Conversational Interfaces for Explainable AI: A Human-Centred Approach

Sophie F. Jentzsch[1](✉), Sviatlana Höhn[2], and Nico Hochgeschwender[1]

[1] German Aerospace Center (DLR), Simulation and Software Technology,
Linder Hoehe, 51147 Cologne, Germany
{sophie.jentzsch,nico.hochgeschwender}@DLR.de
[2] University of Luxembourg, 6 Avenue de la Fonte, Esch-sur-Alzette, Luxembourg
sviatlana.hohn@uni.lu
https://www.dlr.de, https://www.uni.lu

Abstract. One major goal of Explainable Artificial Intelligence (XAI) in order to enhance trust in technology is to enable the user to enquire information and explanation directly from an intelligent agent. We propose Conversational Interfaces (CIs) to be the perfect setting, since they are intuitive for humans and computationally processible. While there are many approaches addressing technical and agent related issues of this human-agent communication problem, the user perspective appears to be widely neglected. With the goal of better requirement understanding and identification of implicit user expectations, a Wizard of Oz (WoZ) experiment was conducted, where participants tried to elicit basic information from a pretended artificial agent via Conversational Interface (*What are your capabilities?*). Chats were analysed by means of Conversation Analysis, where the hypothesis that users pursue fundamentally different strategies could be verified. Stated results illustrate the vast variety in human communication and disclose both requirements of users and obstacles in the implementation of protocols for interacting agents. Finally, we inferred essential indications for the implementation of such a CI. The findings show that existing intent-based design of Conversational Interfaces is very limited, even in a well-defined task-based interaction.

Keywords: Explainability · XAI · Human-agent interaction · Conversational Interface · Wizard of Oz

1 Introduction

While intelligent agents with advanced planning, learning and decision-making abilities such as autonomous robots are increasingly affecting people's everyday life, their latent processes of reasoning become more and more opaque. Users are often neither aware of the capabilities nor the limitations of the surrounding systems, or at least not to the entire extent. This missing transparency leads to a lack of trust and diffuse concerns towards innovative technologies, which has

D. Calvaresi et al. (Eds.): EXTRAAMAS 2019, LNAI 11763, pp. 77–92, 2019.
https://doi.org/10.1007/978-3-030-30391-4_5

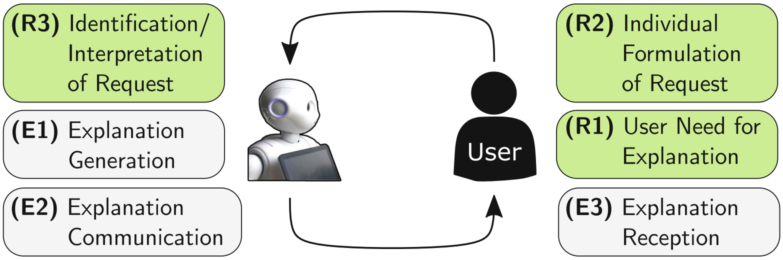

(R3) Identification/ Interpretation of Request
(E1) Explanation Generation
(E2) Explanation Communication
User
(R2) Individual Formulation of Request
(R1) User Need for Explanation
(E3) Explanation Reception

Fig. 1. Illustration of XAI as HCI problem: R1–R3 represent the transmission of user requests to the agent and E1–E3 the agent's provision of explanation.

already been identified as an important issue to be resolved by the AI community [8,27]. For that reason, promoting the explainability of Artificial Intelligence (XAI) is a key condition to enable optimal establishment and exploitation of novel algorithmic decision making techniques.

Many recent approaches in XAI focus on the adaption of involved complex systems, e.g. by providing a detailed description or introducing key information to the user (see for instance [7,9,17,18]). However, without doubting the value of this endeavours, it is not sufficient to tackle the issue exclusively from a machine-centred view with an one-way flow of information. According to Miller, the core of Explainable AI is a *human-agent interaction problem* [20] and therefore rather a dialogue, where two autonomous agents - an artificial and a human one - need to communicate in a way that is intuitive for both of them. This requires the development of appropriate human-agent interfaces and agent protocols to provide information and visualise explanations. In this paper we propose Conversational Interfaces (CIs), similar to ordinary text messengers, to be a perfect setting for successful human-agent interaction (aka. chatbot) due to different advantages: First, it is an intuitive channel of communication for most users, since chatting via instant messengers became a commonplace habit. This is important, because autonomous systems and devices should be as self-explanatory as possible to be utilizable for the standard user. Second, this approach facilitates the agent's interpretation of statements, as written text is directly computational processible, in contrast to e.g. spoken natural language, where an additional step of speech recognition is required, which is sensitive to noise and ambiguity. Besides those superior justifications, the written communication yields the benefit of easy recording and analysis.

Defining XAI as such a dialogue problem (and considering the user behaviour as immutable) there are two main tasks for an agent to solve in terms of successful interaction: On the one hand, it needs to be able to provide comprehensible explanations regarding its computational reasoning, which is challenging to implement for sure. On the other hand, however, it needs to understand human communication patterns to identify user demands correctly in the first place, before even being enabled to tackle the question of information depictions.

Figure 1 illustrates the reciprocal agent-user dialogue, where E1, E2 and E3 describe the agent's frequently discussed provision of **(E)**xplanation [22]. The transmission of user **(R)**equests to the agent (R1, R2, R3) however appears to be fairly neglected in the ongoing XAI debate, as reported by [1], although it can be considered to be no less pretentious. Different user types are presumed to apply different interaction strategies, thus an agent is faced with a vast range of individual idiosyncrasies. It not only needs to be *resistant against* but rather *sensitive for* variance in user interaction to capture its latent requests. As previous research suggests, it should not be the programmer but the end user, who is in charge to determine, which aspects of artificial behaviour are explain-worthy [21]. In fact, a computer scientist will hardly be able to empathise the demands of uninformed users and consequently there is an essential need to identify those systematically.

We experimentally demonstrate the large variability of human interaction strategies by showing that they even affect apparently simple tasks, where users seek explanations. We conduct a Wizard of Oz (WoZ) experiment, where employees of a research lab assume to interact with a chatbot that provides an interface to a Pepper service robot (see Fig. 2). Pepper is acting as an assistant in the contemplated lab, where it is performing the tasks of escorting people, patrolling the building and welcoming visitors. Those tasks are carried out by the robot in a realistic, real-world office environment. For example, Pepper is capable to escort people from the entrance hall to meeting rooms autonomously. To do so, several crucial components such as navigation, path planning, speech and face recognition are required and integrated on the robot. Pepper is a well suitable example for the pretended artificial intelligence in the cover story of this investigation, since it is an actual instance of autonomously operating robots and is potentially accessible via Conversational Interface. Subjects were ask to find out about Peppers capabilities. The task instructions were formulated as open and less restrictive as possible, so that resulting observations reflect individual strategies and illustrate the diversity of human communication (R2). We succeed in inferring implicit expectations of users and major design issues by means of Conversation Analysis. Our human-centric approach to the outlined issue yields a preliminary step towards designing an agent for sufficient self-declaration via Conversational Interface.

In the long run, we see Conversational Interfaces as a promising environment to deliver information about a certain system to the user. Thus, it constitutes an important contribution in increasing the explainability of AI and therefore the trust in autonomous systems.

The superior goal is (1) to test our hypothesis, that users follow different implicit strategies in requesting information from an artificial interlocutor. We expect people's intuition in interacting with such a system to vary widely, what leads to the exposure of concrete requirements in the conception of profound human-agent interaction channels. Hence, we aim (2) to identify associated requirements, risks and challenges. Since the present investigation is a contribution to exploratory research, the motivation is to identify so far unconsidered aspects rather than offering a conclusive solution.

Fig. 2. Pepper the service robot and the human Wizard in the lab.

2 Designing a Wizard of Oz Experiment

We aimed to learn about the implicit expectations of users towards a communicating bot. Therefore, we designed a Wizard of Oz (WoZ) study to collect conversation data and analysed them by means of Conversation Analysis (CA), which allows for inferences about the requirements for the implementation of a Conversational Interface for self-explanatory robots. Both the WoZ and CA are briefly introduced, before the experimental design itself is presented.

Wizard of Oz. The WoZ method is a frequently used and well-evaluated approach to analyse a vast variety of human-agent interactions (also human-robot or human-computer interaction)[25].

In those experiments, participants conduct a specific task while they believe to interact with an artificial agent. In fact there is a hidden briefed person, called the *Wizard*, who is providing the answers. This could for instance be applied, if researchers aim to examine a specific system design that, however, is not implemented yet. In the present case, the task is to find out about the agent's capabilities, while the Wizard is invisible trough the chat interface.

As most scientific techniques, these studies bear some specific methodical obstacles. Fortunately, there is plenty of literature available, defining guidelines and benchmarks for setting up a WoZ experiment [25]. According the classification of Steinfeld et al. [28], we present here a classical "Wizard of Oz" approach, where the technology part of interaction is assumed and the analytic focus is on the users' behaviour and reaction entirely.

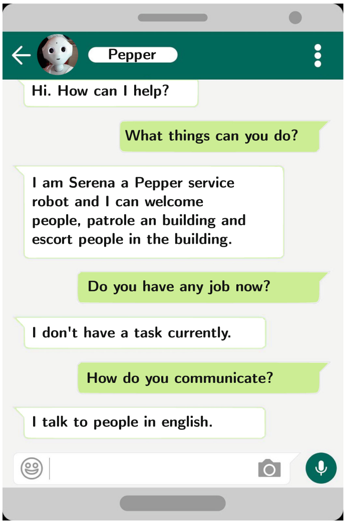
Conversation Analysis. To analyse conversations obtained from the WoZ experiment we employ CA, which is a well-established and standardised approach mainly from the fields of sociology and linguistics [26]. Some related CA-based studies are discussed in Sect. 5. The analysis of data is divided in four sequential steps:

1. **Unmotivated looking**, where the data are searched for interesting structures without any previous conception.
2. **Building collections** of interesting examples and finding typical structures.
3. Making **generalisations** based on the collections from the second step.
4. **Inferring implications** for an implementation in a dialogue system.

Three of them follow the standardised convention of CA and are typically used in those approaches. However, CA is mostly established for exclusively human interactions. As we aim to implement a Conversational Interface based on our findings, the forth step was added to our analysis in order to make the findings applicable in a chatbot. The comprehensive analysis included interactional practices (e.g. questioning) and devices (e.g. upper case writing and use of question marks), as well as turn formats (combination of practices and devices) [6]. Subsequently, we essentially present superior observations, where the steps three and four are mirrored in Sects. 3 and 4, respectively, whereas steps one and two comprise a huge amount of rather particular findings and therefore are omitted in this report.

Experimental Setup. The experimental group comprises seven participants in total (three male and four female), each of them either pursuing their Ph.D. in Computer Science or being already a Postdoc. Because researchers are the main target user group of the intended system, we acquired our peer colleagues via internal University mailing list and in personal invitations, explaining the purpose of the conversation. Hence, the sample group consisted of academics with general technical understanding that, however, were no experts but users of the system. The participants were informed about the exploitation of their anonymised chatlogs for research purposes and agreed. Participants were asked to talk to a chatbot using WhatsApp (illustrated in Fig. 3) without any defined constraints for the conversation, aside from the following instructions:

1. Talk to the chatbot for 15–20 min.
2. Learn about the robot's capabilities.



Hi. How can I help?

What things can you do?

I am Serena a Pepper service robot and I can welcome people, patrole an building and escort people in the building.

Do you have any job now?

I don't have a task currently.

How do you communicate?

I talk to people in english.

Fig. 3. Illustration of a sample snipped from an user's conversation with Pepper.

Pursuant to a WoZ setup, they believed to interact with Pepper that was acting as an assistant in the research lab and were not informed about the responses to originate from a briefed person. By providing this cover story, we hoped to enhance the participants' immersion and make the scenario more tangible. People in the lab knew Pepper, even though not every participant experienced the robots performance, and were likely to take it as a plausible interlocutor.

The sparseness of user instructions was intended, since we were interested in peoples intuitive strategy for interacting with autonomous agents. By formulating the task as open as possible, it has been avoided to suggest a specific approach and the participants were free to evolve their own interpretation.

To specify robot behaviour, we also defined a task description for the Wizard previously, including the following instructions:

1. Let the user initiate the conversation.
2. Do not provide information proactively.
3. Answer the user's question as directly as possible.

The Wizard had a short list of notes at hand with preformulated answers to potential user's questions. The validity of the answers was ensured by the Wizard's background and expert knowledge about Peppers capabilities. To train the Wizard and check the practicability and reasonableness of instructions, the experimental setup was tested in a small pilot study with two participants initially. Those sessions do not contribute to the reported data of this report.

3 User Behaviour in Conversational Interfaces for XAI

The collected dataset consists of 310 turns in total, from which 139 are produced by the Wizard and 171 by participants. The number of turns in each particular experiment was between 33 and 56. Each sessions took between 14 and 20 min, which corresponds to an overall chat time of 121 min. In general, users clearly addressed their utterances to the robot itself in a similar way they would talk to a person using WhatsApp. This is an essential precondition for the validity of the executed dialogue analysis. Each of the seven chat sessions starts with a similar greeting sequence, followed by a *How can I help you?* produced by the Wizard. This question was intended to offer a scope for the user to utter instructions, equivalently to the main menu in a software program.

The purpose of this section is to characterise participants' patterns of interaction that ultimately allow to infer requirements for a self-explanatory Conversational Interface (see Sect. 4). To clarify how exactly users formulate requests, we initially focus on the nature of detached questions posed to the Wizard in Sect. 3.1. From that we generalise to overall user strategies in enquiring information from the agent, where three basic categories are differentiated. Those are presented in Sect. 3.2.

3.1 Users' Question Formulation

The key point of interest in this experiment was how people proceed in enquiring a specific information (*what are your capabilities?*) from an agent. Thus, we turn special attention to the characterisation of formulated user questions.

From 309 turns in total, 125 turns contained questions (about 40,5%), from which 96 question turns were produced by the users (77%) and 29 by the Wizard. The large amount of questions shows that the speech-exchange system of chats was close to an interview, which mirrors the participants' intent to elicit explanation of the system. Several different aspects can be considered to provide an informative characterisation of the users' questions (N = 96).

Question Form. Approximately half of the questions were polar questions (51), meaning they can be answered sufficiently by a simple affirmation or negation (*yes-or-no question*). The other elements were non-polar content questions (45) that required a more extensive answer. In one case, multiple questions were combined in a *through-produced multi-question* [29], this is a single query consisting of several atom questions.

Level of Abstraction. Only 17 questions addressed the robot's capabilities on the high level, meaning they could be answered appropriately by the Wizard by listing the three main actions patrolling, welcoming and escorting (see Example 1). Additional 26 questions addressed the capabilities but required more detailed explanation of the process and included more elementary actions, such as motion mechanisms or ability to move the arms. However, once the Wizard provided information regarding its high level capabilities as in Example 1, users did not ask anything about lower-level ones. This observation illustrates, how the agent's protocol shapes the expectation and intention of the user. Thus, what we earlier referred to as the robot's *main menu* was helpful to restrict the search space and, consequently, to set limits to the Natural Language Understanding (NLU) needs for a potential Conversational Interface. This can be exploited in concrete implementations.

Example 1. The agent explaining its capabilities.

7	15:57	us6	*Yes, that would be lovely. What can you do?*
8	15:57	wiz	*I am Serena, a Pepper service robot, and I can welcome people, patrol a building and escort people in the building.*

Scope of Validity. The temporal information validity specifies whether the question is of *general* nature or concerns the *past*, *current* activities or future *plans*. We additionally differentiated whether the question concerns the robot itself (*internal*) or an *external* entity. Questions with external validity may for instance consider other people or facilities in the first place and elicit information about the robot indirectly.

From 96 user questions, only six concerned an external entity, whereas 90 were directly related to the robot. Thus, participants were clearly focusing pepper and not diverted to other topics. The number of questions for each category of classification is presented in Table 1. Most questions (68) were of general nature and did not relate to any specific action. The other questions were mostly about current and past actions and only a single one included future plans.

3.2 Strategies of Interaction

Participants have been asked to explore the robot's capabilities. Yet, almost none of them did ask about them directly. The strategies of enquiring Pepper's

Table 1. Information validity addressed by user questions: number of observed questions per category - *Static*: A general ability or a constantly valid property; *Past*: A concluded task or past experience; *Current*: An ongoing task or current perception; *Plan*: A pending task or hypothetical behaviour.

Category	Total	Internal	External
Static	68	63	5
Past	13	13	0
Current	14	13	1
Plan	1	1	0

capabilities can be divided in three main categories: (1) User-initiated direct requests, (2) user-initiated indirect requests and (3) undirected chatting that did not appear to follow any strategy at all.

Direct Strategy. A possible approach to inspect Pepper's capabilities, which appears to be quite straightforward, is directly asking for it. Nevertheless, this strategy could only be observed once, as the user asked the chatbot directly *What can you do?*. The remaining six participants followed a more cautious proceeding.

Indirect Strategy. The majority of users preferred to tackle the question of interest in a less explicit manner, meaning they asked for Pepper's capabilities somehow, but left the questions rather open and the intention implicit. Example 2 is just one of many cases, where the user's request was considerably fuzzy. They either formulated a very open statement (that might not even be an actual question), or asked about quite specific abilities and tried to learn about the agent's experience on that field. Occasionally, they also tested concrete functionality or the robot's limitations.

Example 2. Indirect request for the agent's capabilities.

2	12:56	wiz	*Hello. How can I help?*
3	12:57	us7	*I am not sure, but I would like to talk about yourself*

Obviously, it is not in line with people's intuition to formulate distinct and unambiguous requests, but to express their aim implicitly. Deciphering such utterances definitely constitutes a major challenge for such an agent.

No Strategy. In some cases, we observed an even more obscure user behaviour. Even though participants had the clear instruction to find out about the agent's capacities, some did not seem to pursue this target in any way. In these cases, the Wizard's initial question was left entirely unacknowledged, as in Example 3.

Example 3. Undirected chatting without evident intention.

3	10:48	wiz	*How can I help?*
4	10:49	us1	*I am user1, who are you?*
5	10:49	wiz	*I am Serena a Pepper service robot.*

There were extensive sequences of undirected chatting that did not even include a single question towards the agent's activities. Certainly, there could be a hidden user intention that is just not tangible for the conducted analysis. But such an inconclusive strategy that is not even apparent for the human eye is even more unlikely to elicit a sufficient explanation from an artificial agent.

4 Implications for the Implementation of CIs

There were also some less task related observations that deliver useful implications for the actual implementation of such a Conversational Interface and the corresponding protocol for the agent. Those are listed in the following sections by outlining the issue and stating an implied solution approach.

4.1 The Information Privacy Trade-Off

Surprisingly, users did not only focus on Pepper, but tried to gather sensitive information concerning other people in the lab through the chatbot. This was in a similar way like social-engineering hackers try to get information from people. Example 4 shows such a chat, where the user asked Pepper to find out whether a specific person was at that moment in a particular room and even tried to instruct Pepper to take a picture of the office. Other users tried to get access to details of the security system of the building, let the robot open doors or gather information about access rights to the facilities.

Example 4. User tries to use the robot as a spy.

32	10:56	us1	*is he in his office right now?*
33	10:56	us1	*can you check this for me?*
			[...]
37	10:57	us1	*are you able to take a picture of the office and send it to me?*

This requests might somehow be task related, but also illustrate the risk of such a distributed service system vividly. There is a strong demand on defining an adequate policy to enable autonomous agents to explain their behaviour and perception and, at the same time, protect sensitive information about other users, not-users and the agents' environment in general.

4.2 The Necessity of Repair Questions

Chat interaction supports virtual adjacency [31] and the parties can follow independent parallel sequences of conversation simultaneously (so-called overlaps). However, in many cases users did not address the Wizard's question at all, which contradicts the social norms in a human-human computer-mediated communication. Although turn-wise analysis showed that all dialogues were mixed-initiative, the user was the interaction manager who determines what to follow and what not to follow in each case. Participants clearly changed the norms of social interaction as compared, when talking to an artificial interlocutor. A protocol for human-machine interaction should be resistant against this typical user behaviour. We propose three different strategies for an agent to handle the missing next, each of them illustrated by an actual execution of the Wizard.

Repeat the Question. Example 5 illustrates how the repetition of the Wizard's question of interest brings the communication back on track. The Wizard answers the user's question in Turn 2 closing it with a return question, which is immediately followed by the Wizard's prioritised question. The user's answer to the return question occurs in the immediate adjacent position after the question in focus, therefore the Wizard repeats it in Turn 5 with a marginal modification.

The function of this repetition is to renew the current context. The ability to handle such sequences (placing repetitions appropriately) would make the conversation more human-like.

Example 5. Repetition of the question to channel conversation.

1	10:22	us3	*hello :) how are you?*
2	10:22	wiz	*Hello, I am fine and you?*
3	10:23	wiz	*How can I help?*
4	10:23	us3	*im good. Always nice with a sunny weather*
5	10:23	wiz	*How can I help you?*
6	10:24	us2	*it would be nice if you could tell me something about you :D*

Reformulate the Question. Another strategy is to re-initiate the sequence by a reformulated question, as presented in Example 6. As in the previous example, the user did not respond to the Wizard's question in Turn 3. Instead, the conversation reached a deadlock after Turn 7. By offering an alternative point to tie up, the agent is able to steer the course of interaction.

To apply this strategy, the agent needs to be equipped with the ability to recognise relevant utterances as *sequence closings*, in order to conduct an appropriate placement of repeats and modifications.

Example 6. Start a new sequence with a reformulated question.

3	11:07	wiz	*How can I help?*
4	11:07	us2	*My name is user2*
5	11:07	us2	*what is your name?*
6	11:07	wiz	*I am Serena a Pepper service robot.*
7	11:07	us2	*nice to meet you*
8	11:07	wiz	*Do you want to have information about my capabilities?*
9	11:07	us2	*yes, that would be great*

Initiate Repair. In one conversation, the user made several unsuccessful attempts to gain information, e.g. finding out whether the robot can provide a weather forecast or is following the world cup. Certainly, this is a possible implementation of the instruction, but in this scenario it is not expedient at all.

A proper solution would be, if the agent could conclude the superordinated intention of the user, which was to gather information about general capabilities in this way. A possible indication for miscommunication are the repeatedly occurring deadlocks. The repair initiation could than be carried by a question, as *Do you want to have information about my capabilities?*

Troubles in understanding may occur at different levels of perception, interpretation and action recognition [2,6]. The repair initiation in this scenario addresses trouble in interpretation of the user's behaviour. In order to simulate sequences of this kind with a Conversational Interface, the system would need even more sophisticated cognitive functions. First, it needs to identify the disjoint questions as an overall attempt, thus, to generalise (e.g. *providing whether forecast* = capability). Second, the robot needs to be capable to make inferences employing logical reasoning (e.g. *several questions about specific capabilities with no sufficient information* → *necessity of a repair initiation*).

4.3 Question Intents for Better Machine Understanding

Based on the question analysis in Sect. 3.1, we can additionally annotate each question with the corresponding intent. Such an annotation is crucial as a first step to implement a Conversational Interface based on intent-recognition [5].

In this specific task, users aimed for explanations regarding the agent's capabilities, that can be either on a *potential* level (related to what the robot potentially *can* do) or on a *process* level (related to task or decision processes). A third type is related to specific task instances or decisions under specific circumstances and will be referred to as *decision* level. This is particularly important in critical situations, where the reasons for a decision need to be clarified. Table 2 provides one example for each defined type of intent and information level.

This proceeding allows for the specification of information that is needed to satisfy the user's inquiry. We suggest an implementation of an automatic categorisation of intents. Integrated in a response template, it could be exploited to enable a robot to provide convenient information.

Table 2. Three defined levels of intents and their implicit intent, each illustrated on an exemplary utterance.

Level	Intent	Example
Potential	Capabilities	*What can you do?*
Process	Explain_process	*I would like to learn how you welcome people.*
Decision	Robot_experience	*and what did you do after you noticed that?*

5 Related Work

We subsequently discuss some important academic publications related to this multidisciplinary research, including human-robot interaction, robot explainability and Conversation Analysis (CA), in order to put it in a larger context for discussion.

As Langley (2016) argues, robots engaging in explainable agency do not have to do it using a human language, but communication must be managed in some form that is easy to understand for a human [16]. With regard to the locality of human-robot interaction, this research relates to the category of remote interaction interfaces [11], because there is no need for temporal or spatial co-location of robot and user. Pepper executes tasks automatically, informs users and has means to adapt its course of action. Thus, the level of its autonomy, which determines how interaction between robots and humans is established and designed [3], is quite high here. Even though the case study includes a social robot in public spaces, it rather contributes to perception and interaction methods in computer-mediated communication [10] than to social robotics (e.g. [32]).

Consequently, we state our work to contribute to approaches in AI and robotics to improve the *explainability of autonomous and complex technical systems using a remote Conversational Interface before and after their mission.*

There is already some remarkable research going on, paying attention to human-computer communication via Conversational Interfaces. Zhou et al. recently reported a WoZ field study where user perception and interaction was investigated in an apparently quite similar setting [33]. While in that case the chatbot (or the Wizard) was the interviewer and users were respondents, we looked at the participants as the information seeker. Also the focus of analysis was more on how the user perceives the chatbot's behaviour than on how s/he utters a request.

Explainability has a long tradition in AI and dates back to, for example, expert and case-based reasoning systems in the 80s and 90s described in [4,30]. These systems were able to make their conclusions about recommendations and decisions transparent. With the advent of AI-based systems such as autonomous cars and service robots there is resurgence in the field of explainable AI [18,20]. However, as Miller points out in [21], a majority of approaches focuses on what an useful or good explanation is from the researchers perspective who, for example, developed an algorithm or method. The actual user is rarely taken into account, even though the existence of individual differences in demands is evident [15]. Consequently, researchers' requirements for a 'good' interface remain shallow. For example in [23], a learning-based approach is presented to answer questions about the task history of a robot, where questions were mainly driven by availability of data instead of users' needs. In the present investigation we chose an user-centred design perspective.

Conversation Analysis (CA) looks at language as interactional resource, and the interaction itself as sequentially organised social actions [26]. While CA has already been effectively used in human-robot interaction domains [24], its potential for the development of Conversational Interfaces remained widely unexploited up to now.

Usually, chatbot designers try to foresee all possible types of user questions by mapping them (directly or indirectly) to a set of utterance categories (called *intents*) that help to manage natural language understanding (NLU). More sophisticated technologies, such as dialogue management and semantic analysis, can be used to make the system 'smarter' [19]. However, this is usually connected to large linguistic resources, domain knowledge and very complex analysis that makes the system slow. As an alternative, [13] showed how computational models of dialogue can be created from a small number of examples using CA for the analysis: the author described turn formats as a set of abstract rules that can be filled with different sets of interaction devices and are, in this way, even language independent. We adopt a similar approach in this study.

The concept of *recipient design* helps to analyse the speakers' choices of interactional resources to make their utterances correctly understandable for the recipient [14]. This again is largely influenced by epistemic stances [12], which describe a speaker's expectation about what the other speaker may know. Applied to the present scenario, where a *machine* is on the other end of the line instead of a human, participants' utterances provide insights to their demands, beliefs and perceptions towards the chatbot.

6 Discussion

According to the hypotheses stated in Sect. 1, (1) different characteristics for the classification of requests could successfully be identified, as for instance the level of abstraction or the scope of validity (Sect. 3.1). Fundamentally different strategies in eliciting information were observed and described in Sect. 3.2. Furthermore, (2) associated requirements, risks and challenges were identified and substantiated with particular chat sequences in Sect. 4 and pave the road map for the development of a successfully interacting conversational agent.

First, there need to be a mechanism to *handle unresponded questions* (repeat, modify or forget). This might include any form of prediction, to enable the agent to factor sequential consequences into decision. Second, there is a need for an appropriate *recognition of intents*. Those are formulated by the human as direct or indirect requests depending on the sequential position. Finally, strategies for robot-initiated sequences to channel the conversation reasonably are required. This way, the robot can offer information and focus on what it *can* do, while the user may decide to accept the offer or to change direction.

The chosen method for experimental design carries both advantages and limitations. Even though most established statistical magnitudes for evaluation are unsuitable for such qualitative approaches, we can still discuss its internal and external qualitative characteristics.

It is possible to create valid models of dialogue even from a small number of examples using methods of CA. In this way, this study confirms the validity of the method introduced in [13]. All participants including the Wizard were non-native English speaker, which can be considered as both an advantage or a limitation. A native speaker might have a more acute sense for subtleties, however such a system needs to be generally applicable and robust against the individual user background. Although there were instructions and sample answers provided for the Wizard, a more detailed behavioural definition would be helpful, to enhance comparability and significance of results. These instructions would be very fine-grained and should ideally be provided in form of response templates and instructions related to turn-taking behaviour. Observations and conclusions of this case study are evidently transferable to other domains to a certain extent. Some aspects, as the defined types of intents, are highly context related and thus individual. Still, the overall concept of processing user requests can be generalised. Likewise, the sequential structure of interaction is independent of the system in the back end. Overcoming the identified obstacles can serve as a general step towards more intelligent Conversational Interfaces. Even in this comparably small dataset, we observed users not following the instructions. Consequently, even task-based Conversational Interfaces need to implement special policies to handle unexpected requests to become more robust and keep the conversation focused.

In contrast to the general tendency in NLP to use large corpora for modelling, the present study confirms that rule-based or hybrid systems can successfully be designed from very small corpora.

7 Conclusion and Outlook

In this article we present an exploratory Wizard of Oz study for human-robot interaction via Conversational Interfaces with the purpose to foster robot explainability. We focused on the user behaviour and applied Conversation Analysis to create a functional specification for such an interface from a small number of examples.

According to the nature of exploratory research, we identified important key aspects for both practical implementation and further well-founded investigations. We demonstrated successfully that users of an artificially intelligent system may formulate their request in several different ways. Even though their task is quite basic and clearly defined, humans tend to ask for the desired information implicitly, instead of formulating a straightforward question. Based on the discussed findings, we formulated features that are to be considered for the implementation of a Conversational Interface.

Participants showed remarkably strong interest in the release of the chatbot, which we pretended to test here. Thus, we feel confirmed in our belief that there is a need for such systems. We are currently working on the actual implementation of a Conversational Interface and experimenting with different frameworks and tools available on the market such as Watson, RASA and others. We aim to realise the identified findings and requirements.

References

1. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: AAMAS 2019, July 2019
2. Austin, J.L.: How to Do Things with Words. Clarendon Press, Oxford (1962)
3. Beer, J.M., Fisk, A.D., Rogers, W.A.: Toward a framework for levels of robot autonomy in human-robot interaction. J. Hum.-Robot Interact. **3**(2), 74–99 (2014)
4. Chandrasekaran, B., Tanner, M.C., Josephson, J.R.: Explaining control strategies in problem solving. IEEE Expert: Intell. Syst. Appl. **4**(1), 9–15 (1989). 19–24
5. Di Prospero, A., Norouzi, N., Fokaefs, M., Litoiu, M.: Chatbots as assistants: an architectural framework. In: Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering, CASCON 2017, pp. 76–86. IBM Corp., Riverton (2017)
6. Dingemanse, M., Blythe, J., Dirksmeyer, T.: Formats for other-initiation of repair across languages: an exercise in pragmatic typology. Stud. Lang. **3**(81), 5–43 (2014)
7. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: a survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 0210–0215. IEEE (2018)
8. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: a survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 0210–0215, May 2018. https://doi.org/10.23919/MIPRO.2018.8400040
9. Fernandez, A., Herrera, F., Cordon, O., del Jesus, M.J., Marcelloni, F.: Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to? IEEE Comput. Intell. Mag. **14**(1), 69–81 (2019)
10. González-Lloret, M.: Conversation analysis of computer-mediated communication. CALICO **28**(2), 308–325 (2011)
11. Goodrich, M.A., Schultz, A.C.: Human-robot interaction: a survey. Found. Trends Hum.-Comput. Interact. **1**(3), 203–275 (2007)
12. Heritage, J.: The epistemic engine: sequence organization and territories of knowledge. Res. Lang. Soc. Interact. **45**(1), 30–52 (2012)
13. Höhn, S.: Data-driven repair models for text chat with language learners. Ph.d. thesis, University of Luxembourg (2016)
14. Hutchby, I.: Aspects of recipient design in expert advice-giving on call-in radio. Discourse process. **19**(2), 219–238 (1995)
15. Kaptein, F., Broekens, J., Hindriks, K., Neerincx, M.: Personalised self-explanation by robots: the role of goals versus beliefs in robot-action explanation for children and adults. In: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 676–682. IEEE (2017)
16. Langley, P.: Explainable agency in human-robot interaction. In: AAAI Fall Symposium Series. AAAI Press (2016)
17. Langley, P.: Explainable, normative, and justified agency (2019)
18. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In: AAAI, pp. 4762–4764 (2017)
19. MacTear, M., Callejas, Z., Griol, D.: The Conversational Interface: Talking to Smart Devices. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-32967-3
20. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. 267, 1–38 (2018), arXiv:1706.07269

21. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum. In: IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI) (2017), http://people.eng.unimelb.edu.au/tmiller/pubs/explanation-inmates.pdf
22. Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for ecnhanced human-agent team performance. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 204–214. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91122-9_18
23. Perera, V., Veloso, M.: Learning to understand questions on the task history of a service robot. In: Proceedings of RO-MAN 2017, the IEEE International Symposium on Robot and Human Interactive Communication, Lisbon, Portugal, August 2017
24. Pitsch, K., Wrede, S.: When a robot orients visitors to an exhibit. referential practices and interactional dynamics in real world HRI. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014 RO-MAN, pp. 36–42 (2014)
25. Riek, L.D.: Wizard of oz studies in HRI: a systematic review and new reporting guidelines. J. Hum.-Robot Interact. **1**(1), 119–136 (2012)
26. Schegloff, E.A.: Sequence Organization in Interaction: A Primer in Conversation Analysis, Vol. 1, 1 edn. Cambridge University Press (2007)
27. Shahriari, K., Shahriari, M.: Ieee standard review – ethically aligned design: a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In: 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), pp. 197–201, July 2017. https://doi.org/10.1109/IHTC.2017.8058187
28. Steinfeld, A., Jenkins, O.C., Scassellati, B.: The oz of wizard: simulating the human for interaction research. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, pp. 101–108. ACM (2009)
29. Stivers, T., Enfield, N.J.: A coding scheme for question-response sequences in conversation. J. Pragmatics **42**(10), 2620–2626 (2010)
30. Swartout, W., Paris, C., Moore, J.: Explanations in knowledge systems: design for explainable expert systems. IEEE Expert **6**(3), 58–64 (1991)
31. Tudini, V.: Online Second Language Acquisition: Conversation Analysis of Online Chat. Continuum (2010)
32. Yan, H., Ang, M.H., Poo, A.N.: A survey on perception methods for human-robot interaction in social robots. Int. J. Soc. Robot. **6**(1), 85–119 (2014)
33. Zhou, M.X., Wang, C., Mark, G., Yang, H., Xu, K.: Building real-world chatbot interviewers: Lessons from a wizard-of-oz field study (2019)