Usability and User Experience, Vol. 39, 2022, 233–241
https://doi.org/10.54941/ahfe1001712

Heuristic Evaluation of Public Service Chatbots

Marleen Vanhauer and Stephan Raimer

University of Applied Sciences for Administration and Services Altenholz, 24161, Germany

ABSTRACT

In recent years, chatbots have been adopted in business contexts and also for public services at a growing rate. Chatbots provide dialogue interfaces combining visual elements with natural conversation. Good Conversational Design in this context covers the topics of Natural-Language Processing (NLP) and Dialogue Management (DM). Few attention has been paid to the usability evaluation of conversational interfaces (Höhn & Bongard-Blanchy, 2021). The present paper builds upon the work by Höhn & Bongard-Blanchy by applying their framework of conversational heuristics to evaluate a set of public service chatbots operated in the federal state of Schleswig-Holstein. Thus, for each public service chatbot, a usability score is established and typical characteristics of public service chatbots in general are summarized. We discuss results by comparing the overall scores, weaknesses and strengths of each chatbot. In addition, we reflect on our experience in the application of the framework as well as highlight possible optimization potentials. Concludingly, this paper collects UX recommendations for public service chatbots.

Keywords: Heuristic evaluation, Public service chatbots, Conversational UX, Usability evaluation of chatbots, UX recommendations for public service chatbots

INTRODUCTION

Chatbots have been adopted in business contexts and also for public services at a growing rate. As a new way of interaction, chatbots allow 24/7 availability while being scalable to large numbers of users. They can be used to support public administration services, for example to provide specific information in advance. Due to the different responsibilities of the administration, at least three levels must always be considered in the federal structure in Germany (federal, state and municipal level). Following this, a separate technical infrastructure for chatbots was provided for the German federal state of Schleswig-Holstein. Different parts of the administration thus find a basis to quickly implement their use cases (i.e. COVID-19 chatbots and other examples). Generally, chatbots provide dialogue interfaces combining visual elements with natural conversation. Good Conversational Design in this context does not only require dialogues, but also logical and coherent dialogue structures and an iteratively optimized user experience that takes into account the respective context of the user. Essentially, good conversational design covers the topics of Natural Language Processing (NLP, i.e. correctly interpreting user intents) and Dialogue Management (DM, i.e. providing the appropriate content and responses) (Budiu, 2018). While conversational capabilities of chatbots (especially NLP) were improved, fewer attention was paid to the evaluation of the user experience and usability of chatbots (Höhn & Bongard-Blanchy, 2021). Taken this into account, the study underlying this paper aims to answer the following research questions and sub-questions:

Q1: What public service chatbots exist in the federal state of Schleswig-Holstein?

Q2: What are certain characteristics of public service chatbots?

- Which functional classifications for chatbots of public services are there?
- Do chatbots support English or simple language?
- To what extent is reference made to the privacy policy (GDPR)?

Q3: How usable are public service chatbots?

Our focus is to heuristically review existing chatbots for public services in the German federal state of Schleswig-Holstein. Therefore, our approach is to apply the generic evaluation framework by Höhn & Bongard-Blanchy (2021) with 12 heuristics that are adapted to the conversational interface context. Thus, a usability score for each chatbot example is established. In addition, we discuss our experience in the application of the framework as well as highlight possible optimization potentials. Concludingly, this paper collects UX recommendations for public service chatbots.

RELATED WORK

As a basis for our work we have evaluated literature dealing with chatbot evaluation methods or (public service) chatbot evaluations. In addition, we give an overview of which best-practices, recommendations and guidelines for chatbots are there.

In 2018, Budiu and the Nielsen Norman Group performed an *informal tasked-based usability study* with 8 participants to evaluate a set of customer service and messenger chatbots. Among others, they criticized the lack of ability of most chatbots to react to unexpected user inputs deviating from the chatbot's flow. They turned their findings into *UX Guidelines for Designing Chatbots*. Accordingly, a state-of-the-art chatbot should be able to perceive natural-language queries, as well as understand and process these messages intelligently. Maroengsit et al. (2019) reviewed thirty chatbot research articles on chatbot evaluation methods and classified methods into 3 categories: content evaluation methods (e.g. *automated evaluation* or *expert evaluation*), user satisfaction evaluation methods and functional evaluation methods. According to Maroengsit et al., the most common method is the *evaluation of user satisfactio*n. In their study, Ren et al. (2019) searched databases of scientific publications in regard to usability evaluation methods. They state, that common HCI methods for usability testing are being adopted to chatbots, often combining two or more methods, with *questionnaires* and *usability interviews* being the most commonly used ones. Besides *efficiency* and *effectiveness*, they identified *satisfaction* as the usability characteristic being evaluated the most. Moreover, they list corresponding measures (e.g. *accuracy of chatbot reply, expert assessment, ease-of-use* or *user experience*) which determine the previously named characteristics. They also state, the methods *SUS survey* and *follow-up interview* combined are being applied more frequently for chatbot evaluations. Höhn & Bongard-Blanchy (2021) performed a heuristic evaluation on COVID-19 chatbots. They adapted the 10 heuristics by Nielsen (1994a) and proposed a framework consisting of 12 heuristics and 39 sub-heuristics to apply to conversational chatbots applications. Their scoring shew how much heuristics were supported by (health) chatbots.

For designing user interfaces and interactions in general, probably the most cited principles are Nielsen's 10 Usability heuristics (Nielsen, 1994a). These cover the most universal aspects relevant for designing nearly any kind of interface and user interaction, and in turn, as previously mentioned, can serve as a tool for a heuristic evaluation. Klopfenstein et al. (2017) examined chatbots running on popular messaging platforms which they introduced as 'Botplications'. Botplications as a novel form of conversational chatbots follow "principles of simplicity and effectiveness" as opposed to mobile and web (chat) apps. They describe the following features as best-practices: (1) Thread as app, (2) History awareness, (3) Enhanced UI, (4) Limited use of Natural Language Processing (NLP), (5) Message self-consistency and (6) Guided conversation. Amershi et al. (2019) collected and studied HCI guidelines for over 20 years and in their paper introduced a validated set of 18 guidelines applicable for practitioners when creating AI interfaces. They consider four timely stages of interaction when to apply guidelines to AI-systems: (1) Initially, (2) During interaction, (3) When wrong and (4) Over time. Due to always rapidly changing technologies and according to their opinion, the most profound contents can be found within industry resources.

In the context of corporate and industry publications, we refer to the Ethical Guidelines for Trustworthy AI by the European Commission (2019), the blogposts by the regional chatbot development company assono GmbH (2021) as well as to the Conversation Design Guidelines by Google Developers (2021), the Responsible AI principles by Microsoft (2022) and Microsoft's the Guidelines for Human-AI Interaction (2021). The European Commission within their guidelines published a list of 7 requirements to guide the development of trustworthy AI applications, which are: (1) Priority of human action and supervision, (2) Technical robustness and security, (3) Data privacy and data quality management, (4) Transparency, (5) Diversity, non-discrimination and fairness, (6) Social and environmental well-being, and (7) Accountability. Assono GmbH as one of the developers of chatbots in our case study gives recommendations on error handling, designing natural and sympathetic conversations as well as the usage of an avatar, small talk and emojis for chatbots. Google's Guidelines on Conversation Design aim at practitioners who define interactions for *intelligent voice assistants*, especially the Google Assistant. To some extent, these guidelines may also be helpful when composing dialogues for chatbots by introducing *building blocks*, i.e. action patterns, and giving overviews of possible *conversational*

Table 1. The 12 heuristics for analysis of conversational interfaces (Höhn & Bongard-Blanchy 2021).

Heuristic	Sub-Heuristics	
1 Visibility of system status	1.1	Presence of information about the chatbot's state in the entire process
	1.2	Immediate feedback (did the last user action work?)
	1.2	Compel user action (what does the chatbot think the user will do next?)
2 Match between system and the real world	2.1	Chatbot uses the language familiar to the target users
	2.2	Visual components (emojis, GIFs, icons) are linked to real-world objects
	2.3	If metaphors are used, they are understandable for the user
3 User control and freedom	3.1	Chatbot supports undo/redo of actions
	3.2	Chatbot offers a permanent menu
	3.3	Chatbot provides navigation options
	3.4	Chatbot understands repair initiations
4 Consistency and standards	4.1	Chatbot uses the domain model from the user perspective
	4.2	Chatbot has a personality, consistency in language and style
5 Error prevention	5.1	Chatbot prevents unconscious slips by meaningful constraints
	5.2	Chatbot prevents unconscious slips by spelling error detection
	5.3	Chatbot requests confirmation before actions with significant implications
	5.4	Chatbot explains consequences of the user actions
6 Recognition rather than recall	6.1	Chatbot makes the options clear through descriptive visual elements and explicit instructions
	6.2	Chatbot shows summary of the collected information before transactions
	6.3	Chatbot offers a permanent menu and help option
7 Flexibility and efficiency of use	7.1	Chatbot understands not only special instructions but also synonyms
	7.2	Chatbot can deal with different formulations
	7.3	Chatbot offers multiple ways to achieve the same goal
8 Aesthetic and minimalist design	8.1	Chatbot dialogues are concise, only contain relevant information
	8.2	Chatbot uses visual information in a personality-consistent manner to support the user, not just random decoration
9 Help users recognize, diagnose, and recover from errors	9.1	Chatbot clearly indicates that an error has occurred
	9.2	Chatbot uses plain language to explain the error
	9.3	Chatbot explains the actions needed for recovery
	9.4	Chatbot offers shortcuts to fix errors quickly
10 Help and documentation	10.1	Chatbot provides a clear description of its capabilities
	10.2	Chatbot offers keyword search
	10.3	Chatbot focuses its help on the user task
	10.4	Chatbot explains concrete steps to be carried out for a task
11 Context understanding	11.1	Chatbot understands the context within one turn
	11.2	Chatbot understands the context within a small number of turns (usually 2-3 user-bot turn pairs)
	11.3	Chatbot understands the context of a multi-turn conversation
12 Interaction management capabilities	12.1	Chatbot understands conversation openings and closings (e.g., 'hello')
	12.2	Chatbot understands sequence closings (e.g., 'ok' and 'thank you')
	12.3	Chatbot understands repair initiations and replies with repairs
	12.4	Chatbot initiates repair to handle potential user errors

Table 2. Evaluated public service chatbots, developed by [1] Dataport AöR, [2] Govii UG and [3] assono GmbH.

Chatbot	Federal state (Municipality)	**Purpose of use**	**URL**
Cabo[1]	SH	FAQs on COVID19 for citizens and companies of Schleswig-Holstein	https://www.schleswig-holstein.de/DE/Schwerpunkte/Coronavirus/coronavirus_node.html
Govii[2]	SH (Kiel)	Information on public services	https://www.kiel.de/de/politik_verwaltung/service/
Ina[1]	SH	Information on services of the integration office for citizens and employers	https://ina.schleswig-holstein.de/
Nordi[3]	SH (Norderstedt)	Information on public services	https://www.norderstedt.de/
RECKi[1]	SH (Rendsburg-Eckernförde)	Information service on vehicle registrations for citizens and companies	https://chatbot.kreis-rendsburg-eckernfoerde.de/chat/

components (e.g. acknowledgements or questions) as well as *visual components* (e.g. cards, carousels or lists). Microsoft's AI principles reflect a set of 6 rules for developing responsible AI products, namely: (1) Fairness, (2) Reliability & safety, (3) Privacy & security, (4) Inclusiveness, (5) Transparency and (6) Accountability – which approach is similar to the one by the European Commission. Whereas, Microsoft's (2021) Guidelines for Human-AI Interaction honour the 18 guidelines originally published by Amershi et al. in 2019.

HEURISTIC EVALUATION AND CASE STUDY

The heuristic evaluation is a method, originally invented by Nielsen (1994b), to review a prototype or product by several internal or external experts. Hence, it is sometimes also referred to as expert review. This sets them apart from other evaluation methods such as user interviews or user analytics which in contrast involve actual users. Thus, a heuristic evaluation does not reflect the behaviour or attitudes of users. Instead, predefined heuristics, design rules, principles or guidelines serve experts as a basis for evaluation. Nevertheless, it is a well-established method to identify usability issues and to derive novel principles and recommendations.

The present heuristic evaluation was undertaken by two evaluators having backgrounds in Public IT Services, UX Research and Design. For each chatbot product, the authors noted down issues per sub-heuristic, rated each sub-heuristic and derived recommendations. Subsequently, listed are the 12 heuristics for evaluation of conversational UX defined by Höhn & Bongard-Blanchy (2021) which we used to analyze our case study.

Our case study consists of five chatbots developed for the federal state of Schleswig-Holstein (SH) and associated municipalities serving different purposes and use cases.

RESULTS

To identify certain characteristics of public service chatbots, the authors researched public sources and took notes during evaluation.

Which Functional Classifications for Chatbots of Public Services are There?

According to Etscheid et al. (2020) by Fraunhofer IAO, public service chatbots can be classified into three categories, which at the same time reflect their grade of maturity: (1) Providing information services, from answering simple FAQs through keyword search up to handling complex, multi-turn conversations (e.g. informing about services, finding a contact), (2) Connecting to existing processes (e.g. making an appointment), and (3) Integration into existing specialist procedures (e.g. filing applications, handling of complete processes). All chatbots provide information on public services, whereas Cabo presented the least functional maturity as a solely FAQ chatbot (category 1). All others are able to at least some extent handle multi-turn conversations, and connect to existing services (category 2), e.g. an online-appointment-service or the public service web portal. Ina even provides an address input form for ordering brochures or call back-service (category 3).

Do Chatbots Support English or Simple Language?

Serving citizens equal of their nationality public service chatbots should at least provide English as internationally spoken language or a simplified version of the originate language to also address laypersons which most often do not understand legal formulations. At present, none of the chatbots supports English as alternative language to German. Although, Nordi reacts to the English greeting „Hello" with the response that English is not yet available, but planned to be implemented in the future. Of particular note is the chatbot Ina which even offers users to switch to simple language.

To what Extent is Reference made to the Privacy Policy (GDPR)?

A note on the privacy policy is made in different ways. Inside the chatbot application it is only addressed by the chatbots Govii and Ina. Cabo asks for consent outside the chatbot on the corresponding web platform, as well as Nordi which links outside of the chatbot when asking for the privacy policy.

To assess the usability of the public service chatbots, we applied the heuristics and scoring scale by Höhn & Bongard-Blanchy (2021), awarding scores of either 0.0, 0.5 or 1.0 for unsupported, partially supported and fully supported sub-heuristics. The score for each heuristic results from the average score of sub-heuristics by the authors. The total score for each chatbot results from summing up all 12 scores. In addition, usability issues for each sub-heuristic

Table 3. Rated scores per chatbot and heuristic (rounded to one digit after comma).

Public Service Chatbot	Heuristic/ Total score	1	2	3	4	5	6	7	8	9	10	11	12
Nordi	**10.5**	0.8	**1.0**	0.9	0.9	0.6	**1.0**	0.8	**1.0**	0.9	0.9	0.8	0.8
Govii	**10.4**	0.7	**1.0**	0.8	**1.0**	0.7	0.9	0.9	0.6	0.9	**1.0**	0.9	0.9
RECKi	**9.6**	0.8	**1.0**	0.6	**1.0**	0.7	0.5	**1.0**	0.9	0.6	0.8	0.8	0.9
Ina	**7.8**	0.7	0.8	0.5	0.8	0.4	0.5	0.8	0.9	0.7	0.8	0.6	0.5
Cabo	**5.8**	0.7	0.7	0.4	0.8	0.6	0.4	0.3	0.5	0.6	0.4	0.2	0.4

were documented qualitatively based on the UX heuristic evaluation template by Moyes (2022).

DISCUSSION

The rating shows the chatbot Nordi reaching the highest and the chatbot Cabo the lowest overall score. Conversation with Nordi feels most natural due to the fact that Nordi is able to understand and reply to multi-turn conversations by giving a full range of answers, closely followed by Govii. The chatbot Ina stands out by offering its service in simple language. The chatbot RECKi shows a serious problem, due to an expired SSL certificate which most likely leads users not to use the service at all. Cabo scored the worst, not being able to understand or give context-based answers. Also, its keyword-based search delivers way too much results making it hard to find the desired information.

Regarding the practicability and efficiency of the heuristic evaluation, almost all sub-heuristics and heuristics were applicable, except sub-heuristic 6.2 which can only be applied to chatbots collecting personal information (e.g. Ina). Overall, we think that the proposed 39 sub-heuristics for chatbots by Höhn & Bongard-Blanchy (2020) could all fit into the main 10 heuristics by Nielsen (1994a) in order to keep the heuristic framework lean. The sub-heuristics of heuristic 11 (Context understanding) concern the efficiency, which we would therefore deploy to heuristic 7. The sub-heuristics 12.1 and 12.2 reflect communication which we would deploy to heuristic 4 (Consistency and standards), and the sub-heuristics 12.3 and 12.4 concerning a chatbot's intention to repair errors to heuristic 5 (Error prevention). Moreover, we suggest to assign the sub-heuristic 3.4 to heuristic 5 (Error prevention) and the sub-heuristic 6.3 to heuristic 10 (Help and documentation). To the end, heuristic 7 (Flexibility and efficiency of use) could be extended by effectiveness (e.g. accuracy of search results).

Altogether, we found that the heuristic evaluation is a viable method to address usability issues, including those of public service chatbots. The qualitative assessment provides weighting scores which allow to focus on the most important issues. However, it never claims to be complete and depends on the unbias of the evaluators. A successful conduction requires sufficient preparation time to immerse into the abstract heuristics, to use the chatbots and apply the framework to each specific chatbot application.

CONCLUSION

Based on our case study, we derived a set of UX recommendations and principles which complement existing chatbot guidelines and can be applied to public service chatbots, especially to the presented chatbot cases.

- For keyword search, optimize and limit search results to a reasonable number of results.
- Avoid to implement only keyword-based search functionality or linking to external web contents.
- Do not only offer restricted answers and find a balance between open and closed questions.
- Train the chatbot to understand multi-turn dialogues with different dialogue flows (e.g. informing about services, finding a contact), to broaden the context and deliver precise information.
- Connect the chatbot to existing processes (e.g. making an appointment, call-back).
- Integrate existing services into the chatbot (e.g. filing applications, applying for benefits).
- Implement simple language instead of using extended official language (also within FAQs).
- Consider to offer at least one alternative language.
- Ask for consent for privacy policy regulations inside the chatbot application.
- Integrate live search while the user is typing and spell-checking.
- Offer a permanent menu with access to help (e.g. call-back or e-mail).
- Allow to cut off a conversation thread (‘Goodbye’) anytime. Offer to restart a conversation.
- Ask for feedback at the end of conversation thread.

For future work, besides evaluating against abstract heuristics, another approach would be to evaluate against best-practises and recommendations proposed by industry resources. We also want to focus on to what extent a modified rating scale (5 or 7 point) can lead to more actionable results in combination with qualitative data.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.*

Assono GmbH (2021). Error-Handling im Chatbot: So halten Sie den Dialog am Laufen [online]. *Assono GmbH.* Available from: https://www.assono.de/blog/so-halten-sie-den-chatbot-dialog-bei-unbekannten-fragen-am-laufen [Accessed Jan 2022].

Budiu, R. (2018). The User Experience of Chatbots [online]. *Nielsen Norman Group.* Available from: https://www.nngroup.com/articles/chatbots/ [Accessed Jan 2022].

Etscheid, J., Stroh, F., and von Lucke, J. (2020). Künstliche Intelligenz in der öffentlichen Verwaltung [online]. *Fraunhofer IAO.* Available from: https://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-5070158.pdf [Accessed Jan 2022].

European Commission (2019). Ethik-Leitlinien für eine vertrauenswürdige KI. Publications Office. https://data.europa.eu/doi/10.2759/856513 [Accessed Jan 2022].

Google (2021). Conversation Design [online]. *Conversation Design I Google Developers*. Available from: https://developers.google.com/assistant/conversation-design/welcome [Accessed Jan 2022].

Höhn, S. and Bongard-Blanchy, K. (2021). Heuristic Evaluation of Covid-19 Chatbots. *Chatbot Research and Design*, 131–144.

Klopfenstein, L.C., Delpriori, S., Malatini, S., and Bogliolo, A. (2017). The Rise of Bots. *Proceedings of the 2017 Conference on Designing Interactive Systems*.

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., and Theeramunkong, T. (2019). A Survey on Evaluation Methods for Chatbots. *Proceedings of the 2019 7th International Conference on Information and Education Technology - ICIET 2019*.

Microsoft (2021). Guidelines for Human-AI Interaction [online]. *Guidelines Overview - Microsoft HAX Toolkit*. Available from: https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/ [Accessed Jan 2022].

Microsoft (2022). Responsible AI principles from Microsoft [online]. *Responsible AI principles from Microsoft*. Available from: https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3Aprimaryr6 [Accessed Jan 2022].

Moyes, M. (2022). UX Heuristic Evaluation Template (Community) - Figma [online]. *UX Heuristic Evaluation Template*. Available from: https://www.figma.com/file/ffRk8zPbGQiY3NrmuyRDjO/UX-Heuristic-Evaluation-Template-(Community)?node-id=419%3A316 [Accessed Jan 2022].

Nielsen, J. (1994a). 10 Usability Heuristics for User Interface Design [online]. *10 Usability Heuristics for User Interface Design*. Available from: https://www.nngroup.com/articles/ten-usability-heuristics/ [Accessed Jan 2022].

Nielsen, J. (1994b). Heuristic Evaluation: How-To: Article by Jakob Nielsen [online]. *Heuristic Evaluation: How-To: Article by Jakob Nielsen*. Available from: https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/ [Accessed Jan 2022].

Ren, R., Castro, J.W., Acuña, S.T., and de Lara, J. (2019). Evaluation Techniques for Chatbot Usability: A Systematic Mapping Study. *International Journal of Software Engineering and Knowledge Engineering*, 29 (11n12), 1673–1702.