Exploring Interaction with Remote Autonomous Systems using Conversational Agents

David A. Robb[1], José Lopes[1], Stefano Padilla[1], Atanas Laskov[2], Francisco J. Chiyah Garcia[1], Xingkun Liu[1], Jonatan Scharff Willners[1,2], Nicolas Valeyrie[1], Katrin Lohan[1], David Lane[1], Pedro Patron[2], Yvan Petillot[1], Mike J. Chantler[1], and Helen Hastie[1]

[1]Heriot-Watt University, Edinburgh, UK
[2]Seebyte Ltd., Edinburgh, UK
Corresponding author: d.a.robb@hw.ac.uk

ABSTRACT

Autonomous vehicles and robots are increasingly being deployed to remote, dangerous environments in the energy sector, search and rescue and the military. As a result, there is a need for humans to interact with these robots to monitor their tasks, such as inspecting and repairing offshore wind-turbines. Conversational Agents can improve situation awareness and transparency, while being a hands-free medium to communicate key information quickly and succinctly. As part of our user-centered design of such systems, we conducted an in-depth immersive qualitative study of twelve marine research scientists and engineers, interacting with a prototype Conversational Agent. Our results expose insights into the appropriate content and style for the natural language interaction and, from this study, we derive nine design recommendations to inform future Conversational Agent design for remote autonomous systems.

Author Keywords

Natural Language Interfaces; Remote Autonomous Systems; Explainable AI; Multimodal Interfaces; Trust; Transparency.

CCS Concepts

•**Human-centered computing** → **Natural language interfaces;** *Graphical user interfaces;* •**Computer systems organization** → *Robotic autonomy;*

INTRODUCTION

Remote autonomous systems and vehicles are increasingly being used to facilitate operations where it is either impossible or dangerous for humans to go. These systems may be deployed, for example, in the energy sector, search and rescue or military settings [31, 47, 63, 71, 79]. It is expected that, increasingly, these systems will have high levels of autonomy and operate in teams requiring only supervision by human operators [5, 22].

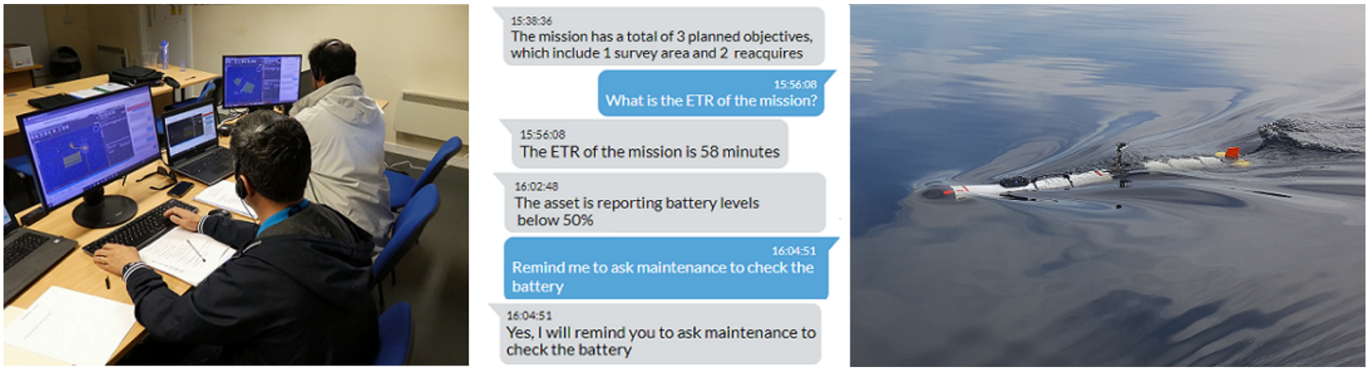
In order for these human-supervised teams to function smoothly, it is essential that there is clear communication and that the human operators maintain high situation awareness. To allow this, robots and autonomous systems need to communicate their world view, system actions and reasoning, developing appropriate levels of trust within those supervising them and enabling sound decision making. The human operators could in future find themselves responsible for multiple, multi-million dollar, pieces of hardware and face high stress situations due to a) the complex goals and objectives which comprise a mission, b) dynamic changes in the environment and operating conditions, and c) occurrence of unexpected but valid autonomous behavior. Unless operators have complete clarity and confidence in their grasp of the situation, these pressures can lead to unnecessarily conservative decisions such as opting to abandon objectives or even entire missions, incurring high costs in rescheduling [20].

As systems gain higher levels of autonomy, transparency will become more important, not only in terms of providing an audit trail but also in terms of human-robot collaboration moving forward. Transparency has been shown to improve understanding of the inner workings of an autonomous system [70, 80] and can also facilitate adoption of technology and help in operator training. Specifically, transparency can improve the user's Mental Model of the system [28] in terms of the system's *functionality*, i.e. understanding what the system can do and *structure*, i.e. understanding how it works. Explainability is one facet of transparency and can help users develop realistic expectations and interact appropriately with the system [81]. Such explanations can be verbal [38, 49], graphical [33] or multimodal [43] in nature.

The user interfaces for remote autonomous systems often include graphical map-like displays for locating mobile systems or robots within their surroundings along with status information to help operators build situation awareness [6, 8]. We believe that incorporating a Conversational Agent (CA) into the interaction with remote autonomous systems, alongside a graphical display, can improve transparency through a fluid natural interaction mode enabling the provision of clear information and explanations. Indeed, recent work in the domain of Autonomous Underwater Vehicles (AUVs) has shown that

DIS '19, June 23–28, 2019, San Diego, CA, USA.

ACM ISBN 978-1-4503-5850-7/19/06.
http://dx.doi.org/10.1145/3322276.3322318



15:38:36
The mission has a total of 3 planned objectives, which include 1 survey area and 2 reacquires

15:56:08
What is the ETR of the mission?

15:56:08
The ETR of the mission is 58 minutes

16:02:48
The asset is reporting battery levels below 50%

16:04:51
Remind me to ask maintenance to check the battery

16:04:51
Yes, I will remind you to ask maintenance to check the battery

Figure 1. Left: Participants interacting with the Conversational Agent (CA) for remote autonomy combined with the SeeTrack graphical interface during our study. Natural language provides a fluid and reassuring mode of interaction with complex multi-million dollar remote systems. Right: The IVER3 autonomous underwater vehicle (AUV), under way before diving, is capable of carrying out submarine survey tasks with a high degree of autonomy (see also Figure 4). It can be monitored and queried during its sub-sea missions using the Conversational Agent described in this paper. Participants launched and retrieved one of these $200,000 AUVs in open water during the study. The center panel shows an example CA interaction.

combining a CA along with a graphical display can improve the situation awareness in operators [59].

Interactive dialog between user and robot is considered to be a characteristic of a transparent system [46, 66]. Incorporating a CA should thus improve accessibility of these systems to users who are not immediately familiar with the operation of particular robots or teams of robots. Additionally, the need for an intelligent interface agent to monitor robots and support human operators when workload, environment and robot capabilities change has already been foreseen [15].

To our knowledge however, there has not to date been a study with the purpose of exploring CA interaction for remote autonomous systems. In this paper, therefore, to enable that exploration, we have chosen the AUV domain as the context for our qualitative investigation of a CA for remote autonomous systems. This domain is ideally suited for the study of interaction with highly autonomous vehicles as the constraints of low bandwidth communications have already motivated the development of systems that facilitate the cooperation of multiple highly autonomous AUVs [44, 53].

We developed a prototype CA with suitable capabilities, including explanations of autonomous behavior, for interaction with dynamic autonomous AUV missions. We used this in our study of a group of marine research scientists and engineers at an autonomous underwater robotics summer school. During a program also including underwater computer vision, sensors and remote operated vehicles, 12 participants experienced handling an AUV and interacting with a CA combined with a graphical interface called SeeTrack (Figure 1). Then, through in-depth semi-structured interviews, the participants provided insights into the kind of CA interaction that they think would work for the operators of teams of remote autonomous vehicles. We have used these insights to develop a set of recommendations for designers of CAs for remote autonomous systems.

The contributions of this paper are summarized as:

- Qualitative insights into effective styles of Conversational Agent (CA) interaction for remote autonomous systems derived from 12 marine research scientists and engineers, with expertise in robotics, AUVs and AUV simulations.
- Nine design recommendations for designers of CAs for remote autonomous systems.

In the rest of this paper, we first describe prior work forming the background to our investigation. Then, we describe the CA we used in our study to inspire participants to discuss styles of CA interaction. We describe the study participant group; the activities they undertook; the equipment and applications used; and the methods for gathering and analyzing the qualitative data. Following that, we discuss in detail the insights in the results and present recommendations for designers of CAs for remote autonomous systems. Finally, we draw conclusions and discuss future work.

BACKGROUND

In this section, we describe other prior work forming the background to our investigation not already discussed in the introduction.

Explanations

As discussed in our introduction, explainability is a facet of transparent robots/systems and can improve the Mental Model of the user [14], as well as increase confidence and performance [33, 38]. Lim et al. [38] described two styles of explanation: "*Why*" and "*Why not*", to explain the functionality and the structure of a system, respectively. They showed that explaining *why* a system behaved a certain way increased both understanding and trust, whilst "*Why not*" showed only an increase in understanding. Successful explanation generation depends on the user, their knowledge and the context, as users will only take the time to process the explanation if the benefits are perceived to be worth it [16]. The content of the explanations should, therefore, be adapted to the user. For example, Kulesza et al. [30] showed that novice users would rather have all the details in the explanations (high soundness/fidelity), whereas Garcia et al. [14] studied experts, who presumably could fill in the gaps and preferred more broad level explanations (low soundness/fidelity). In both studies, being given all the possible explanations was important to the user. As explanations are key to transparency, the CA used in the study described in this paper was equipped with an explanation system, which allowed the levels of detail and amount of explanations to be varied.

Conversational Agents and their Domains

ELIZA [77], one of the first CAs ever developed, aimed at having open conversations with its users. ELIZA's goal was to maintain a coherent conversation. Its range of conversations was limited by it being heavily based on rules. Research then turned to simpler tasks that could easily be modeled with fewer rules and, later, machine learning. A few examples of the domains chosen to implement Conversational Agents were travel planning [74], weather forecasting [83], public transit [13, 55, 67], flight schedule [54, 62], real-estate [18] and restaurant and bar information [29, 61]. These CAs are often named in the literature as 'slot-filling' denoting the finite number of information slots, which need to be filled in order to perform a database query. These CAs were initially implemented as finite-state machines. Later, these were replaced by machine learning methods (e.g. [78]) making implementation less cumbersome, more robust to speech recognition errors, less dependent on the developer's knowledge about the domain and easier to export to new domains.

With the advent of deep machine learning, researchers have worked on creating open-domain CAs that can have social conversations i.e. 'chitchat', using techniques such as neural models [58, 72] and crowdsourcing [3, 24, 34]. However, their dialogs are not always coherent. In addition, their lack of ability to interact in dynamically changing contexts limits their use in certain domains, such as the one described here.

On the other hand CAs with situation awareness in dynamic contexts have been built for in-car navigation. The vehicle location and non-verbal features were used to train the machine learning model for a navigation CA [45]. This scenario shares some similarities with the one presented in this paper, since in both cases the CAs require a perception of the environment that is constantly changing over time (unlike slot-filling CAs). The WITAS [35] framework approached this problem by focusing on control and waypoint specification. Another framework that deals with situation awareness is TRIPS [12]. TRIPS is a combination between a CA (also known as dialog system) and "Specialized Reasoners" that can solve problems such as planning actions, scheduling events or simulating future plans. In addition, just as in our case, both TRIPS and WITAS can be combined with a graphical interface where the environment is represented.

Design Principles for Conversational Agents

Design principles should provide guidelines to improve and maintain the quality of CAs. How to measure the quality of task-orientated dialog systems, as discussed above, has been much researched (see [19] for an overview). Evaluation of non-task oriented social dialog systems, on the other hand, is a new emerging challenge, as there is no clear measure for task success and evaluating whether *rapport* has been established is far from clear-cut [9, 39].

One task-orientated evaluation framework is PARADISE [73], which is founded on decision theory and posits that usability (usually in the form of subjective user satisfaction) can be broken down into two contributing quality criteria of task success (e.g. restaurant booking) and dialog costs (e.g. time on task, dialog length). This framework attempts to capture the multi-dimensionality and complexity of dialog through multi-linear regression analysis (MLR), which can give some insights into the factors contributing to high and low user satisfaction. Deriving design principles from MLR is possible by examining the variable weights. As such, optimization functions for adaptive systems have been derived from linear regression analysis [57]. Other evaluation frameworks, such as those based on hidden Markov models [11], are harder to interpret and form into general design principles.

With slot filling systems, the community has adopted design principles based on minimizing dialog length as was shown in the DARPA communicator evaluation using PARADISE [74]. For social dialog, design is focused on lengthening dialogs to reflect user engagement [52]. For example, in the criteria of the Amazon Alexa Challenges students can win USD1M if their system can engage the user for 20 minutes. For interaction with remote autonomous systems, it is unclear what metrics to optimize for. For example, does a long dialog mean that the user is fully immersed and reflect high situation awareness? Or does it imply inefficiencies and the CA should try to minimize dialog length, especially in emergency response situations?

One recent issue is transparency and it has emerged in several interviews reported by Jain et al. [26] that CAs are often not transparent to their users, meaning that users are not aware of their capabilities as they interact with the agent for the first time. Most users seem to dislike the trial procedure that they usually perform to become aware of the system capabilities [25]. The suggested design principle is that the CA should be able to create the awareness of its capabilities both at the beginning and during the interaction. However, doing so in a natural and appealing manner is non-trivial.

Some obvious design principles involve minimizing errors and misunderstandings [74]. In addition, failing gracefully is important in terms of understanding when the CA is in trouble [82] and providing mitigating strategies to minimize user frustration, such as asking for clarification. In multimodal systems, there is the opportunity to provide visual (graphic) information and selection menus to the user [25] but this may come at a cost in terms of user-initiative interaction [68], where the user should be able to drive the interaction in a free flowing manner.

Unlike most of the CAs used in the above-mentioned studies, in our study the CA aims to be an instrument, which is assistive to the operator of a remote autonomous system. Therefore, our contribution is in the design principles from this relatively new domain.

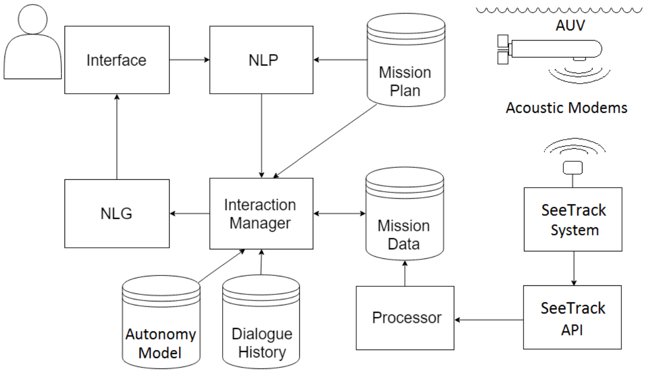


Figure 2. System architecture of the MIRIAM CA. NLP/G are Natural Language Processing/Generation. The architecture is explained in the text.

THE MIRIAM CONVERSATIONAL AGENT USED FOR THE STUDY

A Conversational Agent (CA) was developed known as Multimodal Intelligent inteRactIon for Autonomous systeMs (MIRIAM). MIRIAM a) closely integrates with an AUV software system that provides a particularly high level of autonomy such as is needed for the teaming of multiple remote autonomous vehicles and b) uses a mixed initiative approach in both generating its own notifications about important changes in status and also allowing users to query specific information.

Figure 2 shows the system architecture, which we describe here: In the context of AUVs, such vehicles undertake objectives such as surveying a patch of seabed and several of these objectives together comprise a Mission Plan. The CA uses a rule-based Natural Language Processing (NLP) engine that contextualizes and parses the user's input for intent, formalizing it as a semantic representation. The Interaction Manager can process both static and dynamic data, such as vehicle and objective names. It uses Natural Language Generation (NLG) to present output to the user. It uses the Dialogue History and Mission Data for context. It is able to give explanations of certain autonomous behaviors based on an interpretable Autonomy Model described by Garcia et al. [14]. This allows it to describe why a given behavior is occurring based on the current mission status and history and the possible reasons. It can then provide a list of reasons along with an estimate of the likelihood that a given reason explains the current behavior. The Processor gathers the Mission Data from the SeeTrack System through its Application Programming Interface (API). The SeeTrack[1] commercial AUV software system (comprising a graphic user interface and an autonomy system) communicates with the autonomous vehicles, in this case an AUV. SeeTrack is described later in the "Activities, Equipment and Applications" subsection in "STUDY" below.

STUDY

The aim of the study was to explore Conversational Agent interaction for remote autonomous systems. In this section, we first describe the participants and the setting in which the study took place. We then detail the activities undertaken and the equipment and applications used by the participants. Finally, we describe the interview method used to gather qualitative data from the participants and the analysis of that data.

Participants and Setting

We chose to conduct our study at a robotics, sensors and remote underwater vehicle summer school as a) the attendees would have experience in robotics and engineering and b) it would provide an immersive environment for the study of autonomous systems interaction. It took place at a specialist residential underwater training facility in Scotland on the shores of a deep water sea inlet or *sea loch*. Ethical approval was obtained from our institution.

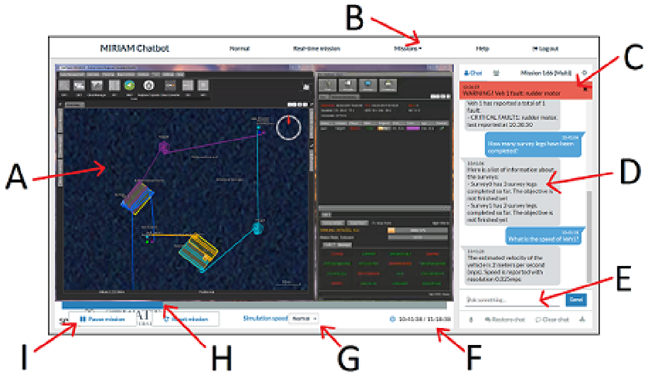
After giving their informed consent to take part, the participants completed a short demographic questionnaire, providing details of their occupation, education level, areas of expertise and experience with AUVs and AUV simulations. There were twelve (ten male, two female, similar to current gender proportions of UK engineering and technology sector employees, 9% female [69]). They aged from 24 to 39 (M 31.4, Med 31, SD 4.9). All were non-native English speakers with English as a Foreign Language skill as a minimum. They all reported their occupation as researcher with five indicating a level of seniority such as assistant professor or senior researcher. Eleven reported from 1 to 15 years expertise in robotics (M 5.3 yrs, Med 5 yrs, SD 4.3 yrs). They had, between them, expertise in other areas such as sensors, unmanned airborne vehicles (UAV), embedded systems, marine biology and computer vision. They were asked to rate their expertise with AUVs using a 5-point Likert type item, with opposing semantic anchors, ranging from 1- "Novice, I know nothing about them" to 5 - "Expert, I have a deep understanding of them". One participant reported level 2, nine level 3 and two level 4. None reported levels 1 or 5, i.e. all had some AUV expertise. When describing their knowledge of the SeeTrack system for mission planning, three stated they had limited knowledge while nine had none. When asked if they had previously worked with AUV simulations, six stated that they had and six had not. Therefore, we describe our participant group as consisting of 12 marine researchers and engineers with expertise in robotics, AUVs and AUV simulations but with little or no experience of the SeeTrack software. This leads us to expect that, while they would be approaching the combined SeeTrack and CA as a new interface, their opinions on the AUV interaction they would experience in the study would be well-informed.

Activities, Equipment and Applications

In this subsection, after briefly listing the four main blocks of activities that summer school attendees undertook, we describe in detail what our study participants did and the equipment and applications they used during the **AUV activity** directly connected with the CA study.

Participants experienced four blocks of summer school activities: 1) remote underwater vehicle pilot simulator training, 2) underwater vision and mapping, 3) sensor networks, and **4) autonomous underwater vehicles (AUVs)**. Participants

[1] Made by Seebyte Ltd., the SeeTrack user interface and mission planning system interfaces with Seebyte's Neptune autonomy system. For simplicity, we refer to both collectively as SeeTrack.



WARNING! Veh 1 fault: rudder motor

Veh 1 has reported a total of 1 fault:
- CRITICAL FAULT 1: rudder motor, last reported at 10:20:30

How many survey legs have been completed?

Here is a list of information about the surveys:
- Survey0 has 3 survey legs completed so far. The objective is not finished yet
- Survey1 has 2 survey legs completed so far. The objective is not finished yet

What is the speed of Veh1?

The estimated velocity of the vehicle is 2 meters per second (mps). Speed is reported with resolution 0.025mps.

Figure 3. The CA web application. A) Synchronized streamed video capture of mission in SeeTrack showing vehicle traces, predicted tracks, objectives not yet completed, and a table of vehicle status; B) Drop-down menu of missions; C) Critical alerts are pinned, e.g. ***"WARNING! Veh1 fault: rudder motor"*****; D) Scrolling chat history panel; E) User enters query here; F) Mission clock e.g.** ***10:41:38/11:18:38*** **indicating there are 37 minutes of mission time remaining; G) Mission speed-up drop-down menu - choose from 1x to 8x speed; H) Fast Forward - click and drag slider; I) Play/Pause mission.**

would experience the CA within this fourth activity (the AUV activity). Day 1 consisted of lectures on the different blocks' topics. At the end of the AUV lecture, focusing on mission planning, the CA study was described and the 12 participants signed appropriate consents to take part. On the second and third days, participants undertook practical activities associated with the four blocks. The participants experienced each block of activities in a group of four individuals and, where appropriate, rotated through subsidiary activities to experience everything within each block.

The AUV activity lasted three and a half hours excluding a 30 minute break and involved **two parallel activities:** 1) hands-on launching and retrieving an AUV, and 2) a classroom session where they would experience AUV mission planning and monitoring software and the CA. These two activities ran in parallel with two participants each and then they swapped round after a break. The AUV activity was divided this way to allow an appropriate ratio of safety trained staff to participants when afloat handling the AUV.

The AUV mission planning and monitoring software, which we used is SeeTrack (shown within Figure 3). It is commercial software and combines a chart and tabular user interface (UI) allowing complex mission planning on a PC using an autonomy system, which also runs on the AUVs' embedded computers. The AUVs communicate with the UI via networks. The networks can be wifi or acoustic for real vehicles or wired for simulated vehicles. While the autonomy system has been designed to allow missions consisting of sub-sea, surface and air vehicles in the marine domain, the context of our study was sub-sea missions involving sonar survey and object reaquisition objectives. The SeeTrack software provides an API allowing the CA system access to the mission and vehicle data (Figure 2). From this it builds a mission database, which grows as the mission progresses and on which the CA can run queries in response to user interaction, generating alerts and notifications in response to mission and vehicle events.

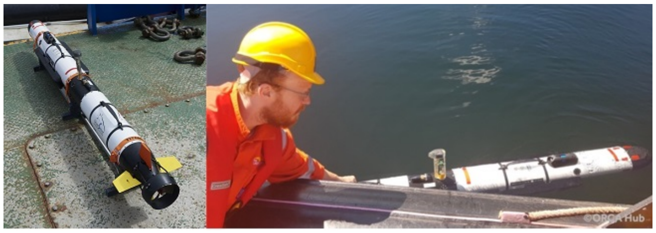


Figure 4. Left, the IVER3 AUV on the quayside and right, just after being placed in the water. (See also Figure 1, right).

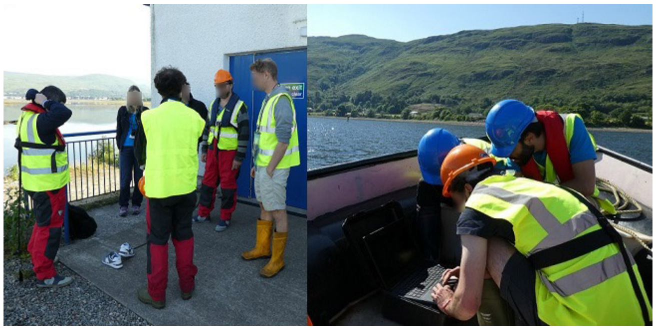


Figure 5. Left: Four participants receiving a safety briefing before two undertake launching and retrieval of an AUV. The other two, first, would do the classroom activity and later launch the AUV. Right: Participants monitor the AUV on a laptop in the boat using the SeeTrack software.

The two parallel AUV activities were designed to give the participants experience of: A) handling an AUV capable of highly autonomous behavior (Figure 4), and B) first planning and executing missions for a team of two AUVs using the SeeTrack system and simulated AUVs, and then interacting with the combined SeeTrack and CA system in a number of single and multiple vehicle mission scenarios.

AUV Activity A: This allowed participants to experience handling, launching and recovering an IVER3 AUV. The OceanServer IVER3 is designed for sonar survey and object acquisition. It has endurance of 8-14 hours and speed of 1 to 4 knots. Its main sensor is high resolution side-scan sonar. It has wireless and acoustic communications and GPS and Doppler velocity log (DVL) location. It is compatible and can operate autonomously with the SeeTrack UI and autonomy system. A typical mission for a single IVER3 might comprise surveying areas of the seabed with its sonar, visiting specific locations and then using sonar to reacquire (i.e. revisit) a previously located object. For our participants to experience a complete launch and recovery of the AUV, a short simple pre-planned mission consisting of a single small survey area and launch and recovery points was used. Participants were given a safety briefing and geared-up in personal protective equipment (Figure 5, left). They participated in loading and launching the AUV from a small boat and monitored what the AUV was doing from a laptop running SeeTrack in the boat (Figure 5, right).

AUV Activity B: During the classroom session two participants at a time were shown how to plan a mission with a researcher demonstrating before taking turns to create and edit a survey objective. The mission was then executed in real time on a simulator, simulating two AUVs running the autonomy system. The participants watched the early stages of the mission unfold including observing the division of the objectives between the two simulated AUVs, an important aspect of the cooperative autonomy that the SeeTrack system enables [44, 32]. At the mission planning stage, operators can suggest which AUV in a team does which objectives but once the mission starts, the autonomy system takes over allowing each AUV to undertake objectives in an order that the system calculates to be optimal, given the current conditions such as prevailing currents and vehicle availability.

After seeing their planned mission unfold in its early stages, participants were introduced to the CA, which was added as a web browser window panel beside the SeeTrack UI display. The participants could use the CA to query the system about what each vehicle was doing and the progress of the mission that they themselves had planned.

The first option would be for the participants to experience the CA during a real mission. However, these missions can last for a number of hours with activity spread thinly throughout. Therefore, in order for the subjects to experience all the aspects of the CA within a shorter time period, we created a special purpose interface for simulated missions, which we will refer to as the CA web application (see Figure 3). To create this application, simulated missions were run and the SeeTrack UI display was video captured while the back-end of the CA system captured the mission and vehicle data in real time and stored it in a database for use later. When a particular mission is selected from the web application menu that mission's video is streamed while allowing chat through the CA to have access to the timestamped data synchronized with mission progress. It allows mission time to be sped up in increments e.g. x4 or x8, fast forwarding through uneventful periods in a mission or directly accessing particular points in a mission to experience what CA interaction is like during a particular type of mission event.

While using the simulations in the CA web application, participants did three sub-activities involving text chat with the CA. 1) a short quiz based on a specific mission asking questions at specific timed points, which could be answered using the CA, 2) a directed activity (a set of steps on paper) to show the CA's explanation capability and how the amount of detail in explanations can be changed using the CA's settings menu and 3) free exploration of the different scenarios and interaction using the CA. During these 3 sub-activities, participants were able to refer to a 3-page printed table showing what information the CA can be asked about. While participants used keyboards for text input in the text chat, they also wore headphones to listen to speech output (Figure 1, left-hand side).

Below is an example interaction from the quiz sub-activity. It shows mission time, which was sped up x8. Here, ETR refers to Estimated Time of Retrieval and Target0 is a re-acquisition objective for the AUV to visit and search:

```
USER:'How many row scans has the vehicle done'

SYSTEM:'Survey1 has 4 survey legs completed so far.
The objective is not finished yet'

SYSTEM:'The vehicle is reporting a change in the
mission ETR of more than 5 minutes.  The vehicle now
estimates it will finish the mission by 15:45:24'

SYSTEM:'The vehicle is in transit to Target0'

USER:'what is the total mission progress'

SYSTEM:'The vehicle is reporting that it has
completed 48.48% of the mission'
```

Method

After the participants had completed the activities detailed above, they were interviewed either the same day or the following day [23]. For consistency, there was one interviewer, an author. We used semi-structured interviews [7, 65], the method used, for example, by Le Bras et al. to develop a set of design recommendations for automated concept map layout and explanation systems [33]. This method allows particular topics of interest to be explored while also enabling the pursuit of interesting participant views as they arise. To stimulate and focus discussion, the interviewer used the following resources: a picture of the IVER3 AUV, printouts of screen shots of the interfaces they had used, the printed materials to which they had referred during the classroom activity, and a laptop running the CA web application to allow display of the CA's output during an interaction involving an explanation of AUV behavior. Participants were paid US$20 as compensation for their time in interview.

While developing the semi-structured interview script, initial topics for questions were drawn from a prior mixed-methods study of the effect on operator situation awareness of a CA [59]. That study reported a brief analysis of the comments of a group of participants who were experts in the SeeTrack software, gathered in a post-task questionnaire. The topics raised in those comments included usefulness for less experienced operators, number of notifications and succinctness and a wish for added multimodality. In addition, prior work on explanation styles [30, 38] and, more recently, Garcia et al.'s quantitative study, also using experts in SeeTrack and the AUV domain [14], prompted us to further explore CA explanation styles qualitatively with our group of more realistic potential users.

Each interview started with warm up questions asking participants to talk about their previous experience with robots and AUVs using their answers to the demographic questionnaire as a starting point. They were asked about using the IVER3 AUV in particular and how they felt when operating and being responsible for AUVs. Further questions included: asking their views on the frequency and length of alerts, and the amount of information they contained; the format of explanations and their level of detail; the CA's ability to set reminders and whether and how this should be exploited; managing missions involving multiple vehicles (experienced by participants in simulation during the study) and over multiple domains i.e. sub-sea, surface and air (this being a theoretical capability of SeeTrack and the CA). The interview setting is shown in Figure 6.

Figure 6. Interview setting. The participant was wearing personal protective equipment having recently completed the last of their AUV activities. This activity included launching and retrieving an AUV from a boat launched from a pier facility.

Coding

Each interview lasted approximately 30 minutes. Audio recordings were made and these were professionally transcribed. The coder (an author) then listened to all the audio a) to complete the transcriptions at the few points where jargon or lack of clarity had defeated the transcriber, and b) to become fully familiar with all the interviews. We followed a thematic qualitative analysis methodology used in other qualitative human interface studies (e.g. [60]). The data from the interviews was categorized using qualitative data analysis software (NVivo10) [64]. A grounded theory (or inductive) approach with open coding was used [7, 65]. The final code book contained 103 codes and 811 coding instances and from these the overarching themes were identified.

The themes are set out below along with discussion of their significance and relation to prior work. Design recommendations are developed in response to the themes.

RESULTS, DISCUSSION AND DESIGN RECOMMENDATIONS

Below we detail and discuss the themes, setting design recommendations with each, using a structure similar to Padilla et al. [51]. We quote from the data to illustrate the themes and readers are reminded that our participants were non-native English speakers.

There were eight themes found in the data:

1. **Users**: Adapting CA behavior for different user roles and preferences.
2. **Style and content of information**: What information to present and how to present it, including relevance filtering.
3. **Presentation to enable focus**: Highlighting facts and differentiating individual agents in multi-agent scenarios (missions with two or more vehicles).
4. **Correct interpretation of user input**: Sensitivity and robustness to conversation context.
5. **Hands-free use**: Desirability of hands-free in some circumstances such as difficult environments.
6. **Multimodality**: Using a CA with a graphic interface.
7. **Reassurance, explanations and trust**: Expensive equipment and complex missions as a source of anxiety. Desirability of explanations.
8. **Post-conversation follow-up**: Summarizing and communicating unresolved items post-conversation.

Theme 1: Users

This theme touched most of the other themes as participants expressed opinions about the various aspects of CAs in the context of their own experience. They might state their own preference but would also add that it depends on the user and proceed to give their view of what users in other roles might find appropriate. This is summed up by participant 9 (P9) when commenting on the amount of information content (itself a subsequent theme): *"Some information is too much. Some may be too low, I don't know. That depends on the end user, so who is inquiring what"[P9].* Our participants suggested, in all, six user roles:-

1. **Trainee operators:** *"[for] an operator [. . .] getting the courses for being an operator [. . .] Training. It could be useful to have this type of information"[P5]*
2. **Operators unfamiliar with the usual control system:** *"MIRIAM is good for people with not so much knowledge about the system. Because you can ask her everything"[P7], "an untrained user[. . .] that user wouldn't recognize that something is a bit off from the traditional interface, right?"[P2].*
3. **Non-technical operators unfamiliar with robots:** *"I think for a not cyber person [it] is a good tool"[P4].*
4. **Researcher\Engineer\Developer :** *"It depends on the personal level. [. . .] if it is a researcher or a developer, you want to see everything."[P12]*
5. **Experienced operators:** *"I think it depends on the user you are working with, or if the person is experienced with the system or not"[P1]*
6. **Commanders or supervisors of complex missions:** *"For example the captain [. . .] in the ship that wants to know how the mission is going "[P4].*

Our participant group clearly thought that a CA should cater for a number of user roles and, hence, modify the interaction in style and information content relative to the user. Indeed, there has been recent work on response generation that can improve the quality of the interaction by simply adding the user ID into the neural model [1] or by integrating complex persona models [36]. Various aspects of user adaptation are discussed further in the themes below.

R1: We recommend that CA designers implement adaptive systems, in which important aspects of interaction are tailored to suit a given user's role, better meeting their needs and achieving a closer match between expectation and execution [48]. Those aspects to be adjusted depending on the user can include, for example, style and content of information and frequency of updates and are addressed further in subsequent themes.

Theme 2: Style and Content of Information

This theme relates to the well-studied challenge of information presentation in natural language generation (NLG) and to many aspects of the CA interaction including explanations of behavior and alerts and notifications. Here, we discuss this theme in terms of the two traditional aspects of NLG, i.e. *what* to present, and *how* to present it [56].

With respect to *how* to present information, P12 wished for output from the CA to be personalized to the user: *"Obviously it's very good to have the updates. I don't know about the phrasing, depending on the subject who's getting it"*. This reflects previous work such as Janarthanam & Lemon [27], who adapt referring expression generation for technical instructions based on the level of expertise of the user. Studies have also looked at adapting wording style to the user by inferring their preferences [10] or by converging to the user's style during the interaction, known as entrainment [4, 40].

With respect to *what* information should be given, P9 pointed out that this would also need to be adjusted depending on the user's role (already quoted in "Users" theme). P9 went on to suggest that user interest about specific parameters could be used to infer what they may wish to receive in future unprompted updates: *"Or it [the CA] could learn, like I ask battery and from like you ask for battery I [it] will tell you battery even though you don't ask"*. Previous work has looked at adapting content to a user model [75], allowing for memory of interaction and preferences across dialogs. This is useful in the case of an operator who has to run multiple missions over time.

Information presentation work in the field of NLG has made frequent use of summaries, e.g. for restaurant recommendations [75]. The need for summaries of mission activities were pointed out here by P2: "*If I remember correctly, MIRIAM would go vehicle by vehicle if it's finished [its tasks] [. . .] maybe there could be some kind of summary of like, "50 per cent of the vehicles have finished their tasks" or something like that.*". Summaries have been generated for reporting of AUV post-mission reports [21, 76], as well as, in-mission as seen here.

Finally, this theme also relates to intelligent alerting, in terms of what information the system should offer up and its relevance to that user at that point in time in a specific context. The CA would need to take into account user preference for information and the interaction history (as illustrated in the above quote from P9) but also the user's immediate surroundings e.g. in a small boat after launching an AUV (see the "Multimodality" and “hands-free” themes) and even their current workload [41].

R2: Information presentation (**how and what to say**) from robot missions should be adjusted based on user role, preferences, interaction history and user context to better match user expectations, as stated in R1.

R3: CAs should include mechanisms allowing information content to be **filtered for relevance** and **summarized** based on a) user role, b) individual user preferences, c) particular mission and environment conditions, and d) dynamically modified through system monitoring of user queries.

R1 to R3 reflect Grice's cooperative principle, particularly the maxims about quantity i.e. providing neither more nor less information than is needed by a conversation partner [17]. While it may have been possible to predict that this would be the case, we have here empirical evidence that users expect from the CA what they would expect from a human.

Theme 3: Presentation to Enable Focus

Also related to Grice's maxims on quantity, P7 pointed out that often the output would be lengthy with the sought for fact or value being provided but embedded within a sentence. *"sometimes the answer is a long sentence. And the information, the most important parts, don't pop up. [. . .] We need like a bold or color thing."*. Here, the desire is to bring emphasis to the specific facts in the CA's output.

P3 saw a need for differentiating between utterances concerning different robots in a multi-robot mission conversation and additionally to be able to focus the conversation temporarily on one specific robot by selecting that robot:*"you have only one, one MIRIAM [. . .] you have a MIRIAM, a Josie, a Peter and [each] corresponds to each vehicle."*, and then, *"You want to speak to MIRIAM, okay, you select MIRIAM. You want to speak to Peter . . . "*. Here, the suggestion is to allow focus on a specific robot in a team by facilitating the interaction to be filtered down to just that robot and give it an individual identity. Here, work on CA personality and social interaction would be relevant and perhaps contribute to differentiating between robots. This may also heighten engagement when the conversation branches in this manner [37]. Perhaps different voices in speech output could also be used. P8 also wished for clear graphic differentiation of vehicles: *"In the text [. . .] give a color for vehicle one, vehicle two, vehicle three. With the colors we can understand quickly if [it] is vehicle one or two"*.

R4: CA designers should a) emphasize facts and values within utterances formed as sentences with highlighting in the CA's text output, and b) consider taking steps to differentiate output associated with different robots e.g. by using color in text output and c) allowing the conversational focus to be switched to a specific robot/agent within a robot team temporarily, assigning alternate personalities to the individual robots or alternate voices in speech output enabling differentiation and heightened engagement.

Theme 4: Correct Interpretation of User Input

This theme addresses the issues of correctly interpreting user input by resolving and/or avoiding ambiguity through autocorrection, context-sensitivity, and context clarity.

Participant P3 expected the CA to be more robust in the face of typographical errors: *"When for example I type a question* *and I write a wrong word, [. . .] MIRIAM answers "sorry but I couldn't answer" to that and why not suggest a question, like: "Did you mean . . . ?" Okay?"*. P7 described her interaction in which she found that the CA had too short a memory about what had come before:*"MIRIAM gives me a warning, I don't know, about batteries, something like that. And I was writing a message, so, when I pressed the ENTER button I didn't see the warning. So, she was asking me if I want to know more information about the warning. And I send the message. And [then] when I reply "Yes", MIRIAM doesn't know [to] what I am answering, yes."*.

This is an example of a gap between system context and user context, a phenomenon noted in CAs from other domains [25, 26, 42]. Here, it arises from the CA not correctly managing sub-dialogs in a way that the user might have expected. The user might have been better served by the CA seeking to look back in the conversation record to find the most recent utterance for which "Yes" would have been an appropriate user response and then asking a question to clarify the current context. The use of visual (or graphic) feedback to maintain a match between system context and user context is discussed later in the "Multimodalty" theme.

R5: We advise CA designers to a) provide context-sensitive auto-correction for input and b) provide as long a conversation context as possible within which users can implicitly refer to entities. (See also the Multimodality theme where we recommend providing context feedback graphically.)

Theme 5: Hands-free Use

P2 was experienced at launching and monitoring AUVs and talked about speech output: *"If you don't want to be looking at the computer screen on a rib [a small boat], which you definitely don't want to do, it's really helpful."* and then *"So if you can have something, this [the CA], on your earphone and you're just getting updates from this, it's quite helpful, I think. So this higher level mission status or some warnings about some fault or something like that. "*. This reflects what was found in a recent investigation of interaction with CAs, which concluded that "hands-free" was seen as the principle use case [42]. The reasons found in that study for using speech included: hands-dirty or otherwise engaged; device not within reach; speech felt to be faster; and attention divided between activities. Most of these would apply when operating in the field or in a boat, which was the situation our participant was describing.

R6: Provide voice input and output to allow hands-free use when needed e.g when used in problematic environments.

Theme 6: Multimodality

Here, we address the rationale of using chat alongside a graphical interface and their further integration. Three subthemes were identified.

Subtheme - The cognitive difference between visual (graphical) and verbal information: P3 pointed out that there are individual differences in regard to consuming information visually and verbally : *"Some people prefer the verbal, the language and some people prefer the visual, the visual things."*. P7 stated outright: *"I am more visual."*. Here, our participants were acknowledging differences in cognitive style. This is the psychological construct that explains individuals' differing preferences in the mental processing of information. Recent models of the visual and verbal dimensions of cognitive style take account of advances in neurophysiology [2]. The salience of this aspect of cognition to multimodal interfaces is recognized [50]. Recent work involving combining a CA and a graphic information display showed that, while cognitive style was a factor affecting the success with which users extracted information from the combined interface, operator situation awareness was still improved by using a CA and a graphical display in combination, irrespective of cognitive style [59].

Subtheme - The practical limitations of what can be conveyed either graphically or verbally (and hence their complimentarity): P1 felt that a graphical interface had its place but could not provide all the required information and suggested that some explanations might have to be verbal: *"I think the visual information helps a lot, but sometimes I understand that visual information is not enough for you to have a complete picture of what's going on. [. . .] sometimes, robots do some things that you won't expect or you would not understand what's going on, and in that case I thought it [the CA] was useful."*.

Subtheme - The desirability of tighter integration between graphical and verbal interaction: Participants wanted to be able to interact with the map on the graphic display to indicate vehicles, points or regions to be referred to in conversation, e.g. *"Also like if I'm the boat captain I can tell you [MIRIAM] I'm at this position [indicating clicking on the map] [...] am I interfering with any of the AUVs on the water?"* [P9]. Here, such use of multimodality could not only provide information for the conversation but also help to maintain a match between the CA's current context and that of the user in the case of selecting a vehicle on the map. Use of graphical feedback was found to reduce the occurrence of mismatch between system and user context in a study by Jain et al. [25]. Such steps would help to avoid gaps between system execution and user expectation as described by Norman [48].

R7: If the situation affords, CAs should not be used in isolation. Instead, **to exploit both visual (graphical) and verbal cognition**, combine them with graphical representations of the autonomous system state. If possible integrate the CA and the graphical components as a combined multimodal interface. The CA's current conversational context can be displayed graphically in a multimodal interface. If possible allow users to manipulate the conversation context by multimodal input.

Theme 7: Reassurance, Explanations and Trust

Participants expressed concern about the risk of losing expensive equipment. When asked how they felt when putting the

IVER3 AUV into the water P5 replied *"Some fear! It's complicated because we always don't know if we can get the robot back again."* and similarly *" I feel scared because it is a lot of money in the robot. In our robots also [it] is like that."[P8]*. Participants thought access to good monitoring information helped to alleviate such anxieties: *"If [. . .] I am operating this vehicle I have some assurance that it will operate okay, and with MIRIAM I'll just check if it's okay. "[P8].*

P4 addressed the aspect of autonomous behavior giving rise to anxiety: *"If the robot starts to act in a strange way I will be scared a little of that behavior right? And the trust will decrease obviously"*. Then, referring to the CA's explanations, P4 indicated that the CA could give reassurance: *"I think that information for example that MIRIAM tells, I think it's good information. Like now the robots did that because of that. [Indicating steps in an explanation on the stimulus laptop]"*. It is clear that participants felt that reassurance could be gained from the CA and that explanations helped in this. The style of explanations was also probed. When asked about the desirability of just being given the most likely or all possible explanations (low vs. high completeness [30]), participants wanted to see them all e.g. *"For me I want to know all the reasons"* [P4]. These findings align with previous work on a different sample of AUV operators [14]. However, previous work has not compared information needs of various personnel in the AUV domain. P2: *"If it's someone that is not the operator, [. . .] wants to have some insight of what's going on, maybe he doesn't want to see [all] that."* Thus, different user roles may require different levels of completeness.

R8: Firstly, simply relaying facts via a CA is reassuring. CAs should be equipped with explanation capabilities to reassure users when system status might be uncertain or behavior is obscure. The detail provided in the explanations should be varied depending on the user role and preference (see also R2 and R3).

Theme 8: Post Conversation Follow-up

The CA used in the study had the capability to dynamically set reminders, e.g. to contact maintenance about a fault at a later point. It also took the initiative and alerted the user to equipment faults. P9 thought that a post-mission email would be useful: *". . . you can have the reminder that when you finished if you have an oil leakage at the end of the mission, if you send an email on everything, okay: "Attention this robot is having an oil leakage, cannot operate.""*. P11 also suggested that any outstanding items of business from the conversation should be summarized in a post-conversation email including any faults flagged for maintenance: *"If something goes wrong a reminder for example at the end of the mission generates some sort of mission resumé [. . .]"* and then *"[. . .]depending on the number of errors that happened and faults and the severity, for example if it's critical or low. Generate a report and recommend some sort of maintenance. That might be very useful. "*.

One aspect of this, concerning the reminders, echoes existing CA behavior in commercial CA products such as Siri and Google Home. The other aspect, summarizing and collating particular events from a conversation, such as fault reports has been the focus of work on natural language post-mission reporting [76].

R9: CAs for remote autonomous systems should generate natural language reports to summarize mission events and collate user requests for action post-conversation. These could be communicated by email to individual users as an aide-memoire to facilitate action.

CONCLUSIONS AND FUTURE WORK

In this paper, we investigated interaction with a Conversational Agent for remote autonomous systems. The context for our study was Autonomous Underwater Vehicles (AUVs) capable of collaborating on multi-objective missions and operating with a high degree of autonomy. 12 research scientists and engineers experienced handling a high value, highly autonomous, marine survey AUV and using a prototype, mixed initiative Conversational Agent integrated with a commercial map-based autonomy user interface. During semistructured interviews, they revealed qualitative insights into what would be the desirable attributes of a Conversational Agent for interaction with teams of remote autonomous robots. We discussed these insights, related them to relevant prior work and derived **nine design recommendations** applicable to Conversational Agents for remote autonomy. These range from adaptive systems tailoring information filtering and frequency of updates to fit the user role (R2), through intelligent context sensitivity (R5), hands-free use (R6), exploiting multimodality (R7) and providing explanations of autonomous behavior customized for different user roles (R8) to collating unresolved items into post-conversation reports (R9).

In future work, the recommendations we would wish to prioritize in our research are R7 on Multimodalty and R8 on Explanations, i.e. we plan to investigate the benefits and practicalities of implementing the use of multimodal input to control conversation context (both user perception and system sensing), and also intend investigating what different levels of detail in explanations are suitable for which of the various user roles.

We hope our recommendations will inspire the design of Conversational Agents for remote autonomy thus will increase adoption and empower users to interact successfully with the complex autonomous systems of the future.

ACKNOWLEDGMENTS

We thank all those who took part in the Strongmar 2018 Summer School in underwater marine environment perception and the staff of the Underwater Centre, Fort William, Scotland, UK for their generous cooperation during the study. We also thank the anonymous reviewers for their helpful suggestions. The summer school was supported by the Strongmar Project H2020-TWINN-2015, 692427. This work was supported by UK MOD Dstl ACC101939 and the EPSRC EP/R026173/1 ORCA Hub.

REFERENCES

[1] Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational Contextual Cues: The Case of Personalization and History for Response Ranking. *arXiv preprint* (2016), 10. http://arxiv.org/abs/1606.00372

[2] Olesya Blazhenkova and Maria Kozhevnikov. 2009. The new object-spatial-verbal cognitive style model: Theory and measurement. *Applied Cognitive Psychology* 23, 5 (2009), 638–663.

[3] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. 2013. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction* 2, 1 (2013), 82–111.

[4] Susan E Brennan. 1998. The vocabulary problem in spoken dialogue systems. *Automated Spoken Dialog Systems, Luperfoy (Ed.). MIT Press, Cambridge, MA* (1998).

[5] Jessie Y. C. Chen and Michael J. Barnes. 2014. Human-agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (2014), 13–29.

[6] Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19, 3 (2018), 259–282.

[7] Juliet Corbin and Anselm Strauss. 2008. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

[8] Alastair. Cormack, David M. Lane, and Jon Wood. 2010. Operational experiences for maritime homeland security operations. In *Proceedings of OCEANS'10 IEEE SYDNEY*. 1–7.

[9] Amanda Cercas Curry, Helen Hastie, and Verena Rieser. 2017. A Review of Evaluation Techniques for Social Dialogue Systems. In *Proceedings of the ICMI Workshop on Investigating Social Interactions with Artificial Agents*.

[10] Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Getting to Know Users: Accounting for the Variability in User Ratings. In *Proceedings of the Workshop on Semantics of Dialogue (SemDial), Edinburgh, UK*.

[11] Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with Hidden Markov Model. In *Proceedings of Special Interest Group on Discourse and Dialogue (SIGDIAL'09)*.

[12] George Ferguson, James F Allen, and others. 1998. TRIPS: An integrated intelligent problem-solving assistant. In *Proceedings of AAAI/IAAI*. 567–572.

[13] George Ferguson, James F Allen, Bradford W Miller, and others. 1996. TRAINS-95: Towards a Mixed-Initiative Planning Assistant. In *Proceedings of AIPS*. 70–77.

[14] Francisco J. Chiyah Garcia, David A. Robb, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models through a Multimodal Interface. In *Proceedings of the International Conference on Natural Language Generation (INLG'18)*.

[15] Michael A. Goodrich and Alan C. Schultz. 2007. Human-robot Interaction: A Survey. *Found. Trends Hum.-Comput. Interact.* 1, 3 (Jan. 2007), 203–275. DOI: http://dx.doi.org/10.1561/1100000005

[16] Shirley Gregor and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *Management Information Systems Quarterly* 23, 4 (Dec. 1999), 497–530. DOI: http://dx.doi.org/10.2307/249487

[17] H Paul Grice. 1975. *Logic and conversation*. Academic Press, New York, Book section 3: Speech acts, 41–58.

[18] Joakim Gustafson, Linda Bell, Jonas Beskow, Johan Boye, Rolf Carlson, Jens Edlund, Björn Granström, David House, and Mats Wirén. 2000. AdApt-a multimodal conversational dialogue system in an apartment domain. In *Proceedings of the Sixth International Conference on Spoken Language Processing*.

[19] Helen Hastie. 2012. Metrics and evaluation of spoken dialogue systems. In *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Springer New York, 131–150.

[20] Helen Hastie, Francisco Javier Chiyah Garcia, David A. Robb, Pedro Patron, and Atanas Laskov. 2017a. MIRIAM: A Multimodal Chat-Based Interface for Autonomous Systems. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI'17*. ACM, 495–496. https://doi.org/10.1145/3136755.3143022

[21] Helen Hastie, Xingkun Liu, Pedro Patron, and Yvan Petillot. 2017b. Talking Autonomous Vehicles: Automatic AUV Mission Analysis in Natural Language. In *Proceedings of OCEANS MTS/IEEE, Aberdeen, UK* (2017).

[22] Helen Hastie, Katrin Lohan, Mike J. Chantler, David A. Robb, Subramanian Ramamoorthy, Ron Petrick, Sethu Vijayakumar, and David Lane. 2018. The ORCA Hub: Explainable Offshore Robotics through Intelligent Interfaces. In *Proceedings of the HRI Workshop on Explainable Robotic Systems, HRI'18*.

[23] Lyn Henderson, Michael Henderson, Scott Grant, and Hui Huang. 2010. What are users thinking in a virtual world lesson? Using stimulated recall interviews to report student cognition, and its triggers. *Journal For Virtual Worlds Research* 3, 1 (2010).

[24] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 295.

[25] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018a. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–6.

[26] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018b. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 895–906.

[27] Srinivasan Janarthanam and Oliver Lemon. 2014. Adaptive Generation in Dialogue Systems Using Dynamic User Modeling. *Comput. Linguist.* 40, 4 (Dec. 2014), 883–920. DOI: http://dx.doi.org/10.1162/COLI_a_00203

[28] Philip Nicholas Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive Science* 4, 1 (1980), 71–115.

[29] Filip Jurčíček, Blaise Thomson, and Steve Young. 2012. Reinforcement learning for parameter estimation in statistical spoken dialogue systems. *Computer Speech & Language* 26, 3 (2012), 168–192.

[30] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. *2013 IEEE Symposium on Visual Languages and Human-Centric Computing* (2013).

[31] Y. S. Kwon and B. J. Yi. 2012. Design and motion planning of a two-module collaborative indoor pipeline inspection robot. *IEEE Transactions on Robotics* 28 (2012), 681–696. DOI: http://dx.doi.org/10.1109/TRO.2012.2183049

[32] David Lane, Keith Brown, Yvan Petillot, Emilio Miguelanez, and Pedro Patron. 2013. *An Ontology-Based Approach to Fault Tolerant Mission Execution for Autonomous Platforms*. Springer New York, 225–255.

[33] Pierre Le Bras, David A. Robb, Thomas S. Methven, Stefano Padilla, and Mike J. Chantler. 2018. Improving User Confidence in Concept Maps: Exploring Data Driven Explanations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 1–13. DOI: http://dx.doi.org/10.1145/3173574.3173978

[34] Iolanda Leite, André Pereira, Allison Funkhouser, Boyang Li, and Jill Fain Lehman. 2016. Semi-situated Learning of Verbal and Nonverbal Content for Repeated Human-robot Interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, 13–20. DOI: http://dx.doi.org/10.1145/2993148.2993190

[35] Oliver Lemon, Anne Bracy, Alexander Gruenstein, and Stanley Peters. 2001. The WITAS multi-modal dialogue system. In *Proceedings of the Seventh European Conference on Speech Communication and Technology*. 1559–1562.

[36] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. *arXiv preprint* (2016), 10. http://arxiv.org/abs/1603.06155

[37] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do?: Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, 264–275.

[38] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128. DOI: http://dx.doi.org/10.1145/1518701.1519023

[39] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* (2016).

[40] José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language* 31, 1 (2015), 87–112.

[41] José Lopes, Katrin Lohan, and Helen Hastie. 2018. Symptoms of Cognitive Load in Interactions with a Dialogue System. In *ICMI'18 Workshop on Modeling Cognitive Processes from Multimodal Data.*

[42] 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*. ACM, 5286–5297.

[43] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human-agent teaming for Multi-UxV management. *Human Factors* 58, 3 (2016), 401–415.

[44] Emilio Miguelanez, Pedro Patron, Keith E Brown, Yvan R Petillot, and David M Lane. 2011. Semantic knowledge-based framework to improve the situation awareness of autonomous underwater vehicles. *IEEE Transactions on Knowledge and Data Engineering* 23, 5 (2011), 759–773.

[45] Teruhisa Misu, Antoine Raux, Ian Lane, Joan Devassy, and Rakesh Gupta. 2013. Situated Multi-modal Dialog System in Vehicles. In *Proceedings of the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction (GazeIn '13)*. ACM, New York, NY, USA, 25–28. DOI: http://dx.doi.org/10.1145/2535948.2535951

[46] E.T. Mueller. 2016. *Transparent Computers: Designing Understandable Intelligent Systems*. CreateSpace Independent Publishing Platform. https://books.google.co.uk/books?id=NhFRjwEACAAJ

[47] Keiji Nagatani, Seiga Kiribayashi, Yoshito Okada, Kazuki Otake, Kazuya Yoshida, Satoshi Tadokoro, Takeshi Nishimura, Tomoaki Yoshida, Eiji Koyanagi, and Mineo Fukushima. 2013. Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots. *Journal of Field Robotics* 30 (2013), 44–63.

[48] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Constellation.

[49] Florian Nothdurft, Felix Richter, and Wolfgang Minker. 2014. Probabilistic human-computer trust handling. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 51–59.

[50] Sharon Oviatt and Philip Cohen. 2000. Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the ACM* 43, 3 (2000), 45–53.

[51] Stefano Padilla, Thomas S. Methven, David A. Robb, and Mike J. Chantler. 2017. Understanding Concept Maps: A Closer Look at How People Organise Ideas. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 815–827.

[52] Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondrej Dušek, Verena Rieser, and Oliver Lemon. 2017. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. *Alexa Prize Proceedings* (2017).

[53] Yvan Petillot, Chris Sotzing, Pedro Patron, David Lane, and Joel Cartright. 2009. Multiple system collaborative planning and sensing for autonomous platforms with shared and distributed situational awareness. In *Proceedings of the AUVSI's Unmanned Systems Europe, La Spezia, Italy*.

[54] Patti J Price. 1990. Evaluation of spoken language systems: The ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

[55] Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let's Go Public! Taking a spoken dialog system to the real world. In *Ninth European Conference on Speech Communication and Technology*.

[56] Ehud Reiter and Robert Dale. 2000. *Building Natural Lang Generation Systems (Studies in Natural Language Processing)*. Cambridge University Press.

[57] Verena Rieser and Oliver Lemon. 2008. Automatic Learning and Evaluation of User-Centered Objective Functions for Dialogue System Optimisation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'08)*.

[58] Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of EMNLP*. Association for Computational Linguistics, 583–593.

[59] David A. Robb, Francisco J. Chiyah Garcia, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Keep Me in the Loop: Increasing Operator Situation Awareness through a Conversational Multimodal Interface. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM. https://doi.org/10.1145/3242969.3242974

[60] David A. Robb, Stefano Padilla, Britta Kalkreuter, and Mike J. Chantler. 2015. Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, 1355–1364. DOI: http://dx.doi.org/10.1145/2702123.2702470

[61] Stephanie Seneff and Joseph Polifroni. 1996. A new restaurant guide conversational system: Issues in rapid prototyping for specialized domains. In *Proceedings of the Fourth International Conference on Spoken Language, ICSLP 96*, Vol. 2. IEEE, 665–668.

[62] Stephanie Seneff and Joseph Polifroni. 2000. Dialogue management in the Mercury flight reservation system. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems-Volume 3*. Association for Computational Linguistics, 11–16.

[63] Amit Shukla and Hamad Karki. 2016. Application of robotics in onshore oil and gas industry-A review Part I. *Robotics and Autonomous Systems* 75 (2016), 490–507.

[64] David Silverman. 2010. *Doing qualitative research: A practical handbook* (third ed.). SAGE Publications Limited.

[65] Anselm L Strauss. 1987. *Qualitative analysis for social scientists*. Cambridge University Press.

[66] Kristen Stubbs, Pamela J Hinds, and David Wettergreen. 2007. Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems* 22, 2 (2007).

[67] Marc Swerts, Diane Litman, and Julia Hirschberg. 2000. Corrections in spoken dialogue systems. In *Proceedings of the Sixth International Conference on Spoken Language Processing*.

[68] Ella Tallyn, Hector Fried, Rory Gianni, Amy Isard, and Chris Speed. 2018. The Ethnobot: Gathering Ethnographies in the Age of IoT. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Article 604, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3174178

[69] The Institute of Engineering and Technology. 2015. *Engineering and Technology Skills and demand in industry 2015 survey. Overview of issues and trends from 2015 survey.* Technical Report. https://www.theiet.org/factfiles/education/skills2015-page.cfm Accessed on 6th Aug.2018.

[70] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson. 2016. Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. In *Proceedings of AISB Workshop on Principles of Robotics*. University of Bath.

[71] James Trevelyan, William R. Hamel, and Sung-Chul Kang. 2016. *Robotics in Hazardous Applications*. In Springer Handbook of Robotics, Springer International Publishing, Chapter 58, 1521–1548.

[72] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *CoRR* abs/1506.05869 (2015). http://arxiv.org/abs/1506.05869

[73] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: a framework for evaluating spoken dialogue agents. In *Proceedings of the Association for Computational Linguistics (ACL'97)*. 271–280.

[74] Marilyn A. Walker, Alex Rudnicky, Rashmi Prasad, John Aberdeen, Elizabeth Owen Bratt, John Garofolo, Helen Hastie, Audrey Le, Bryan Pellom, Alex Potamianos, Rebecca Passonneau, Salim Roukos, Greg S, Stephanie Seneff, and Dave Stallard. 2002. DARPA Communicator: Cross-System Results for the 2001 Evaluation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'02)*. 269–272.

[75] Marilyn A Walker, Stephen J Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science* 28, 5 (2004), 811–840.

[76] Zhuoran Wang and Helen Hastie. 2015. A prototype for AUV post-mission debrief generation from metadata. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15)*. International Foundation for Autonomous Agents and Multiagent Systems, 1959–1960.

[77] Joseph Weizenbaum. 1966. ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9, 1 (1966), 36–45.

[78] Jason D Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language* 21, 2 (2007), 393–422.

[79] C. Wong, E. Yang, X. T. Yan, and D. Gu. 2017. An overview of robotics and autonomous systems for harsh environments. In *Proceedings of the 2017 23rd International Conference on Automation and Computing (ICAC)*. 1–6. DOI: http://dx.doi.org/10.23919/IConAC.2017.8082020

[80] Robert H. Wortham, Andreas Theodorou, and Joanna J Bryson. 2017. *Robot transparency: Improving understanding of intelligent behaviour for designers and users*. Springer, 274–289. DOI: http://dx.doi.org/10.1007/978-3-319-64107-222

[81] Robert H. Wortham and Andreas Theodorou. 2017. Robot transparency, trust and utility. *Connection Science* 29, 3 (2017), 242–248. DOI: http://dx.doi.org/10.1080/09540091.2017.1313816

[82] Helen Wright-Hastie, Rashmi Prasad, and Marilyn Walker. 2002. What's the Trouble: Automatically Identifying Problematic Dialogues in DARPA communicator dialogue systems. In *Proceedings of ACL*. ACL, 384–391.

[83] Victor Zue, Stephanie Seneff, James R Glass, Joseph Polifroni, Christine Pao, Timothy J Hazen, and Lee Hetherington. 2000. JUPlTER: a telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing* 8, 1 (2000), 85–96.