buttons and require direct manipulation) and conversation-centric (similar to natural conversation) [14, p. 16]. The styles are not disjoint: a content-centric chatbot for document retrieval that understands free text input can use buttons for short replies. Buttons increase the speed and efficiency of use; and these two factors have been reported to be the most important reasons for using chatbots [2].

While the conversational UX Design community formulated many guidelines on how to design chatbots [20, 14, 8], only a few researchers have so far undertaken conversational UX evaluation [7, 9, 21]. Nielsen's (2005) ten heuristics are frequently used to analyse the usability of user interfaces [17] and they have already been employed for chatbot UX analysis [23]. However, the applicability of this UX evaluation approach to conversational interfaces is subject of scientific debate. While Holmes et al. (2019) [9] found the conventional usability evaluation methods not suitable for the evaluation of chatbots, Sugisaki and Bleiker (2020) argue that Nielsen's (2005) approach provides a sound basis for the chatbot domain [21]. Their, most recent, detailed framework for the evaluation of conversational UX contains 53 so-called checkpoints that cover the ten Nielsen heuristics adapted to conversational interfaces.

The framework proposed by Sugisaki and Bleiker (2020) explicitly excludes chatbots that mainly use visual elements for interaction or only accept a precisely defined set of commands. For certain tasks and use cases, natural conversation is indeed the preferable interaction style. However, in other cases, the UX benefits from additional shortcuts, such as buttons and short replies. Many chatbots use both natural conversation and visual elements. That is why an evaluation framework that covers all types of chatbots, as proposed in this paper, is preferable because it allows the comparison of chatbots with different interaction styles.

3 Method

3.1 The 12 Heuristics for Conversational UX Analysis

Chatbots can combine visual elements with natural conversation. They hence require an adapted approach to usability evaluation. We defined the following

[5] https://researcher.watson.ibm.com/researcher/view_group.php?id=7539

12 heuristics to assess Conversational UX based on Nielsen's ten heuristics [17], Shevat's chatbot design guidelines [20] and Conversational UX design guidelines formulated by Moore and Arar [14].

1. **Visibility of system status**
   (a) Presence of information about the chatbot's state in the entire process
   (b) Immediate feedback (did the last user action work?)
   (c) Compel user action (what does the chatbot think the user will do next?)
2. **Match between system and the real world**
   (a) Chatbot uses the language familiar to the target users
   (b) Visual components (emojis, GIFs, icons) are linked to real-world objects
   (c) If metaphors are used, they are understandable for the user
3. **User control and freedom**
   (a) Chatbot supports undo/redo of actions
   (b) Chatbot offers a permanent menu
   (c) Chatbot provides navigation options
   (d) Chatbot understands repair initiations
4. **Consistency and standards**
   (a) Chatbot uses the domain model from the user perspective
   (b) Chatbot has a personality, consistency in language and style
5. **Error prevention**
   (a) Chatbot prevents unconscious slips by meaningful constraints
   (b) Chatbot prevents unconscious slips by spelling error detection
   (c) Chatbot requests confirmation before actions with significant implications
   (d) Chatbot explains consequences of the user actions
6. **Recognition rather than recall**
   (a) Chatbot makes the options clear through descriptive visual elements and explicit instructions
   (b) Chatbot shows summary of the collected information before transactions
   (c) Chatbot offers a permanent menu and help option
7. **Flexibility and efficiency of use**
   (a) Chatbot understands not only special instructions but also synonyms
   (b) Chatbot can deal with different formulations
   (c) Chatbot offers multiple ways to achieve the same goal
8. **Aesthetic and minimalist design**
   (a) Chatbot dialogues are concise, only contain relevant information
   (b) Chatbot uses visual information in a personality-consistent manner to support the user, not just random decoration
9. **Help users recognise, diagnose, and recover from errors**
   (a) Chatbot clearly indicates that an error has occurred
   (b) Chatbot uses plain language to explain the error
   (c) Chatbot explains the actions needed for recovery
   (d) Chatbot offers shortcuts to fix errors quickly
10. **Help and documentation**
    (a) Chatbot provides a clear description of its capabilities
    (b) Chatbot offers keyword search

(c) Chatbot focuses its help on the user task
(d) Chatbot explains concrete steps to be carried out for a task

11. **Context understanding**
    (a) Chatbot understands the context within one turn
    (b) Chatbot understands the context within a small number of turns (usually 2-3 user-bot turn pairs)
    (c) Chatbot understands the context of a multi-turn conversation
12. **Interaction management capabilities**
    (a) Chatbot understands conversation openings and closings (e.g., 'hello')
    (b) Chatbot understands sequence closings (e.g., 'ok' and 'thank you')
    (c) Chatbot understands repair initiations and replies with repairs
    (d) Chatbot initiates repair to handle potential user errors

3.2 COVID-19 Chatbots

Starting with ten English webchat symptom checkers analysed in [16], we searched on the Internet for "COVID-19 chatbots" and "Coronavirus chatbot". In this way we found 14 chatbots working also in messengers (Whatsapp, Telegram, Viber, Facebook Messenger) and added German, Russian, French and Ukrainian (languages spoken by authors of this paper). The following bots were inspected:

(1) **Ada** https://ada.com/COVID-19-screener/
(2) **Apple** https://www.apple.com/COVID19
(3) **Babylon** https://www.babylonhealth.com/ask-babylon-chat
(4) **Bobbi** https://www.berlin.de/corona/faq/chatbot/artikel.917495.php
(5) **CDC** https://www.cdc.gov/coronavirus/2019-nCoV/index.html
(6) **Cleveland Clinic** http://COVID19chat.clevelandclinic.org/
(7) **Corona Bot** CoronaBot.tn im Facebook Messenger
(8) **HSE Coronavirus Selfchecker** https://www.hse.ie
(9) **Covid-19 Chatbot** https://www.chatbot.com/COVID19-chatbot/
(10) **Docyet** https://corona.docyet.com/client/index.html
(11) **Dubai Department of Health** https://doh.gov.ae/COVID-19
(12) **e-Bot**[7] https://e-bot7.de/coronachatbot/
(13) **German Red Cross** WhatsApp +49(30)85404106
(14) **HealthBuddy** https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-COVID-19/healthbuddy
(15) **Infermedica** https://symptomate.com/COVID19/checkup/en/
(16) **Ivan Mask** t.me/ivanmaskbot
(17) **Martha** https://COVID19.app.keyreply.com/webchat/
(18) **MTI Singapore Chat for Biz** https://www.mti.gov.sg/Chatbot/chat
(19) **Providence** https://coronavirus.providence.org/
(20) **Russian Ministry of Health** WhatsApp +7(495)6240168
(21) **Suve** https://eebot.ee
(22) **Symptoma** https://www.symptoma.com/COVID-19
(23) **WHO** WhatsApp +41(79)8931892
(24) **Your.MD** https://webapp.your.md/login

Name	Channel	Content	Language	Interaction style
Ada	webchat	symptom checker	EN, DE	visual
Apple	webchat	symptom checker	EN	visual
Babylon	webchat	symptom checker	EN	visual
Bobbi	webchat	FAQ	EN, DE	content/ conversation
CDC	webchat	symptom checker	EN	visual
Cleveland Clinic	webchat	symptom checker	EN	visual
Corona Bot	FB Messenger	symptom checker, FAQ	FR	visual/ conversation
Covid-19 Chatbot	webchat	symptom checker	EN	visual
German Red Cross	Whatsapp	FAQ	DE	system/ conversation
Docyet	webchat	symptom checker, FAQ, mental support	DE	visual
Dubai Department of Health	webchat	FAQ	EN	content / conversation
e-Bot[7]	webchat	FAQ	DE	content
HealthBuddy	webchat	FAQ	EN, DE, FR, RU	conversation / content
HSE Corona Self-checker	webchat	symptom checker	EN	visual
Infermedica	webchat	symptom checker	EN, DE	visual
Ivan Mask	Telegram	FAQ	UK	visual/ conversation
Martha	webchat	symptom checker, FAQ	EN	content/ visual
MTI Singapore Chat for Biz	webchat	FAQ	EN	conversation/ visual
Providence	webchat	symptom checker	FR	visual
Russian Ministry of Health	Whatsapp	FAQ	RU	system/ content
Suve	webchat	symptom checker, FAQ	EN, ET	conversation/ visual
Symptoma	webchat	symptom checker	EN	visual
WHO	Whatsapp	FAQ	EN	system
Your.MD	webchat	symptom-checker, FAQ	EN	visual

Table 1: Chatbots for COVID-19 matters by channels, content, language and interaction style.

3.3 Expert Review Method

Two experts (one with a PhD degree in UX and one with a PhD degree in chatbots) scored each chatbot from Table 1 on all sub-heuristics (Sec. 3.1) as 0

- 'unsupported', 0, 5 - 'partially supported', and 1 - 'fully supported'. If a sub-heuristic did not apply to the particular chatbot in its particular context, the experts marked it with "n/a". The inter-rater agreement was substantial (Kappa Cohen 0.7245). For the final scoring, we picked the more optimistic value of both raters for non-agreement cases.

To establish a usability score for each chatbot, we first summed the values of the sub-heuristics for each heuristic and then divided the sum by the number of applicable items inside the heuristic. Secondly, we summed the scores for the twelve heuristics. An ideal chatbot would score 1 for each heuristic, hence reach a usability score of 12.

To get an impression which heuristics were overall well implemented compared to others, we summed the sub-heuristic and heuristic scores for all tested chatbots and divided them by the number of applicable items. Given that we looked at 24 chatbots, the highest possible sum per heuristic would have been 24 (=100%). We discuss the results per (sub-)heuristic in Section 4.2.

4 Results

4.1 COVID-19 Chatbots: Channels, Topics and Conversation Styles

19 of 24 tested COVID-19 chatbots work in webchat: they simulate a messenger-like interface on a website. Only five of 24 bots work in messengers: three in WhatsApp, one in Telegram and one in Facebook messenger. However, some messenger bots are available in multiple messengers. For instance, users can reach the WHO bot in WhatsApp and Viber, and Ivan Mask works in Telegram, Viber and Facebook Messenger. We excluded Viber versions from our benchmark but used them for a qualitative cross-channel comparison.

The choice of a particular channel influences interaction. While Viber and Telegram messengers offer similar interfaces, WhatsApp provides a different set of interactional resources. While Viber provides a standard set of widgets for messenger bots (i.e.; permanent menu, buttons, short replies), WhatsApp requires typing text messages. As a consequence, WhatsApp chatbots have to *simulate* a visual-centric interaction style by introducing number codes. To compare, Ivan Mask chatbot working in Telegram and Viber shows a very similar look-and-feel in both messengers.

In contrast to messengers, webchat allows more freedom in the implementation of the graphical user interface (GUI). Some webchat bots offer an attractive GUI (namely Apple, Infermedica, Symptoma, Docyet, Ada), as reflected by the high scores (cf Section 4.2) for heuristic 2 Match between system and the real world, 4 Consistency and standards, and 8 Aesthetic and minimalist design. However, webchat bots often only use a small part of the screen for the chat window, while the rest stays unused. Furthermore, the chat window cannot be moved or resized and the information is usually presented as text only.

Two most popular services in COVID-19 chatbots are symptom-checking and frequently asked questions (FAQ). As Table 1 shows, five chatbots offer

both services, and one of them also offers mental support. Although COVID-19 pandemic dramatically affected national and international businesses, we found only one chatbot that addresses business-related topics. FAQ bots frequently cover COVID-19 myths. Instead of a list of the topics, the WHO Viber bot offers a quiz asking the user to answer the bot's questions. Such a strategy helps increase user engagement and support learning [24].

Most of the bots offer visual-centric or content-centric conversation styles. The FAQ bots mostly offer a selection of topics, and the users can only choose among items from a list, without the possibility to type a question.

Chatbot	**Score**	1	2	3	4	5	6	7	8	9	10	11	12
IDEAL BOT	12	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Suve	**7.8**	0.5	**1.0**	0.15	**1.0**	0.8	0.5	0.8	**1.0**	0.6	0.6	0.3	0.5
Apple	**7.1**	0.7	**1.0**	0.3	**1.0**	**1.0**	0.5	0.0	**1.0**	0.8	0.8	-	-
Infermedica	**7.0**	**1.0**	**1.0**	0.6	**1.0**	**1.0**	0.7	0.0	**1.0**	-	0.8	-	-
Symptoma	**7.0**	**1.0**	0.7	0.6	0.8	0.8	0.8	0.5	**1.0**	0.10	0.8	-	-
HealthBuddy	6.6	0.8	**1.0**	0.1	0.8	0.0	0.5	0.5	0.8	0.9	0.5	0.3	0.5
Docyet	6.5	0.8	**1.0**	0.3	**1.0**	0.5	0.5	0.3	**1.0**	0.8	0.4	-	-
Ada	6.3	0.7	**1.0**	0.3	**1.0**	0.5	0.3	0.0	**1.0**	0.8	0.7	-	-
Germ. Red Cr.	5.9	0.5	**1.0**	0.4	0.5	0.0	0.5	0.8	0.8	0.3	0.6	0.3	0.3
Dub. Dpt. of H.	5.8	0.8	**1.0**	0.0	0.5	0.5	0.3	0.8	0.5	0.4	0.3	0.3	0.5
Corona Bot	5.4	0.5	0.8	0.0	0.8	0.8	0.0	0.3	**1.0**	0.4	0.1	0.3	0.5
Ivan Mask	5.3	0.8	0.8	0.3	0.8	0.3	0.5	0.2	0.5	0.5	0.3	0.3	0.3
Martha	5.3	0.7	**1.0**	0.0	0.8	0.5	0.2	0.2	0.8	0.1	0.5	0.2	0.5
WHO	5.3	0.8	**1.0**	0.4	0.5	0.0	0.8	0.7	0.8	0.1	0.4	0.0	0.0
Bobbi	5.3	0.8	0.7	0.1	0.3	0.3	0.2	0.5	0.8	0.6	0.6	0.2	0.4
MTI	5.1	0.5	0.5	0.1	0.5	0.5	0.2	0.8	**1.0**	0.5	0.4	0.0	0.1
Babylon	4.9	0.7	**1.0**	0.3	0.8	**1.0**	0.0	0.0	0.8	-	0.5	-	-
Covid-19	4.9	0.5	0.7	0.0	**1.0**	**1.0**	0.5	0.0	**1.0**	0.0	0.3	-	-
CDC	4.4	0.5	**1.0**	0.0	0.8	**1.0**	0.2	0.0	0.5	-	0.5	-	-
HSE	4.4	0.3	**1.0**	0.0	0.8	**1.0**	0.0	0.0	**1.0**	-	0.3	-	-
e-bot7	4.1	0.8	0.5	0.1	**1.0**	0.5	0.5	0.0	0.5	0.0	0.3	-	-
Providence	3.9	0.5	0.5	0.0	0.8	**1.0**	0.0	0.0	0.8	0.0	0.4	-	-
Your.MD	3.8	0.7	0.7	0.0	0.8	0.3	0.3	0.2	0.3	0.4	0.4	-	-
Clevel. Clinic	3.7	0.5	0.5	0.0	0.8	**1.0**	0.2	0.0	0.5	-	0.3	-	-
Russ. M. of H.	2.3	0.5	0.8	0.2	0.3	0.0	0.3	0.0	0.0	0.3	0.1	0.0	0.0
Score/heuristic		*64%*	*83%*	*17%*	*74%*	*58%*	*34%*	*28%*	*75%*	*40%*	*44%*	*21%*	*32%*

Table 2: Expert review scores for 24 COVID-19 chatbots on 12 heuristics: 1 Visibility of system status, 2 Match between the system and the real world, 3 User control and freedom, 4 Consistency and standards, 5 Error prevention, 6 Recognition rather than recall, 7 Flexibility and efficiency of use, 8 Aesthetic and minimalist design, 9 Help users recognise, diagnose, and recover from errors, 10 Help and documentation, 11 Context understanding, 12 Interaction management capabilities; on the scale 0 unsupported, 0,5 partially supported, 1 fully supported, - not applicable

Some bots also accept free text entry (e.g. German Red Cross and Ivan Mask) but very few bots are capable of information extraction from user utterances. The most disappointing experience appeared when a bot offered a text input line, but the function was disabled (e.g., e-Bot[7]).

Some bots offer both, buttons and free text input for interaction, but are in most cases not able to understand free text input. At least, the perceived experience for non-recognised inputs improves when the bot explains that it is still learning (e.g. Ivan Mask) and recommends using buttons, or when bots are capable of performing simple conversation management, such as recognition of openings and closings (e.g. Suve).

4.2 Results by Heuristic

Our second research question concerns the usability of COVID-19 chatbots. Table 2 presents the results of the expert review. The highest score in our sample is 7.8 out of 12. It was achieved by the Suve chatbot. The following top-scoring bots are the symptom checkers of Apple (7.1), Infermedica (7.0), and Symptoma (7.0). Infermedica and Symptoma were also the best two in the accuracy evaluation of COVID-19 symptom checkers [16].

Nearly all tested COVID-19 chatbots **scored well** on heuristics **1** Visibility of system status, **2** Match between system and the real world, **4** Consistency and standards, and **8** Aesthetic and minimalist design. **Ambivalent scores** were observed for the heuristics **5** Error prevention, **9** Help users recognise, diagnose, and recover from errors, and **10** Help and documentation. **Unsatisfying scores** were found for the remaining heuristics **3**. User control and freedom, **6** Recognition rather than recall, **7** Flexibility and efficiency of use, **11** Context understanding, and **12** Interaction management capabilities (cf. last row Tab. 2 for the percentage of the main heuristic).

The best scores were achieved by bots running in webchat and offering a combination of visual-centric interaction and natural conversation. Those bots also showed well-implemented heuristics 1, 2, 4, 6, 8 and 10. The details of each heuristic give a clearer picture of what specific functionalities (here represented by the sub-heuristics) require further design effort.

1. Visibility of system status was rather well implemented throughout all tested chatbots (64%). Sub-heuristic *(b) Immediate feedback* was close to entirely covered (98%), meaning that the user quickly knows that their input has been received and is treated. Nearly all chatbots showed efforts to *Compel user actions (c)*(56%). However, heuristic *(a) Information about the chatbots status in the process* was a weakness (20%) for all except two top-scoring bots.

2. Match between system and the real world has been very well implemented (83%). The good scoring comes from the high scores for sub-heuristic *(a) Chatbot uses the language familiar to the target user group* (90%), meaning that most chatbots employ easily understandable language. Many bots did neither employ visual components nor metaphors to enhance the communication with the users. For this reason, the two other sub-heuristics, namely *(b) Visual components of the messages (emojis, GIFs, icons) are linked to real-world objects*

and *(c) If metaphors are used, they are understandable for the user*, were not applicable for more than half of the tested bots.

3. User control and freedom scored very low throughout all tested chatbots (17%). None of the bots *understood repair initiations (d)* and only a few, among the best scoring bots, provided *navigation options (c)*, offered a *permanent menu (b)*, or *partially supported undo/redo of actions (a)*.

4. Consistency and standards scored well for most of the tested chatbots (74%). They *use the domain model from the user perspective (a)*, and have a *personality with a language is consistent throughout all interaction paths (b)*.

5. Error prevention shows an ambivalent scoring (58%). While the *prevention of unconscious slips by meaningful constraints (a)* is implemented to a basic degree (61%), only a few bots prevent these through *recognition of typos and spelling error correction (b)* (37%). The other two sub-heuristics for error prevention *(c) Chatbot requests confirmation before action with significant implications for the user* and *(d) Chatbot explains consequences of the user action* were not applicable for any of the 24 tested bots.

6. Recognition rather than recall is a usability principle that has not sufficiently found its way into the chatbots we tested (34%). About half make the options at least partly clear by adding *descriptive visual elements and clear instructions (a)* (52%). Few provide a *summary of the collected information before transactions (b)* (32%) which, however, in one-third of the tested chatbots was not even an applicable use case. None of the chatbots offered both a *permanent menu and help option (c)*, although about half had either one or the other.

7. Flexibility and efficiency of use was another low scoring heuristic (28%). Only about half of the sampled chatbots partly understands not only special instructions but also *natural synonym phrases (a)* (38%). One third can to some degree deal with *different formulations of the same intent (b)* (25%) and offer *multiple ways to achieve the same goal for more and less proficient users (c)* (21%). However, only six bots reached score 1 for at least one sub-heuristic here, leaving room for improvement.

8. Aesthetic and minimalist design is among the well-implemented heuristics (75%). Most chatbots reach scores of 1 or 0,5 for their *use of visual information in a personality-consistent manner (b)* (87%), as well as for *concise and precise dialogues (a)* (65%).

9. Help users recognise, diagnose, and recover from errors is a heuristic that was not sufficiently established in our chatbot sample (40%). None of the bots scores 1 for clearly *indicating that an error has occurred (a)* (33%). Only some *use plain language to explain the error (b)* (56%) or *explain the actions needed for recovery (c)* (45%). Even less offer a *shortcut to quickly fix the error (d)* (22%).

10. Help and documentation shows an ambivalent scoring (44%). While nearly all chatbots provide a *clear description of their capabilities (a)* (85%), very few offer *keyword search (b)* (34%). Only the high-ranking bots *focus their help on the user task (c)* (26%) and *explain concrete steps to be carried out for a task (d)* (25%).

11. Context understanding was a non-applicable heuristic for half of the tested chatbots and not very well implemented in the applicable cases (44%). When applicable, most *understood the context within one turn (a)* to some extent (55%). However, almost none *understood the context within a small number of turns (b)* (9%), let alone the context of a *multi-turn conversation (c)* (0%).

12. Interaction management capabilities too was a non-applicable heuristic for about half of the tested chatbots and scored low in the applicable cases (32%). If applicable, most of the bots understood *conversation openings and closings (a)* (68%) as well as *sequence closings (b)* (55%) to some extent. However, they neither *understood repair initiations or replied with repairs (c)* (0%), nor did they *initiate repairs to handle potential user errors (d)* (5%).

4.3 Non-applicable Heuristics

Only 11 of 24 chatbots implemented at least one element per heuristic. 13 of 24 chatbots did not implement any conversational functionality, and therefore, heuristics 11 and 12 were not applicable for them. Five of the chatbots are strictly visual-centred (interaction only via buttons), so that heuristic 9 was not applicable, either. We furthermore find sub-heuristics that have not been applicable for any of the 24 chatbots in our sample. Among them are *(5c) Error prevention - Chatbot requests confirmation before action with significant implications for the user* and (5d) *Error prevention - Chatbot explains the consequences of the user action in chat*. However, the experts did not encounter any situation that would have required these features.

5 Discussion and Limitations

The experience a user lives with a product or service materialises from the interplay of various dimensions [1]. Despite the widespread opinion that natural conversation with chatbots is the ultimate goal in chatbot research, this review shows that the right balance of interaction flexibility and pace can be achieved by merging natural conversation with visual-centric interaction. In this way, a satisfying conversational UX can be ensured for users who value efficiency [2].

Channels, content, and interaction style show mutual dependencies. Messengers such as WhatsApp, are potent communication channels because of their extensive number of users. However, the dominance of webchat channels can be explained by two aspects: 1) security and data protection considerations; 2) greater freedom for design - webchats offer more possibilities to personalise the design and to add visual elements as compared to bots running in messengers. The fact that the highest scores in this study have been earned by webchat bots does not mean that webchat as a channel is per default the best one.

The WhatsApp API does not provide any visual elements, and therefore, it is better suitable for conversation-centric style. The three WhatsApp chatbots (WHO, German Red Cross (GRC) and Russian Ministry of Health (RMoH))

in our sample chose different message-based interactions that *simulate* visual-centric style. RMoH chatbot understands only number codes (scored 2.3). The WHO chatbot understands number codes and keywords presented in bold in the bot messages (scored 5.3). The GRC chatbot understands the first two variants plus it can extract keywords from natural phrases (scored 5.9).

Intent-based natural language understanding (NLU) is the state of the art in current chatbot building platforms (e.g. Watson, DialogFlow and RASA). Surprisingly, less than half of the chatbots in this study made use of NLU methods. Although almost none of the symptom checkers implemented conversation management or context understanding, this is not necessarily negative for symptom-checkers that simulate a form-filling interaction (Apple, Symptomate, Ada).

Indeed, the advantages of using a chatbot for the sake of informing people about COVID-19 are in many cases unclear. Both FAQ and form-filling (symptom checking) tasks can be presented more user-friendly on a "traditional"website. Building a chatbot just for the sake of having a chatbot may harm the service because of the less optimal UX.

Expert reviews based on usability heuristics are only one among the various tools of UX evaluation [11]. An expert usability review usually analyses only one service/product in-depth and explicitly lists all identified issues - including screenshots, description, and proposed solution. This study sought to give an overview of usability problems in Covid-19 chatbots in general. The heuristics were therefore only used to establish a usability score for each bot. The scoring for each heuristic highlights design rules that are not sufficiently taken into account, and serve to trace specific usability issues. Observing real users during their interaction with the bot will reveal the most critical shortcomings of the system, as well as provide an impression of the user satisfaction with the interaction - insights a heuristic review cannot produce.

Finally, this heuristic review, unfortunately, did not yield conclusions specific to channels, topics, and interaction styles because the different types were not evenly represented in our sample - mostly webchat, mainly symptom checkers, principally visual-centric conversation style.

6 Conclusions and Recommendations

This study shows that our conversational UX evaluation framework is applicable to chatbots of different conversation styles [14]. We can conclude that natural conversations with chatbots are in general not mandatory for good conversational UX. Because the analysed COVID-19 chatbots show a large redundancy in topics and types of service, but are diverse in UX scores, we conclude that conversational E-health applications would be more attractive to users if they invest in UX from the beginning. The following concrete steps need to be taken in order to make pragmatically motivated use of chatbots [2] also satisfying in terms of UX:

1. Before starting, think of having a conversation with a real person in that chat. Will the chat format be efficient and effective to solve the user problem? If yes, start with the chatbot. If not, choose another channel.
2. Does the bot have to share large pieces of text or even documents with the user? In this case, a chat window might not be the right place. Break down the large text pieces or reconsider whether the chatbot is the right way of communication.
3. Implement basic conversation management capabilities. Many chatbot building platforms call it "small talk"and offer ready-to-use conversational components for it.
4. Think how to implement repair functionalities and shortcuts in order to increase the bot's usability. In conversation-centric interaction it should be close to the repair system [14]. In visual-centric interaction, other interactional resources must be chosen.
5. If the chatbot channel supports visual-centric interaction, think carefully, where visual elements can improve the UX, and where text input is more effective. If the bot accepts free text, be prepared that people will use it.

Further research questions arose from this study: What sorts of chatbots would be really helpful in the context of pandemics, going beyond accuracy and UX? Which conversational e-health applications offer real added value to their users? Which topics and services beyond FAQ and symptom checkers can be explored within in the context of the COVID-19 crisis?

References

1. Bongard-Blanchy, K., Bouchard, C.: Dimensions of user experience-from the product design perspective (2014)
2. Brandtzaeg, P.B., Følstad, A.: Why people use chatbots. In: International Conference on Internet Science. pp. 377–392. Springer (2017)
3. Brandtzaeg, P.B., Følstad, A.: Chatbots: Changing user needs and motivations. Interactions **25**(5), 38–43 (2018). https://doi.org/10.1145/3236669
4. Budiu, R.: The user experience of chatbots. Nielsen Norman Group logoNielsen Norman Group (November 2018)
5. Chaix, B., Delamon, G., Guillemasse, A., Brouard, B., Bibault, J.E.: Psychological distress during the covid-19 pandemic in france: a national assessment of at-risk populations. medRxiv (2020)
6. Danilava, S., Busemann, S., Schommer, C., Ziegler, G.: Towards Computational Models for a Long-term Interaction with an Artificial Conversational Companion. In: Proc. of ICAART'13 (2013)
7. Fadhil, A., Schiavo, G.: Designing for health chatbots. arXiv preprint arXiv:1902.09022 (2019)
8. Höhn, S.: Artificial Companion for Second Language Conversation. Springer (2019)
9. Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., McTear, M.: Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In: Proceedings of ECCE. pp. 207–214 (2019)
10. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (sassi). Natural Language Engineering **6**, 287–303 (2000)

11. Lallemand, C., Gronier, G.: Méthodes de design UX: 30 méthodes fondamentales pour concevoir des expériences optimales. Eyrolles (2018)
12. Luger, E., Sellen, A.: "Like Having a Really Bad PA": The gulf between user expectation and experience of conversational agents. In: Proceedings CHI'16. p. 5286–5297. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858288
13. McTear, M., Callejas, Z., Griol, D.: The Conversational Interface: Talking to Smart Devices. Springer (2016)
14. Moore, R.J., Arar, R.: Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework. ACM Books (2019)
15. MTI-Singapore: Ministry of trade and industry singapore, MTI Chat for Biz. https://www.mti.gov.sg/Chatbot/chat (2020)
16. Munsch, N., Martin, A., Gruarin, S., Nateqi, J., Abdarahmane, I., Weingartner-Ortner, R., Knapp, B.: A benchmark of online COVID-19 symptom checkers. medRxiv (2020)
17. Nielsen, J.: Ten usability heuristics. http://www. nngroup. com/articles/ten-usability-heuristics/(acc-essessed 25.06.2020) (2005)
18. Semigran, H.L., Linder, J.A., Gidengil, C., Mehrotra, A.: Evaluation of symptom checkers for self diagnosis and triage: audit study. bmj **351**, h3480 (2015)
19. Shariat, J., Saucier, C.S.: Tragic Design: The Impact of Bad Product Design and How to Fix It. O'Reilly (2017)
20. Shevat, A.: Designing Bots: Creating Conversational Experiences. O'Reilly (2017)
21. Sugisaki, K., Bleiker, A.: Usability guidelines and evaluation criteria for conversational user interfaces: a heuristic and linguistic approach. In: Proceedings of the Conference on Mensch und Computer. pp. 309–319 (2020)
22. TARS: Chatbot templates to fight coronavirus (covid-19) pandemic. https://hellotars.com/chatbot-templates/coronavirus-covid19-fight/ (06 2020)
23. Verma, V.: 10 usability heuristics every chatbot company should follow. UX Collective, Medium (2019)
24. Vijayakumar, B., Höhn, S., Schommer, C.: Quizbot: Exploring formative feedback with conversational interfaces. In: International Conference on Technology Enhanced Assessment. pp. 102–120. Springer (2018)
25. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: Evaluating spoken dialogue agents with PARADISE: Two case studies. Comp. speech & lang. **12**(4) (1998)