Heuristic Evaluation of Conversational Agents

Raina Langevin
Human Centered Design and Engineering,
University of Washington
rlangevi@uw.edu

Ross Lordon
Microsoft
rolordon@microsoft.com

Thi Avrahami
Rulai
thi@rul.ai

Benjamin Cowan
School of Information and Communication
Studies, University College Dublin
benjamin.cowan@ucd.ie

Tad Hirsch
Department of Art + Design, Northeastern
University
tad.hirsch@northeastern.edu

Gary Hsieh
Human Centered Design and Engineering,
University of Washington
garyhs@uw.edu

ABSTRACT

Conversational interfaces have risen in popularity as businesses and users adopt a range of conversational agents, including chatbots and voice assistants. Although guidelines have been proposed, there is not yet an established set of usability heuristics to guide and evaluate conversational agent design. In this workshop paper, we introduce a set of heuristics for conversational agents adapted from Nielsen's heuristics and based on expert feedback. We validate the heuristics through two rounds of evaluations conducted by participants on two conversational agents, one chatbot and one voice-based personal assistant. When using the conversational agent heuristics to evaluate both interfaces, evaluators were able to identify more usability issues than when using Nielsen's heuristics. We propose that the heuristics successfully identify issues related to dialogue content, interaction design, help and guidance, human-like characteristics, and data privacy.

CCS CONCEPTS

• **Human-centered computing** → **Heuristic evaluations**; **User interface design**.

KEYWORDS

heuristic evaluation, conversational agents, user interface design

INTRODUCTION

Conversational agents are growing in popularity, through the uptake of text based and voice based conversational systems such as chatbots and Intelligent Personal Assistants (IPAs) respectively. Unlike other forms of human-computer interfaces, there is little consensus as to best practice for the design of conversational agents [2]. Recently there have been strides towards consolidating and validating guidance in related areas, such as human-AI interaction [1], and human-like chatbot experiences [10]. Our work looks to build upon recent efforts towards heuristics for specific modalities, like voice interactions [7][11], to develop a comprehensive set of heuristics for conversational agent based interactions.

In this paper, we use Nielsen's heuristics [8] as a foundation upon which to build, adapting these for conversational agent based interaction. We contribute a set of validated heuristics that researchers and practitioners may use in their formative evaluation of conversational agents. By demonstrating their effectiveness in real world system evaluations, we propose that our heuristics can be applied to text and voice-based conversational agents. More broadly, our work contributes to existing research on heuristic evaluation and further highlights how this technique may be adapted for new and future interfaces.

DESIGN PROCESS

We sought to expand on Nielsen's heuristics using a four phased design process and utilized a similar design process used in prior work to develop heuristics for ambient displays [6]. We first developed a set of heuristics for the design of conversational agent interfaces using prior research findings as well as our own experiences in developing these interfaces [5] [3]. Second, we presented these heuristics to 9 experts in conversational agent design and heuristic evaluation, and incorporated their feedback. In the third phase, participants evaluated our heuristics on two interfaces, a voice assistant on the Amazon Echo and an online chatbot. We compared our heuristics with Nielsen's heuristics to observe their effectiveness in identifying usability issues with conversational agents. After finding that the conversational agent heuristics performed well on the voice interface, but not the chatbot interface, we further iterated on the heuristics. Finally, in the fourth phase, we validated our heuristics on the chatbot interface by comparing them to Nielsen's heuristics. We invited freelance professionals in user interface design on Upwork to conduct online heuristic evaluations. From this, we determined that the conversational agent heuristics performed more effectively than Nielsen's heuristics.

		Phase 3		Phase 4	
Participant set	Experts	CA	N	CA	N
voice	9	30	23	–	–
chatbot-all	31	33	34	35	28
chatbot-bal	31	22	24	35	28

Table 1: Number of usability issues found by the experts, conversational agent (CA) and Nielsen (N) groups in Phase 3 and 4.

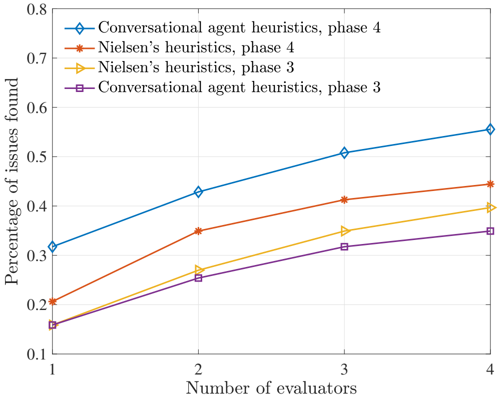

Figure 1: Percentage of issues for the chatbot found by the top four evaluators using the conversational agent heuristics and Nielsen's heuristics in Phase 3 and 4.

RESULTS

Two of the authors iterated on the master list of usability issues for the chatbot from Phase 3 and merged in issues from Phase 4. Since Phase 4 had only 8 participants, we selected 8 participants from Phase 3 (the top 4 in the Nielsen group and top 4 in the conversational agent group) who had identified the most issues from the master list. In Table 1, we refer to the balanced set of 8 participants in Phase 3 and 8 participants in Phase 4 as *chatbot-bal*. We refer to the set of all participants in Phase 3 and 4, 16 participants in Phase 3 and 8 participants in Phase 4, who evaluated the chatbot as *chatbot-all*. We also include the set of 12 participants from Phase 3 who evaluated the Alexa skill as *voice*.

Figure 1 shows that evaluators using the revised conversational agent heuristics identified more usability issues than evaluators using Nielsen's heuristics. In the conversational agent group, a single evaluator found 20 issues, while a single evaluator found 13 issues in the Nielsen group. Four evaluators in the conversational agent group were able to find 56% of the usability issues, compared to four evaluators in the Nielsen group who found 44% of the issues.

We propose that the proportion of unique issues found by the conversational agent group is higher than those found by the Nielsen group. To test this hypothesis, we used a statistical test to compare the proportion of unique issues found by each evaluator. We consider a unique issue to be an issue found only by one heuristic set, Nielsen or conversational agent, and not found by both sets. We found that evaluators using the conversational agent heuristics found significantly more unique issues (M= 0.42, SD = 0.17), than evaluators using Nielsen's heuristics (M= 0.19, SD = 0.09), t(6) = 2.47, 95% CI = [-0.461,-0.002], $p<0.05$. Evaluators using Nielsen's heuristics found on average 19% unique issues.

We then grouped the usability issues to better understand the types of issues that the heuristic sets cover.

Content. The revised heuristics address 4 out of 8 issues related to the content of the dialogue, while Nielsen's set only identified 3 of the issues in Phase 4. The conversational agent heuristics may better identify issues related to the comprehensibility of the chatbot dialogue, such as issues with the wording of questions and explanations of acronyms. There were two issues identified solely by the experts: "dialogue is written at an advanced reading level" and "too many chatbot messages in a row". The dialogue content should not only be understandable, but it should be adapted to the reading level of the conversational agent users.

Answer interaction. The revised heuristics address 8 out of 10 issues related to interactions with questions and responses. The conversational agent heuristics may encourage the designers to consider intuitive and free-form ways to respond to the conversational agent. These issues included users being limited to answer options that might not describe their circumstances, lack of answer validation and confusion about the "explain" feature of the chatbot. One issue, "unclear how to submit text input", was only identified by a participant in the Nielsen group, but they did not assign it one of Nielsen's heuristics and instead labeled it as having "no heuristic".

Guidance. The revised conversational agent heuristics identify all of the 6 usability issues sorted under help and guidance. We speculate that due to the development of the heuristic *Help and guidance*, evaluators using the conversational agent heuristics were able to generate more issues in this area.

Data Privacy. The heuristic *Trustworthiness* was used to identify issues related to data privacy. The revised heuristics identified 2 out of 3 issues, including one issue that data was downloaded at the end of the conversation without notifying the user.

Settings. The issues related to the chatbot's settings highlights an area in which the conversational agent heuristics faced limitations. The revised heuristics identified only 2 of the 6 issues related to the conversational agent settings. The heuristic *Help and guidance* emphasizes that guidance should be provided during the conversation. This may lead evaluators to focus less on other forms of help that exist in the interface, like the settings menu. Potential revisions could be made to address providing user guidance and feedback outside the dialogue in conversational agents with GUIs.

DISCUSSION

In this work, we propose and validate a set of 11 heuristics for conversational agents that can be generalized to text, voice and multi-modal conversational agents. We found that the conversational agent heuristics are useful for identifying more usability issues than Nielsen's. While usability heuristics traditionally focus on providing a clear and efficient experience, the design of conversational agent interfaces may need to go beyond usability. Providing a good user experience may require an evaluation of the conversation as well as user interactions. By explicitly calling out new design principles, evaluators consider new usability issues that may not be prioritized using Nielsen's heuristics. It is important for designers to support user expectations of context preservation [4]. Participants often noted that the chatbot seemed confused when it asked unnecessary follow-up questions. Though conversational agents may have varying levels of context handling, storing the user's recent state would help to maintain relevance in the conversation. Additionally, the conversational agent should be truthful in its interactions to foster trustworthiness [9]. The conversational agent should not mislead users about its identity, nor withhold important information about how user data will be used. Additionally, while the user may require information on how to interact with the conversational agent, they should not be overwhelmed with too much information. In particular, it may be difficult to recognize the system status and remember instructions when using a voice interface. Thus, *Help and documentation* has been removed from the heuristic set and it has been adapted, along with *Recognition rather than recall*, into *Help and guidance*. Users may need feedback and guidance throughout the conversation to better

Conversational Agent Heuristics	
Visibility of system status The system should always keep users informed about what is going on, through appropriate feedback within reasonable time, without overwhelming the user.	**Help and guidance** The system should guide the user throughout the dialogue by clarifying system capabilities. Help features should be easy to retrieve and search, focused on the user's task, list concrete steps to be carried out, and not be too large. Make actions and options visible when appropriate.
Match between system and the real world The system should understand and speak the users' language—with words, phrases and concepts familiar to the user and an appropriate voice—rather than system-oriented terms or confusing terminology. Make information appear in a natural and logical order. Include dialogue elements that create a smooth conversation through openings, mid-conversation guidance, and graceful exits.	**Flexibility and efficiency of use** Support flexible interactions depending on the use context by providing users with the appropriate (or preferred) input and output modality and hardware. Additionally, provide accelerators, such as command abbreviations, that are unseen by novices but speed up the interactions for experts, to ensure that the system is efficient.
User control and freedom Users often choose system functions by mistake and will need an option to effortlessly leave the unwanted state without having to go through an extended dialogue. Support undo and redo.	**Aesthetic, minimalist and engaging design** Dialogues should not contain information which is irrelevant or rarely needed. Provide interactional elements that are necessary to engage the user and fit within the goal of the system. Interfaces should support short interactions and expand on the conversation if the user chooses.
Consistency and standards Users should not have to wonder whether different words, options, or actions mean the same thing. Follow platform conventions for the design of visual and interaction elements. Users should also be able to receive consistent responses even if they communicate the same function in multiple ways (and modalities). Within the interaction, the system should have a consistent voice, style of language, and personality.	**Error prevention** Even better than good error messages is a careful design of the conversation and interface to reduce the likelihood of a problem from occurring in the first place. Be prepared for pauses, conversation fillers, and interruptions, as well as dialogue failures, deadends or sidetracks. Proactively prevent or eliminate potential error-prone conditions, and check and confirm with users before they commit an action.
Help users recognize, diagnose and recover from errors Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.	**Trustworthiness** The system should convey trustworthiness by ensuring privacy of user data, and by being transparent and truthful with the user. The system should not falsely claim to be human.
	Context preservation Maintain context preservation regarding the conversation topic intra-session, and if possible inter-session. Allow the user to reference past messages for further interactions to support implicit user expectations of conversations.

Table 2: The final set of conversational agent heuristics.

understand the status of the system, how they can search for help and what options are available to them.

WORKSHOP GOALS

In this workshop paper, we present the design of heuristics for conversational agent interfaces to identify usability issues that may not be addressed by Nielsen's heuristics. We hope to participate in this workshop to receive feedback and share the outcomes of this work with the research community in academia and industry. Our goal is to engage in dialogue on the opportunities, challenges and future work in conversational agent design.

ACKNOWLEDGMENTS

We would like to thank all of the experts who reviewed the heuristics and our study participants for their invaluable time and feedback. This work was in part supported by the Science Foundation Ireland ADAPT Centre (13/RC/2106).

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2018. The state of speech in hci: Trends, themes and challenges. *arXiv preprint arXiv:1810.06828* (2018).

[3] Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58.

[4] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 895–906.

[5] Ross James Lordon. 2019. *Design, Development, and Evaluation of a Patient-Centered Health Dialog System to Support Inguinal Hernia Surgery Patient Information-Seeking*. Ph.D. Dissertation. University of Washington.

[6] Jennifer Mankoff, Anind K Dey, Gary Hsieh, Julie Kientz, Scott Lederer, and Morgan Ames. 2003. Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 169–176.

[7] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45.

[8] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 249–256.

[9] Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons* 62, 6 (2019), 785–797.

[10] Nina Svenningsson and Montathar Faraon. 2019. Artificial Intelligence in Conversational Agents: A Study of Factors Related to Perceived Humanness in Chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*. 151–161.

[11] Zhuxiaona Wei and James A Landay. 2018. Evaluating Speech-Based Smart Devices Using New Usability Heuristics. *IEEE Pervasive Computing* 17, 2 (2018), 84–96.