An Explainability Case-Study for Conversational User Interfaces in Walk-Up-And-Use Contexts

Tim Schrills
University of Lübeck
Lübeck, Germany
schrills@imis.uni-luebeck.de

Leon Schmid
University of Lübeck
Lübeck, Germany
leon.schmid@me.com

Hans-Christian Jetter
University of Lübeck
Lübeck, Germany
jetter@imis.uni-luebeck.de

Thomas Franke
University of Lübeck
Lübeck, Germany
franke@imis.uni-luebeck.de

ABSTRACT

Current research shows that interactions with conversational user interfaces (CUI) miss requirements for good usability, e.g. sufficient feedback regarding system status. Within a user-centred design process we created different design approaches to explain the CUI's state. A prototypical explainable conversational user interface (XCUI) was developed, which explains its state by means of representations of (1) confidence, (2) intent alternatives, (3) entities, and (4) a context time line. The XCUI was then tested in a user study (N = 49) and compared with a conventional CUI in terms of user satisfaction and task completion time. Results indicated that completion time and satisfaction improvement were dependent on specific task characteristics. The effects of the implemented XCUI features potentially resulted from task-specific needs for explanation. This could be based on the tasks' different complexity indicating the potential need for adaptive presentation of explainability features.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Conversational User Interfaces, Explainable Artificial Intelligence, Explainable Conversational User Interfaces

1 INTRODUCTION

Boosted by techniques commonly referred to as artificial intelligence (AI), end-user devices with integrated conversational user interfaces (CUI) [34] such as intelligent speakers or smartphones have been growing quickly in popularity during recent years. Although this technology promises some advantages (such as operation without the use of hands for voice user interfaces, or natural language input in general), various studies find low usability or satisfaction with CUI applications [38]. For example, a common strategy amongst users is to try out different phrases as instructions to achieve their goal - and consequently, users either have to learn how their system works [43] or are not able to use all functionalities [37].

These barriers to successful interactions become particularly apparent in walk-up-and-use systems [20], which focus on spontaneous and short-term use by different users - e.g. in reception areas in hospitals, offices or administration.

We address these challenges by providing additional explanations which grant users information about the system's state and resulting interaction possibilities. Such procedures can be summarized as explainable artificial intelligence (XAI) [25]. However, there has not been sufficient research on how key explainability features can be integrated into an existing CUI to form an explainable conversational user interface (XCUI). In this study, we particularly focus on walk-up-and-use-systems, where users can neither prepare for the use of the system nor have the possibility to take advantage of optimizations over time (e.g. by adapting the system or through learning effects from the user). To this end, we conducted a theoretically grounded design process for explainability features as well as an online user study.

Our three main contributions are: **(1) initial designs** of an explainability approach, consisting of four explanation features which supplement the primary speech modality of a CUI in the context of a walk-up-and-use situation with visual explanations. These were derived from a user-centered design process and built up on existing approaches such as [35] in terms of providing more detailed information to the user, enhancing understanding by relying on contrastive explanations (c.f.[41]) and combining information from natural language processing (NLP) as well as intent recognizing modules; **(2) a comparative user study** of a walk-up-and-use CUI with and without our designs of explanation features. N = 49 users were assessed in regard to task completion time and user satisfaction; **(3) design-relevant findings** about the intricate and non-trivial relations between user tasks and the design and benefit of the different explanation features, e.g., our results revealed that the depiction of entities identified by the CUI is better suited for complex tasks while competing intents should be displayed for easy tasks.

We will first describe relevant concepts from the field of CUI and explain which interface improvements could be applied to current barriers and how the researched context differs from previous studies. We then explicate the conducted design process and the

Mensch und Computer 2021, Workshopband, Workshop on User-Centered Artificial Intelligence (UCAI '21)

https://doi.org/10.18420/muc2021-mci-ws02-377

created visual explainability features for an XCUI. Subsequently, we present our research strategy to test the explainability features as well as the empirical results. Finally, we discuss the meaning of the feature-related results for further research and CUI designs. Hence, we discuss the following research questions along our research process:

(1) **RQ1:** Which explainability features appear promising to provide information about a system's state?
(2) **RQ2:** Which impact do these explainability features have on task completion time and user satisfaction (comparison CUI vs. XCUI)?

2 RELATED WORK

2.1 Interacting with Conversational User Interfaces

[39] defines Conversational User Interfaces (CUI) as interfaces or frontends, accepting speech, text and touch input, built e.g. for chat bots or conversational assistants [40]. CUI differ from other interfaces in their ability to understand natural language and therefore communicate with people in a natural way [40]. Other characteristic features are the ability to remember conversational content [40] or to include environmental information [44].

CUI enable applications commonly referred to as Virtual Personal Assistants/Voice Assistants [39], Intelligent Assistants [6], Intelligent Personal Assistants [13] or Conversational Agents [38]. In this paper we use the term CUI, since it includes all of the above. We also distinguish our research from studies dealing with Embodied Conversational Assistants (ECA), which focus on imitating a human conversational partner as authentically as possible on different levels (language selection, intonation, but also body language and gestures or facial expressions), see [44]. In contrast, our work focuses only on the communication of information and does not aim to represent emotional or affective states in order to achieve a higher resemblance to humans.

However, based on the promise of a natural way of communication, CUI are required to meet the high expectations that stem from non-digital communication between people, e.g. the ability to integrate context information into an answer and, if necessary, to state which context information has been used for an answer [23]. To manage users' expectations, previous approaches to optimize CUI often try to support users during the development of valid mental models or to correct faulty mental models. Mental models can be defined as an abstract cognitive representation of systems consisting of relevant elements needed to perform a given task [9]. Optimizing them can be achieved e.g. by well-designed feedback [19], or by explicitly explaining certain functions of the system. In a walk-up-and-use context those systems aim to provide a lightweight decision without prior training (as in [5, 20, 31] or researched in [18]). Here, previously described approaches fall short because interaction is not frequent or continuous enough - in the worst case, there is only one conversation/interaction. In these cases it makes sense to design a user interface that requires as few assumptions as possible about the mental model of users and displays information that can be used by different users without previous training or tutorials. As opposed to training users and their mental models over a longer period of time and usage, our approach aims at immediately communicating system state and interaction possibilities by the explicit addition of explainability features.

2.2 Challenges for Improving CUI Interaction

The problem of users falling back on trial-and-error in interactions is often attributed to the fact that feedback from a system does not allow users to sufficiently observe (1) the current state of a system (i.e. observability as a key facet of explainability) and to understand (2) possible actions of the system and associated commands [37, 43]. Critical requirements for usability and user experience in human-machine interaction (such as overcoming the gulfs of execution and evaluation [42]) are therefore not met. Instead, usability problems are further reinforced by system characteristics: communication, e.g. in the auditory domain, imposes strong limitations on how much information can be output simultaneously [17]. Overall, information transmitted during a conversation has a shorter time span within which it persists [48]. As [45] and [39] show, supplemental visual support can be utilized to display additional information, thus mitigating the effects of the auditory domain. Although previous research has shown that users may feel frustrated having to deal with visual information when they expected to communicate in spoken language (e.g. [6]), we expect additional and information-rich interaction possibilities to be helpful especially in walk-up-and-use systems (see also [39]).

Users expect CUI to understand their spoken requests [32]. There are various techniques aiming to understand the user's intent, for instance tokenization, bag of words and regular expressions, but machine learning techniques such as Deep Learning (c.f. [28]) are increasingly being used for this purpose [39]. While decisions of rule-based systems can be presented in an adequate manner, doing the same is more difficult for machine learning systems [53]. Understanding which processes happen in e.g. a deep neural network (DNN), accessing, analysing and expressing these processes is a current challenge in research [46]. This lack of comprehensibility for the user has been criticized by several researchers [1, 14, 25]. In their design guidelines for Artificial Intelligence Amershi and colleagues [2] strongly advocate for an explicit explanation of the AI systems' behavior, such as the display of contextually relevant information or the scoping of services when the user's intent is not sufficiently clear.

If a system is not explainable the user may remain indecisive even if the results are correct and the system is trustworthy enough [10]. Missing explanations on the side of e.g. CUI impedes the development of user acceptance and prevents users from building up trust for a system [11, 27, 30].

2.3 Explanations in AI Systems and CUI

Research shows that a variety of AI systems' users trust a system more when it is able to explain its decisions [29]. Since people form a mental model of a system's functionality, this helps them in their later usage through valid explanations [51]. Additionally, the explanation of a system's behaviour increases both the understanding of and the trust in the system, especially when a user doesn't know how the system works [36]. The explanation can also serve as a

mechanism to check if the system is working correctly [15]. Furthermore, explanations are necessary when a system intelligently includes the context and uses machine learning models so that users can't tell how exactly the system reached its decision [29]. Especially in walk-up-and-use systems, so called local explanations, that focus on explaining the reasons for a special action or decision of a system rather than the general functionalities, may prove helpful to avoid trial-and-error situations [47]. Therefore, as a step towards higher transparency and explainability of AI systems [14] the goal of this study is to shed light on the effects of selected features introducing explainability into walk-up-and-use CUI.

In previous research, different approaches to achieve explainability have been developed and, in part, tested: for example, [22, 24] provide insights on displaying information on why a system didn't choose an alternative, e.g. categorization of an image recognition system. Informing users about very likely, yet not chosen alternatives is called contrastive explanation [41] or counterfactual explanation [7]. Users benefit from contrastive explanations because they allow a variety of cognitive manipulations that help to interact with AI systems [7].

Another idea from explainable artificial intelligence is to highlight the input features accordingly to their importance for the system's result. In the field of image recognition, for example, heatmaps are used to show which pixels have been most relevant (or irrelevant), leading to different XAI-techniques such as sensitivity analysis [3] or layerwise relevance propagation [4]. Looking at CUI, displaying important keywords which were used in order to extract the meaning of an inquiry has already been established [35]. However, additional features such as displaying the individual relevance of certain inputs (i.e. words) or to what extent some inputs may reduce general confidence, still need to be implemented and tested.

A common technique in systems relying on deep neural networks, e.g. in image recognition, is the display of probability of a chosen outcome (i.e. a target class) given the presented stimuli. Information about the confidence value of results aim to enable users to reflect the reliability of, e.g. an image categorization. As studies showed, confidence ratings are able to influence human machine cooperation [16, 50]. Yet, when it comes to CUI, especially in those relying on spoken language, many systems do not provide information about the confidence regarding intent detection or speech recognition. One can argue that single-turn conversations do not need to provide confidence information, since users may be satisfied with an answer or not.

In summary, while holding much potential for further improvements of CUI, explainability features need to be designed carefully and by following a user-centered design. Therefore, in our present research we chose to integrate potential users' ideas about tasks, context and features and evaluate the developed prototype accordingly.

2.4 Use Case

For this research, the selected use case is exemplary for the use of a walk-up-and-use system: the reception of a collaborative work space. Typical tasks are the request of schedules, directions, information about meetings or arrival and departure options. On one hand, this use case represents different response types (dynamic information e.g. about meetings or static information e.g. about rooms) and, on the other hand, only a short interaction span.

2.5 Design Process

To begin the design process, the tasks that the CUI in our walk-up-and-use case would be able to work on were defined in a workshop with potential users. These tasks were grouped into user requests regarding (1) transportation connections, (2) appointments, (3) spatial searches and (4) simple tasks. Afterwards, a 3-step design process was conducted in order to create an XCUI that would explain its state.

2.5.1 Identifying the elements of a CUI's state. The goal of the first step was to identify elements that could be used to explain the state of a CUI. [44] and [39] describe the functionality of CUI and specify the elements *intent*, *entity* and *context*. Gupta et al. give an example of a CUI system including those elements and add *dialog management* and *speech recognition* [26]. The state of a CUI can therefore be described using these elements. Thereafter, different services (Rasa, Google Dialogflow, IBM Watson Assistant) were analyzed to find out what information they provide that could be used to explain the elements above. It turned out that Rasa gave the most information, especially about intents and entities when compared to other services like Google Dialogflow or IBM Watson Assistant. As a speech recognition service, IBM was to be selected since it also provided the most information on *speech recognition*.

2.5.2 Ideation workshop. Having identified the elements of a CUI state and what information is provided about these elements, in the next step an ideation workshop was conducted to collect possible ways of displaying the information and thereby the state of a CUI. The workshop followed the design studio method and was carried out with six participants over a video conference tool. In two iterations with an ideation phase and feedback phase each, the participants firstly created ideas on their own on how to explain a specific CUI element using the existing information and secondly discussed as a group how these explanations could be integrated in a concept for the whole state with all elements. The guiding question was which explainability features can provide information about the system status.

In the first iteration, the following ideas for each element were created: *Intents* could be explained by displaying alternatively detected intents of an inquiry as buttons. By clicking on a button, the intent could be changed accordingly. The recognized *entities* of an inquiry could be highlighted in color. To display the variables that are saved in the *context* during an interaction, the participants thought of a timeline where each variable could be presented. If the CUI takes a variable into account to answer a question during an interaction, the corresponding variable should be highlighted on the timeline. Explaining the *dialog management* was difficult for the participants due to insufficient information about that element. By displaying the transcribed inquiry, the element *speech recognition* is already partially explained. To give even more information, the participants suggested reporting the confidence score of this recognition process by using an avatar or emoji.

Explaining the elements together in one concept, in the second iteration the participants agreed on explaining the confidence value

of *speech recognition* together with the confidence value of the *intent* detection process. *Dialog management* was left out in the concept as well as the element *intent* because users could also correct their intent by formulating another inquiry.

2.5.3 RQ1: Creating an XCUI concept. Based on the results of the ideation workshop, an XCUI concept was created that consisted of four features which aim to explain the system's state. The concept mostly followed the results of the ideation workshop and therefore contained explanations of *entities* and *context*, leaving out the explanation of *dialog management*. As suggested, the concept also included the explanation of the confidence scores of *intent* and *speech recognition* by changing the facial expression of an avatar. The functionality of viewing and selecting intent alternatives was added even though the workshop's participants didn't consider it, since the element would help to reduce the task completion time and therefore affect RQ2. After being trained, the system was able to identify 24 different intents. All in all, the explainability features of the XCUI are named: (1) intent alternatives, (2) confidence, (3) entities and (4) context timeline. Figure 1 gives an example of how intent alternatives are displayed using buttons, while Fig. 2 shows the implementation of the context timeline and a highlighted entity. In addition, both figures include avatars with different facial expressions representing different confidence levels.

Figure 1: Example of Intent Alternatives

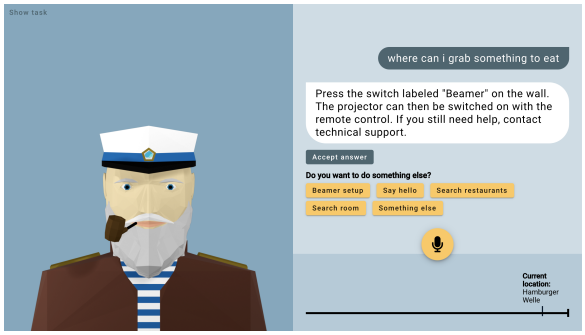
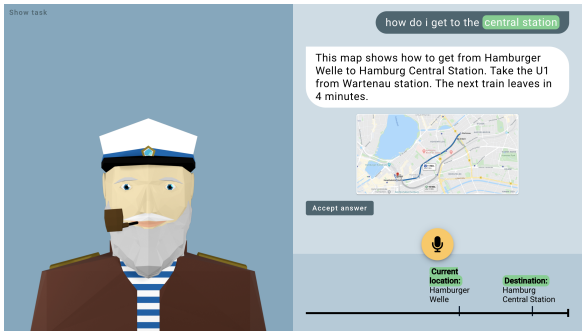


Figure 2: Example of Explainability Features: Context Timeline and Entities

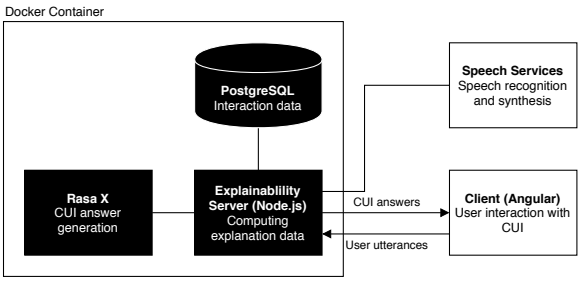


The concept also included ways to support the tasks which would be carried out with the CUI. For instance, a traffic connection task would be displayed along with a map visualizing the traffic route as seen in Fig. 2 and search results of spatial tasks would be presented on a suitable map. Following [49] we enhanced the CUI by displaying spatial tasks, or tasks including traffic connections on the screen instead of reporting them acoustically, thereby reducing the user's cognitive workload.

2.6 System Implementation

In order to evaluate the XCUI in the next chapter, the concept was first implemented. Fig. 3 shows the architecture of the system consisting of three applications and four connected services.

Figure 3: XCUI architecture



2.6.1 Rasa X. The core of the XCUI system is Rasa X, the framework that was used to build the CUI. By using training sentences and sample conversations, a model was built that can be used to give answers to user requests. Rasa X runs on multiple Docker containers, versions of its model used Git and can be accessed via an https API.

2.6.2 Client. The client implements the XCUI, which the users interact with. When speaking to the XCUI the spoken inquiry of the user is sent to the explainability server where the request is handled. To gather data for the study, the client also collects interaction data such as satisfaction scores and completion time of tasks from the users.

2.6.3 Explainability server. This instance is responsible for handling user requests from the client. Speech is transcribed using the connected IBM speech-to-text service. The written request is then sent to Rasa X to get an answer from the CUI, which is played to the user on the client by first synthesizing it as spoken output using the Google text-to-speech service. In order to display the explanations of the XCUI on the client side, this server is also responsible for getting the explanation information including data about the recognized intent, entities and context from Rasa X and then computing the needed explanation data which is sent to the client. Any interaction data that is collected by the client is saved by the explanation server into a PostgreSQL database.

3 USER STUDY

To examine the effects of the designed XCUI features we designed an experimental user study focusing on the comparison of CUI vs. XCUI regarding user satisfaction and task completion time (within-subject design). Due to COVID-19 restrictions, the study had to

be transferred into an online experiment. However, particular care was taken to maximize external validity and quality control (e.g. checking for screen size or performing a functionality test before the experiment) in order to achieve valid results. In sum, the conditions were not much different from the originally planned lab setting.

3.1 Methods

3.1.1 Participants. For the present study, N = 49 participants (29 female, 19 male, 1 neither male nor female; *Mean* = 31.2 years; *SD* = 11.8) were recruited via university e-mail lists and social media. 69.4 % of the participants had an academic degree. Because research in human-computer interaction is at risk of a biased sample selection (e.g. because this kind of research attracts participants with high technological affinity, see [52]), we checked our sample by assessing affinity for technology interaction (ATI, [21]). Our sample had a wide range (from 1.70 to 6.00) with an average value of 3.90 being close to the answer scale mean of 3.50. Additionally, the standard deviation of the study sample (*SD* = 0.97) falls within the range of prior studies (*SD* = 0.87 to 1.09, see [21]).

3.1.2 Setting and Procedure. A biphasic design was chosen for the experiment:

In the first phase **(P1)** participants had to check their requirements (display size, microphone and speaker). To prevent further technical problems, participants read given phrases aloud and the speaking velocity was measured in order to ensure that everyone accomplished the tasks in the same velocity.

In the following phase **(P2)** participants performed nine simple tasks which resulted from the task analysis. Participants were divided into two groups, A performing tasks (1-4) with XCUI and B (5-9) respectively. Accordingly, A performed tasks 5-9 with the CUI and B tasks 1-4 with the CUI. In the end the sequence of the tasks was randomized to avoid sequence effects. For each task, the task completion time was tracked and user satisfaction was elicited on a six-level scale. The differentiation in two groups only existed in phase 2. In the second phase the following dependent variables (DV) were collected: **(DV1)** task completion time (ms) as a proxy for efficiency and **(DV2)** user satisfaction (6 item likert scale). For further analysis, means for time and satisfaction were calculated. After the second phase, the Explanations Satisfaction Scale [29] was conducted to evaluate the XCUI. All metrics are intended to measure the system's usability.

3.2 Results

Our general results including user satisfaction and completion time are shown in Table 1. The time values were z-standardized for better comparability. Neither user satisfaction nor task completion time showed a significant difference on the overall sample.

Group A, likewise the overall sample, showed no differences regarding satisfaction or task completion time. The effect sizes indicate that when using the XCUI participants of Group A were less satisfied and needed longer to complete the task. In Group B no significant effects for task completion time were found. However, Group B showed a significantly higher satisfaction using the XCUI. The general assessment of satisfaction results through the ESS were in the middle range of the scale (*Mean* = 3.34) with a rather low variance (*SD* = 0.88).

4 GENERAL DISCUSSION

In the present study, we designed a conversational user interface for a walk-up-and-use system, therefore demanding a high level of explainability. Four interaction features (intent, confidence, entities and context timeline) were developed and tested against a system without these features.

Techniques to Imbue Explainability. Four different concepts were initially identified as particularly promising to improve the explainability of a CUI system (c.f. *RQ1*: Which explainability features appear promising to provide information about a system's state?). Within a user-centered design process, four matching explainability features were created based on the developed concepts: Firstly, (1) recognized intents should be displayed as well as other possible intents and should be selectable. In the sense of a contrastive explanation, especially obvious intents are presented. Furthermore, (2) the recognized entities were color-coded to express the relevance of single recognized words. Since these entities can retain relevance over a longer period of time in the conversation, one (3) timeline marking all used context information was added. Finally, the (4) confidence with which the CUI interprets the statements was also shown in a conceptual way via the facial expression of an avatar. However, for unfamiliar systems, this representation of confidence might was too subtle and participants did not recognize it. Within our experimental user study we found our features in general to increase subjective preference without having a negative impact on task completion time or satisfaction.

Finding 1) We showed exemplary, how intent, confidence, context and entities can be incorporated into a visual display for CUI systems without having a negative impact on task completion time or user satisfaction.

Higher Satisfaction for Complex Tasks. The results indicate that the effects of the explanatory components depended on the specific tasks given and the reception of those components was influenced by task characteristics. In regards to the second research question (*RQ2*: Which impact do explainability features have on task completion time and satisfaction?) there did not appear to be significant differences in task completion time or user satisfaction across the whole sample. However, when looking at the two groups (A and B), we found no significant results in Group A, which mainly carried out easy tasks with the XAI. This contrasts Group B, where users were faster and showed higher satisfaction.

Finding 2) For complex tasks assisted by a walk-up-and-use voice assistant with a display, highlight recognized entities relevant for the given task and present them in a timeline in order to demonstrate the system's state and capabilities. For easy tasks assisted by a walk-up-and-use voice assistant with a display, show information regarding recognized intent alternatives.

It can be concluded that the added components of explainability possibly have different effects under different task conditions, i.e. task complexity. [12] found in their study on the use of information for task processing that different information is used by users -

Table 1: Users' task completion time and satisfaction (on a scale of 1-6)

	Descriptives Statistics				Paired T-Test			
	M_{XCUI}	SD_{XCUI}	M_{CUI}	SD_{CUI}	t	df	p	d
Overall								
TCT	28851	17311	29666	18512	0.28	48	.781	0.04
Satisfaction	4.63	1.09	4.45	1.11	1.16	48	.125	0.17
Group A								
TCT	30395	13547	22971	14753	-1.91	19	.071	-0.43
Satisfaction	4.34	1.30	4.58	1.17	-0.93	19	.818	-0.21
Group B								
TCT	27787	19655	34284	19639	1.69	28	.101	0.32
Satisfaction	4.84	0.881	4.35	1.07	2.56	28	.008	0.48

Note. $N_{\text{overall}} = 49$; $n_{\text{Group A}} = 20$; $n_{\text{Group B}} = 29$. TCT = task completion time (in ms). TCT values were z-transformed before testing.

depending on how complex the task is . Since the XCUI explanations of its status can be interpreted as additional information, it is possible that this additional information was also used for more complex tasks in the present study. [8] describes how more divergent approaches are more often used for complex tasks than for simple ones.

On a technical level, [33] already showed that the consideration of semantically similar intents is a good possibility to develop an efficient but adaptive assistance system. Here, the AidMe system from [33] was developed to add user specific intents to the already trained ones. Focusing on only one feature, such as intents in walk-up-and-use systems, can also help to reduce the cognitive load, which should be investigated in future studies.

5 LIMITATIONS AND FURTHER WORK

Even though we kept the tasks as unspecific as possible in the context of this research, some limitations have to be taken into account. For example, the XCUI system did not take into account any prior knowledge of the user. This can be especially relevant for the representation of used entities. In the present case, this was not included in our research, since personalization of answers in walk-up-and-use systems is based on direct interactions most of the time.

The analysis of the results showed clear effects, which are probably due to certain task types. While this was not considered before the development of the experiment, future investigations should explicitly investigate task parameters. For example, it should be investigated whether the adaptive (and exclusive) use of explainability features for more complex tasks and not for easy tasks leads to higher overall satisfaction. Although the results found in our evaluation clearly indicate this, the experiment itself is not designed to analyze satisfaction variation based on task differences.

Based on the feedback we gathered during our user-centered design process, we chose a symbolic representation of confidence by facial expression of a virtual character. This stood in contrast to the design of other features such as the abstract timeline or visual highlights in the transcribed text. Future work should focus on the question if more consistency in terms of abstraction or symbols could increase recognizability and usability. Finally, the sample used in the evaluation consists mainly of people with a higher educational background, whereby a wide distribution of ATI values in particular indicates that the sample is sufficiently diverse.

6 CONCLUSION

One reason for low usability of CUI is missing feedback of the system's state to users. Within our research, we deducted, designed and implemented four different explainability features to challenge this shortcoming. First of all, we discovered that CUI in walk-up-and-use systems such as reception desks have special requirements: CUI must be able to display enough information on system status and action options due to limited learning resources. Guidelines for the design of human-centered AI and CUI were used to reduce ambiguity and, based on these guidelines, first concepts for the improvement of CUI were developed. We focused on four central components of a conversation that we wanted to represent: 1) the recognized entities, 2) using contextual information, 3) the recognized and optional intents and 4) the confidence of the recognition.

Within the framework of a user-centered design process, we further developed and implemented the results in a prototypical XCUI that processes information about the course of the conversation and presents it. Based on our results, it can be assumed that use of the XCUI in more complex tasks may leads to higher satisfaction. At the same time, it shows that further research is needed on the connection between the concrete task and additional explanations of the CUI. Therefore, the given research design could be developed to integrate further measurements regarding task complexity (e.g. perceived workload) and design tasks in order to evoke different explanatory demands.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. (2018), 1–18. https://doi.org/10.1145/3173574.3174156

[2] Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, and Paul N. Bennett. 2019. Guidelines for Human-AI Interaction. (2019), 1–13. https://doi.org/10.1145/3290605.3300233

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *The Journal of Machine Learning Research* 11 (2010), 1803–1831.

[4] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*. Springer, 63–71.

[5] Andrew Bragdon, Robert Zeleznik, Brian Williamson, Timothy Miller, and Joseph J. LaViola. 2009. GestureBar: Improving the Approachability of Gesture-Based Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2269–2278. https://doi.org/10.1145/1518701.1519050

[6] Raluca Budiu and Page Laubheimer. 2018. Intelligent assistants have poor usability: A user study of Alexa, Google assistant, and Siri. *Nielsen Norman Group. Available online at https://www. nngroup. com/articles/intelligentassistant-usability/(last accessed 4/12/2019)* (2018).

[7] Ruth M. J. Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 6276–6282. https://doi.org/10.24963/ijcai.2019/876

[8] Donald J Campbell. 1988. Task complexity: A review and analysis. *Academy of management review* 13, 1 (1988), 40–52.

[9] John M Carroll and Judith Reitman Olson. 1988. Mental models in human-computer interaction. In *Handbook of human-computer interaction*. Elsevier, 45–65.

[10] Davide Castelvecchi. 2016. Machine learning is becoming ubiquitous in basic research as well as in industry. But for scientists to trust it, they first need to understand what the machines are doing. (2016), 4.

[11] Ioannis Chalkiadakis. 2018. A brief survey of visualization methods for deep learning models from the perspective of Explainable AI. (2018), 20.

[12] Bogeum Choi, Austin Ward, Yuan Li, Jaime Arguello, and Robert Capra. 2019. The Effects of Task Complexity on the Use of Different Types of Information in a Search Assistance Tool. *ACM Trans. Inf. Syst.* 38, 1, Article 9 (Dec. 2019), 28 pages. https://doi.org/10.1145/3371707

[13] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?": infrequent users' experiences of intelligent personal assistants. (Sept. 2017), 1–12. https://doi.org/10.1145/3098279.3098539

[14] K. Darlington. 2017. Explainable AI Systems: Understanding the Decisions of the Machines.

[15] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]* (March 2017). http://arxiv.org/abs/1702.08608 arXiv: 1702.08608.

[16] Na Du, Kevin Y. Huang, and X. Jessie Yang. 2020. Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming. *Human Factors* 62, 6 (2020), 987–1001. https://doi.org/10.1177/0018720819862916 arXiv:https://doi.org/10.1177/0018720819862916 PMID: 31348863.

[17] John Duncan, Sander Martens, and Robert Ward. 1997. Restricted attentional capacity within but not between sensory modalities. *Nature* 387, 6635 (1997), 808–810.

[18] Jennifer L. Dyck. 1995. Problem Solving by Novice Macintosh Users: The Effects of Animated, Self-Paced Written, and No Instruction. *Journal of Educational Computing Research* 12, 1 (1995), 29–49. https://doi.org/10.2190/XU45-HAMB-7L4P-KF4X arXiv:https://doi.org/10.2190/XU45-HAMB-7L4P-KF4X

[19] Malin Eiband, Charlotte Anlauff, Tim Ordenewitz, Martin Zürn, and Heinrich Hussmann. 2019. Understanding Algorithms through Exploration: Supporting Knowledge Acquisition in Primary Tasks. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. Association for Computing Machinery, New York, NY, USA, 127–136. https://doi.org/10.1145/3340764.3340772

[20] Scott Elrod, Richard Bruce, Rich Gold, David Goldberg, Frank Halasz, William Janssen, David Lee, Kim McCall, Elin Pedersen, Ken Pier, John Tang, and Brent Welch. 1992. Liveboard: A Large Interactive Display Supporting Group Meetings, Presentations, and Remote Collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 599–607. https://doi.org/10.1145/142750.143052

[21] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[22] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451* (2019).

[23] David Griol, Nayat Sánchez-Pi, Javier Carbó, and José M Molina. 2010. An architecture to provide context-aware services by means of conversational agents. In *Distributed Computing and Artificial Intelligence*. Springer, 275–282.

[24] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. 2019. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.

[25] David Gunning. 2017. Explainable Artificial Intelligence (XAI). (2017), 36.

[26] N. Gupta, G. Tur, D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M. Gilbert. 2006. The AT&T spoken language understanding system. 14, 1 (2006), 213–222. https://doi.org/10.1109/TSA.2005.854085

[27] Taehyun Ha, Sangwon Lee, and Sangyeon Kim. 2018. Designing Explainability of an Artificial Intelligence System. (2018), 1–1. https://doi.org/10.1145/3183654.3183683

[28] Simon S. Haykin and Simon S. Haykin. 2009. *Neural networks and learning machines* (3rd ed ed.). Prentice Hall, New York. OCLC: ocn237325326.

[29] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. (2019), 50.

[30] Andreas Holzinger. 2018. Explainable AI (ex-AI). *Informatik-Spektrum* 41, 2 (April 2018), 138–143. https://doi.org/10.1007/s00287-018-1102-5

[31] Shahram Izadi, Harry Brignull, Tom Rodden, Yvonne Rogers, and Mia Underwood. 2003. Dynamo: A Public Interactive Surface Supporting the Cooperative Sharing and Exchange of Media. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada) *(UIST '03)*. Association for Computing Machinery, New York, NY, USA, 159–168. https://doi.org/10.1145/964696.964714

[32] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. (2016), 121–130. https://doi.org/10.1145/2854946.2854961

[33] Nicolas Lair, Clement Delgrange, David Mugisha, Jean-Michel Dussoux, Pierre-Yves Oudeyer, and Peter Ford Dominey. 2020. User-in-the-Loop Adaptive Intent Detection for Instructable Digital Assistant. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 116–127. https://doi.org/10.1145/3377325.3377490

[34] J. A. Landay, N. Oliver, and J. Song. 2019. Conversational User Interfaces and Interactions. *IEEE Pervasive Computing* 18, 02 (apr 2019), 8–9. https://doi.org/10.1109/MPRV.2019.2921176

[35] Toby Jia-Jun Li. 2017. Designing a Conversational Interface for a Multimodal Smartphone Programming-by-Demonstration Agent.

[36] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. (2009), 2119. https://doi.org/10.1145/1518701.1519023

[37] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[38] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. (May 2016), 5286–5297. https://doi.org/10.1145/2858036.2858288

[39] Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-32967-3

[40] Michael F. McTear. 2017. The Rise of the Conversational Interface: A New Kid on the Block? 10341 (2017), 38–49. https://doi.org/10.1007/978-3-319-69365-1_3 Series Title: Lecture Notes in Computer Science.

[41] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]* (Aug. 2018). http://arxiv.org/abs/1706.07269 arXiv: 1706.07269.

[42] Donald A. Norman. 2002. *The Design of Everyday Things*. Basic Books, Inc., USA.

[43] Christiane Opfermann and Karola Pitsch. 2017. Reprompts as error handling strategy in human-agent-dialog? User responses to a system's display of non-understanding. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 310–316.

[44] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences* (1st ed.). O'Reilly Media, Inc.

[45] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. (2018), 1–12. https://doi.org/10.1145/3173574.3174214

[46] Gabrielle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. *arXiv:1803.07517 [cs, stat]* (March 2018). http://arxiv.org/abs/1803.07517 arXiv: 1803.07517.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[48] Mikko Sams, Riitta Hari, Josi Rif, and Jukka Knuutila. 1993. The human auditory sensory memory trace persists about 10 sec: neuromagnetic evidence. *Journal of cognitive neuroscience* 5, 3 (1993), 363–370.

[49] Ben Shneiderman. 2000. The limits of speech recognition. 43, 9 (2000), 63–65. Publisher: ACM New York, NY, USA.

[50] Randall D Spain. 2009. The effects of automation expertise, system confidence, and image quality on trust, compliance, and performance. (2009).

[51] Aaron Springer and Steve Whittaker. 2018. "I had a solid theory before but it's falling apart": Polarizing Effects of Algorithmic Transparency. (2018), 13.

[52] Daniel Wessel, Moreen Heine, Christiane Attig, and Thomas Franke. 2020. Affinity for Technology Interaction and Fields of Study – Implications for Human-Centered Design of Applications for Public Administration. (2020), 4.

[53] Jürgen Ziegler. 2019. Challenges in User-Centered Engineering of AI-based Interactive Systems. (2019), 5.