Keep Me in the Loop: Increasing Operator Situation Awareness through a Conversational Multimodal Interface

David A. Robb
Heriot-Watt University
Edinburgh, U.K.
d.a.robb@hw.ac.uk

Francisco J. Chiyah Garcia
Heriot-Watt University
Edinburgh, U.K.
fjc3@hw.ac.uk

Atanas Laskov
SeeByte Ltd
Edinburgh, U.K.
atanas.laskov@seebyte.com

Xingkun Liu
Heriot-Watt University
Edinburgh, U.K.
x.liu@hw.ac.uk

Pedro Patron
SeeByte Ltd
Edinburgh, U.K.
pedro.patron@seebyte.com

Helen Hastie
Heriot-Watt University
Edinburgh, U.K.
h.hastie@hw.ac.uk

ABSTRACT

Autonomous systems are designed to carry out activities in remote, hazardous environments without the need for operators to micro-manage them. It is, however, essential that operators maintain situation awareness in order to monitor vehicle status and handle unforeseen circumstances that may affect their intended behaviour, such as a change in the environment. We present MIRIAM, a multimodal interface that combines visual indicators of status with a conversational agent component. This multimodal interface offers a fluid and natural way for operators to gain information on vehicle status and faults, mission progress and to set reminders. We describe the system and an evaluation study providing evidence that such an interactive multimodal interface can assist in maintaining situation awareness for operators of autonomous systems, irrespective of cognitive styles.

KEYWORDS

Multimodal output, natural language generation, autonomous systems, situation awareness, cognitive styles.

ACM Reference Format:

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 *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3242969.3242974

1 INTRODUCTION

Command and control interfaces are typically multimodal, consisting of graphical interfaces showing updates and the location of manned and unmanned vehicles. This type of User Interface (UI) is often combined with multiple human-human chat windows enabling the operator to communicate with personnel in the field, who can give situation updates and discuss logistics [35]. As systems become more autonomous, there is the potential for completely unmanned scenarios. This is desirable in highly hazardous locations, such as in first responder situations or on decommissioned oil and gas rigs [28]. Here, we present such a multimodal command and control (C2) interface, as illustrated in Figure 3, but with an automated conversational assistant and so eliminating the need for a human to be present in the field.

ICMI '18, October 16–20, 2018, Boulder, CO, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5692-3/18/10.
https://doi.org/10.1145/3242969.3242974

Figure 1: Two AUVs enabled with the Neptune autonomy framework that perform activities deep underwater such as surveying areas around oil and gas platforms.

In order for humans and machines to work effectively together as teams in such hazardous environments, it is essential that there is clear communication and high situation awareness. Robots and autonomous systems need to communicate their world view, system actions and reasoning in order to develop trust, avoid unnecessary costly aborts and increase adoption. The domain we address here is Autonomous Underwater Vehicles (AUVs) through collaboration with industry partner SeeByte Ltd, who have developed the Neptune autonomy framework for underwater vehicles (Figure 1). Situation awareness is key in the underwater domain as operators need to keep track of a) multiple and different types of objectives and goals, b) dynamic factors such as the effects of water currents and new objectives arising from new sensor data, and c) the fact that one or more AUVs may not be behaving as provisionally predicted. Interaction through natural language provides an intuitive means of keeping the operator informed and requires little training.

We propose a multimodal system that includes a conversational agent through chat. Conversational and multimodal systems are typically divided into *non-task based*, also known as social conversation or chitchat [42, 48] or *task-based* such as finding a restaurant

Figure 2: A simulator box, which runs a simulated AUV running Neptune (Figure 1). Configuration files allow simulation of factors including the onset of various faults and accelerated battery drain due to adverse environment conditions.

or making travel arrangements [7, 50]. Both of these types of systems tend to have information seeking phases to inform the user and some systems are completely dedicated to this, for example in the tourism domain [45] and conversational search [24]. Conversational systems to date that interface with a database do do mostly with a static database, albeit updated periodically, for example providing recommendations from a static list of restaurants, perhaps updated weekly. The system described here, however, provides information and updates in a fast-moving, evolving and changing world with vehicles providing information whilst performing autonomously. Therefore, there is an element of 'unknown' in terms of their observed behaviours and observed world, which cannot be preprogrammed. In addition, there is no predictable course of events, with the system having to communicate at various points in the dialogue, which also cannot be programmed in advance, for example if the vehicle gets stuck in the seabed. In addition, particularly for more complex missions with multiple vehicles in multiple domains (including air and surface, as well as marine), the dialogue will need to be multi-threaded with sub-dialogues returned to at certain points in the interaction. Finally, the mission goals are dynamic in that they are not known ahead of time and also cannot be preprogrammed.

The contributions in this paper thus are as follows:

(1) A unique multimodal interface with a conversational agent that is able to provide information and updates on underwater autonomous systems in a fast-moving dynamic world model and with mission tasks that are not known in advance.
(2) An evaluation providing evidence that incorporating a conversational agent as part of a multimodal interface can improve situation awareness in operators over a system that uses graphics alone.

We continue, in the rest of this paper, by discussing previous work forming the background to the system's development and evaluation methods used. We describe the prototype system and system architecture. Then, we detail the evaluation study and analyse the results. Finally, we discuss the significance of the results, possible future work and draw conclusions.

2 BACKGROUND

Operators of autonomous systems need to keep track of what their systems are doing and how they are behaving. This is often described as a need for situation awareness (SA). More widely, SA is the maintenance of an appreciation of salient events in one's environment or maintenance of a world or system view. A widely used definition of SA is that introduced by Endsley:

"Situation awareness is the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [13]

SA is important in commonplace activities such as walking, driving, and operating machinery. It is also important in good decision making in domains such as medicine, air traffic control and tactical command in warfare. The psychology of decision making, where maintenance of good SA is important, has led to the devising of models of SA [43, 44]. The model proposed by Endsley [15] describes the involvement of factors including system capability, interface design, stress and workload. SA is usually thought of as applying to human users of systems but can also be applied to the world view as it exists within an autonomous agent or team of autonomous agents and indeed this is an important factor in achieving high levels of autonomy in such systems [1][36]. In this paper, we are primarily concerned, however, with SA in the human operators managing these systems.

Models of SA show it is a complex and context dependent aspect of cognition. There are challenges in measuring it [2, 12, 51]. These include considerations of whether it can be accurately measured by subjective self-reporting techniques or by linking it directly to performance in achieving a given mission objective. These issues have been considered and one technique focusing on capturing the knowledge about a given situation currently possessed by a user has been described by Endsley [14]. This "Freeze Technique" involves having the user undertake a task or activity using a system and then pausing the activity and asking a question about the activity at that point. In the evaluation study described in Section 4, we apply a similar technique and take accuracy in users' answers as a proxy measurement for situation awareness. Increasingly, as the level of autonomy grows in autonomous systems, the problems of operators' reduced operational involvement in (but continued accountability for) tasks that systems do is also a factor in maintaining SA [16, 29].

Multimodality in interface design aims to offer users modalities of interaction, which not only suit human nature but also cater for individual differences. Therefore, when designing the evaluation (described later), we decided to gather data on participants' visual and verbal cognitive styles as they might be a factor. Cognitive styles, the individual cognitive preferences with respect to acquiring and processing information, have been recognised as one factor likely to affect the use of, and be catered for by, multimodal interfaces [37]. Cognitive styles should not be confused with learning styles (or strategies), which are the particular strengths that individuals have in ways of learning and are recognised as a separate construct [41]. Models encompassing both describe cognitive styles as feeding into learning styles along with other factors including working memory, intelligence, and personality [39]. The work of Blazhenkova and Kozhevnikov established that the visual and verbal cognitive styles encompass three mono-polar dimensions:

Object-imagery, Spatial-imagery and Verbal [5, 6, 30]. The object-imagery dimension measured preferences for the represention and processing of *"colorful, pictorial and high resolution images of individual objects"*, the Spatial-imagery scale quantified the preference for *"schematic images, spatial relations amongst objects and spatial transformations"* [6] and the verbal dimension measured preference for conceptualising in language. A three-subscale questionnaire, the *Object-Spatial Imagery and Verbal Questionnaire (OSIVQ)* measures those three monopolar dimensions [5] and has been used in several recent studies e.g. [3, 21, 22, 40].

Very few multimodal interfaces have been built for situation awareness of remote autonomous systems. The WITAS system [32] is one such system that incorporates situation awareness in a conversational interface for controlling airbourne vehicles. The autonomous behaviour lies in its ability to get from one waypoint to the next through route planning with the user able to change these waypoints through the WITAS multimodal system. The user can also get reports on a dynamically changing simulated world (e.g. observing "Truck 8 is turning left") and information on constraints on the system (e.g. "the system won't fly into a burning building"), but not on the health status of the vehicle or alerts/warnings, as in our system. Airbourne vehicles have higher bandwidth communication than AUVs and are required to be controlled by a human. The WITAS dialogue, therefore, focuses on control and waypoint specification (e.g. "then look at the parking lot, the hospital and the building"). Given the nature of the underwater domain, AUVs need greater levels autonomy and our interaction design is more about understanding and being updated on the autonomous systems' behaviours around mission objectives and goals with the assumption that the vehicles can complete them with minimal intervention. For WITAS, the airbourne vehicle plan was incorporated into the Dialogue (Interaction) Manager. While this approach results in interesting and complex dialogue pertaining to system control, it does mean that the WITAS system is highly system dependent whereas our use of an API means that we can interact with any vehicle type (e.g. REMUS, IVER as well as surface and air vehicles), as long as they are running the Neptune autonomy framework. Finally, the WITAS system was not evaluated with human subjects and not with respect to increasing situation awareness through conversation.

3 SYSTEM OVERVIEW

Figure 3 shows the multimodal interface consisting of the SeeTrack graphical interface with the conversational component running on the right hand side. The SeeTrack interface is a commercial product produced by SeeByte Ltd and consists of a map and status table display. The interface can run with real AUVs (Figure 1) running Neptune autonomy software or with a simulator (Figure 2). The latter is used for the evaluation set-up described in Section 4.

After the AUV operator uploads the mission plan to the AUV (or simulator) and gives the command to start, users can query using natural language through a chat interface about the mission plan and vehicle progress in real time as desired for the duration of the mission, simultaneously monitoring progress on the SeeTrack UI.

3.1 Neptune Autonomy Framework

A major part of the motivation behind development of a framework allowing a high level of autonomy is due to AUVs, in particular, facing problems of both continuity and bandwidth of communications. Acoustic connections used for underwater communication are much lower bandwidth than radio above the surface and it is often not possible to exchange all the data required by the operator to make an informed decision. The Neptune autonomy framework makes use of techniques described in [31, 36, 38] to enable the co-operation of multiple unmanned autonomous vehicles operating in the underwater, air or land domains. It allows the planning of missions by defining a set of objectives such as areas to be surveyed by patrolling in a search pattern while collecting sensor data. Another example of an objective would be a specific item of interest, such as a suspected unexploded mine, to be reacquired by revisiting a location and following some reacquisition behaviour suitable for the type of goal and the available sensors. Once the set of objectives is defined, these are entered and a rehearsal track calculated for each autonomous vehicle allocated to the mission. This is displayed on the map area of the UI indicating the autonomy's provisional solution for completing the mission objectives. The operator can accept this or decide to reallocate one or more objectives to the various vehicles. The planned objectives are then uploaded to the vehicles and they can be released to perform their tasks autonomously.

During the mission circumstances may change, e.g. wind or water currents might vary or sensors may reveal some new point of interest, which is then added as a new objective. In addition, within the framework, individual vehicles when undertaking an objective can generate a temporary exclusion zone associated with that objective and this is added to the developing mission status. It is here that the autonomy framework again comes into play. Each vehicle, as an autonomous agent, may at any time deviate from the provisional mission solution in the light of changed circumstances and undertake a different objective. This autonomous behaviour makes the design of the dialogue interaction very challenging.

3.2 Multimodal Interface

As part of the multimodal interface, natural language interactions include asking about the vehicle's current navigational status, the plan and its current objectives, estimated time of retrieval, previous activities, mission progress, hardware fault diagnosis, and estimated time of completion of a specified objective. The system also pro-actively warns users about important events, such as vehicle faults and critical battery status. It recognises the various levels of importance and will pin critical alerts. The interface can be customised for various user-preferences and allows the creation of reminders through chat. Below is an example interaction:

```
15:38:36 SYSTEM: 'Welcome. I can give you updates on a
  mission in progress. Just ask in natural language. If
  you get stuck, type help. Would you like an overview
  of the current mission?'
15:38:36 USER: 'Yes. please'
15:38:36 SYSTEM:'The mission has a total of 3 planned
  objectives, which include 1 survey area and 2 item
  reaquires'
15:39:45 SYSTEM:'Auv1 is in transit to Survey0'
15:44:52 SYSTEM:'Auv1 has reached Survey0'
```

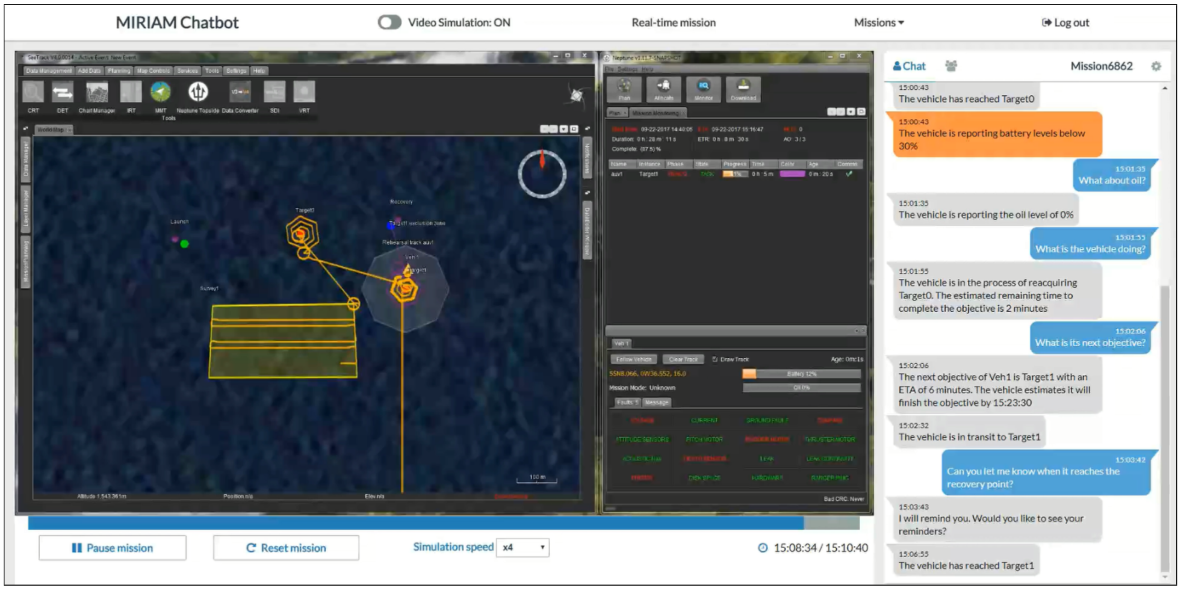

Figure 3: Multimodal interface for autonomous systems. The interface is shown as presented in the evaluation along with controls for progress in the simulated mission. Note the chat interface populated with a typical conversation and including a "low battery" alert initiated by the system.

```
15:56:08 USER:'What is the Estimated Time of Retrieval
   of the mission?'
15:56:08 SYSTEM:'The ETR of the mission is 58 minutes'
16:02:48 SYSTEM:'AUV1 reports battery levels below 50%'
16:04:51 USER:'Remind me to ask maintenance to check the
   battery'
16:02:48 SYSTEM:'Yes, I will remind you to ask
   maintenance to check the battery'
16:04:95 USER: 'What is the progress of the survey?'
16:04:96 SYSTEM: 'The vehicle has completed 20% of
   Survey0'
```

The system architecture is illustrated in Figure 4. Information from Neptune is obtained through an API and processed by methods described in [25, 27] along with the mission plan. This information is continually processed, organised and stored in a database for use by the conversational agent in real time. The conversational agent has three main components: 1) a Natural Language Processing (NLP) Engine that contextualises and parses the user's input for intent, formalising it as a semantic representation in logical form; 2) an Interaction Manager that decides what to say and accesses the database; and 3) a Natural Language Generation (NLG) component.

MIRIAM uses a rule-based NLP Engine that contextualises and parses the user's input for intent, formalising it as a semantic representation. It is able to process both static and dynamic data, such as names and mission-specific words. For example, it is able to reference dynamic objects such as"auv1", the particular name given to a vehicle in the mission plan, without the requirement to hard-code this name into the system. It can handle anaphoric references over multiple utterances e.g. "Where is Vehicle0?" ... "What is its estimated time to completion?". It also handles ellipsis e.g."What is the battery level of vehicle0?" ..."What about vehicle1?". The Interaction Manager is rule-based, receives the output of the NLP

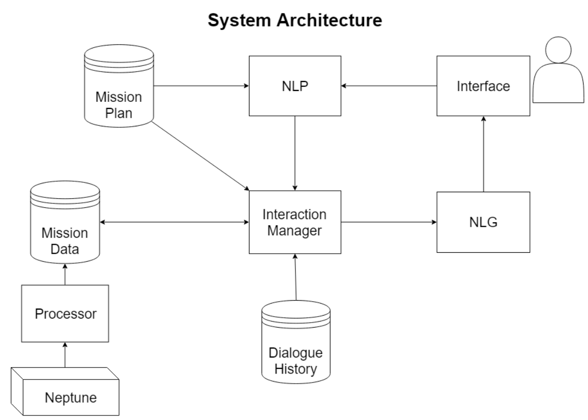

Figure 4: System architecture. Further detail in text.

and decides what dialogue act to do next e.g. getting clarification or providing the relevant information from the dynamic database. It then sends a dialogue act to the NLG component, which then uses a template-based Natural Language Generation component to generate the response or alert "The battery level is 62 percent". Dialogue acts are modified from the ISO standard[1] and include types for requesting and giving information, alerting and discourse structure such as openings/closings and requests for help.

4 EVALUATION

The aim of the evaluation study was firstly to assess the usability of the multimodal interface, gaining feedback from expert users to

[1] https://www.iso.org/standard/51967.html

help increase its effectiveness and secondly to find out if a multimodal interface (graphics + dialogue) provides improved situation awareness (SA) in operators over just graphics.

4.1 Participant Group

Given the esoteric nature of the domain, the multimodal interface for remote AUVs is not a system one could expect to place in the hands of novice users and hope to get meaningful feedback on how the original C2 interface has been improved by adding chat. With such novice users, it might be expected that chat would be used more than the original map and table display, which would inevitably be unfamiliar, unless the novice users were given extensive training.

We decided instead to focus our study on expert users already familiar with the original UI and AUVs. This would be both an advantage and a limitation. In terms of the quality and relevance of any qualitative feedback we could expect to collect, expert participants would be an advantage. At the same time, however, experts are usually few in number and, therefore, we might face the problems associated with lower statistical power in relation to any quantitative data collected. For this reason, rather than frame our evaluation as an experiment with different conditions, we designed it as an observational study in which all participants would do the same activity and we would analyse the quantitative data for correlations.

We recruited participants from within the ranks of the original C2 system's manufacturers. They took part on a voluntary basis with the time spent on participation being part of their professional work that day. There were 16 (14 male and 2 female, exactly reflecting current gender proportions of employees in the engineering and technology sector in the UK, 9% female [47]). They were aged 25 to 40, educated to undergraduate or masters degree level and all worked on developing the original C2 system, and include roles such as development and software engineers.

4.2 Procedure

We were interested in a) situation awareness, b) effect of cognitive styles, c) natural language interface performance, d) usability of the combined multimodal system, e) user satisfaction with chat, and f) feedback on how to improve the multimodal interface. Participants:

(1) completed a pre-task questionnaire for demographic details, and also the OSIVQ standard questionnaire,
(2) did a task in which they viewed and were allowed to interact with a simulated AUV mission in the original UI presented alongside the conversational agent (Figure 3) and periodically answered questions about the mission, and finally
(3) completed a post-task questionnaire to collect feedback on their opinions of the system.

For the task, to ensure consistency of presentation across participants the original UI output of a simulated mission (generated using the Neptune mission simulator) was video captured. The mission activity database was populated from the same simulated mission forming a time-stamped record of every mission event and vehicle status update. A web interface was constructed to present the video alongside the chat interface. Live chat based on the simulated mission was achieved by synchronising and restricting the chatbot's database access time frame with the video's progress (see Figure 3). Normally a mission involving a single AUV can be a relatively sedate affair taking some time to complete. To a) add more challenge to the maintenance of SA, and b) reduce the time on task required and possible fatigue (and consume less of our volunteers' valuable time), we compressed mission time eight fold.

4.3 Objective Measures

We would assess Situation Awareness by having the participants answer questions about mission and vehicle status presented verbally by the experimenter at specific timed points during the mission. At these times, the mission progress would be paused or "frozen" in a similar manner to that described by Endsley [14]. The 8 questions reflected the 3 levels of SA per Endsley's model [15] of 1) perception (e.g. location, status of vehicle, current objective),2) comprehension (e.g. mission timing and status) and 3) projection (e.g. given its status will the vehicle be able to complete the mission). Example questions include *"What is the total mission progress?"* and *"Is the battery sufficient to reach the recovery point?"*. **Situation awareness** was proxied by the percentage of correct answers by each participant. An answer of "Don't know" this was counted as incorrect.

Frequency of information source used would be tallied to establish from which part of the multimodal interface participants were sourcing the answers to the task questions. Specifically, after answering a question during the task, participants were asked to specify whether they got their answer from "Chat, SeeTrack, or Both". For each participant f-Info-Source being Chat (non-exclusively) was taken as "Chat" plus "Both" and likewise for SeeTrack was taken as "SeeTrack" plus "Both". Going forward f-Info-Source will be in terms of these non-exclusive counts.

To allow participants' visual and verbal cognitive styles to be taken into account the validated Object-Spatial Imagery and Verbal Questionnaire (OSIVQ) [5] was used. Each OSIVQ item is a 5 point Likert scale item. 45 items form three subscales. Participants would complete the OSIVQ following its standard instructions with their responses collated into three subscale scores (**OSIVQ Object-imagery**, **Spatial-imagery** and **Verbal**). As the OSVIQ is a validated scale, these scores are taken as ratio data ranging between 1 and 5 (see [5] for details).

Dialogue features similar to those traditionally collected when evaluating a spoken dialogue system [23] including number of user and system turns, frequency of dialogue act types, mean number of words per turn (both system and user) would be gathered from the dialogue system logs. **Concept accuracy** (CA), i.e. how accurately the system interprets the user's utterance, would be calculated by dividing the number of appropriately answered user turns by the total number of user turns.

4.4 Subjective Measures

Using the post-task questionnaire, **usability** would be assessed using the System Usability Scale (SUS)(©Digital Equipment, 1986) as described by Brooke [8], which covers various aspects including learnability, effectiveness, efficiency, aesthetics, system personality and appeal. The SUS questions would be prefaced by *"These questions refer to the combined SeeTrack/MIRIAM system"* so as to measure the overall usability of the system. Also in the post-task questionnaire, **user satisfaction** (USat) would be assessed using a

Table 1: Measures: Frequency of information source used (f-Info-Source), Situation Awareness (SA), OSIVQ Spatial-imagery score (Spatial-I), Usability (SUS), User Satisfaction (USat) and Concept Accuracy (CA).

	f-Info-Source Chat	SeeTrack	SA	Spatial-I	SUS	USat	CA
Mean	**6.3**	**3.9**	**85.2**	**3.56**	**75.6**	**3.0**	**0.79**
SD	1.4	1.3	11.4	0.50	13.0	0.4	0.13
Mdn	6.0	4.0	87.5	3.57	77.5	3.0	0.79
Min	4.0	2.0	62.5	2.53	47.5	2.5	0.58
Max	8.0	6.0	100.0	4.27	97.5	4.1	1.00

scale of 5-point Likert items. They would be prefaced by *"These questions refer to your interaction with MIRIAM only"* so as to measure satisfaction with natural language interaction. The 8 items (listed below) were adapted from the PARADISE evaluation framework as used and reported in [50]:

- Q1: I felt that MIRIAM understood me well
- Q2: I felt MIRIAM was easy to understand
- Q3: I felt it was easy to get the information I needed to answer the questions
- Q4: I found MIRIAM's alerts useful
- Q5: I knew what I could say at each point in the mission
- Q6: I felt that MIRIAM was very quick to respond
- Q7: MIRIAM behaved as I expected she would
- Q8: From my current experience with using MIRIAM, I would use the system regularly

4.5 Study Setting

All participants were fully briefed as to the nature and purpose of the study. Appropriate consents were obtained. As the participants were unfamiliar with the conversational component before doing the task, they were asked to read a short tutorial sheet, which explained the task and the kind of information that could be garnered through conversation. They were also informed that they could take notes[2] and ask questions. Audio recordings were made to allow review and accurate assessment of participant answers.

5 RESULTS

Descriptive results are given in Table 1. Of the three OSIVQ cognitive styles subscale scores (Object-imagery, Spatial-imagery and Verbal), only Spatial-imagery was found to be a significant factor as discussed below and, hence, is the only score shown in this table.

5.1 Information Source

There was a significant difference between the mean f-Info-Source (Chat v.s. SeeTrack) during the task (first two columns of Table 1). The f-Info-Source SeeTrack distribution was significantly non-normal (p <.05) as determined by the Kolmogorov-Smirnov test. Thus, the two distributions were compared in a two-tailed Wilcoxon

[2]Only one participant actually took notes, making a note used later in a qualitative comment.

Table 2: Bivariate correlations with OSIVQ Spatial-imagery subscale score. These statistically significant correlations are shown along with their two-tailed p thresholds.

	OSIVQ Spatial-imagery Spearman's *rho*	p
SA	-0.51	<.05
f-Info-Source Chat	-0.56	<.05

Signed-Rank test. During the task, the frequency with which participants sourced their information from Chat (*Mdn* = 6) was significantly higher than from SeeTrack (*Mdn* = 4), T=4, p<.01, r = -0.69. This is a large effect [18]. It shows rather unexpectedly, that despite our users being experts familiar with the original UI and intimately involved in the original Neptune/SeeTrack UI's development, they nonetheless made use of the conversation component significantly more that they did the original UI when asked for information about the mission and vehicle status. Their qualitative comments, described later, offer illuminating insights into the reasons for this.

5.2 Correlations with OSIVQ Spatial-imagery

Participants' OSIVQ Object-imagery and Verbal subscale scores were not found to be significant factors, however, there were strong and statistically significant negative correlations[3] of OSIVQ Spatial-imagery subscale score with both SA and f-Info-Source being Chat (see Table 2). The higher a participant's Spatial-imagery score the more likely they were to have used Chat less and relied on SeeTrack more often. This was not unexpected as cognitive styles theory would predict that people with high Spatial-imagery scores would be attracted to consuming information diagrammatically rather than textually. There was also a negative correlation with their situation awareness (SA). These correlations with OSIVQ Spatial-imagery score affected how we went on to analyse the relationship between SA and Information Source.

5.3 Information Source and Relation to Situation Awareness

Due to the significant negative correlation of OSIVQ Spatial-imagery score with both SA and f-Info-Source Chat, we used partial correlation controlling for OSIVQ Spatial-imagery to analyse the relationship between SA and f-Info-Source. i.e. we can report the correlations without the influence of OSIVQ Spatial-imagery [18] (Table 3). There was a strong and statistically significant positive correlation between SA and f-Info-Source Chat while controlling for OSIVQ Spatial-imagery score. This represents a large effect and provides good evidence that augmenting the existing C2 interface with chat supports situation awareness irrespective of cognitive style. Conversely, the strong negative correlation between SA and f-Info-Source SeeTrack shows that those relying more on the original UI had lower situation awareness.

[3]Spearman's Rho was used as f-Info-Source SeeTrack, which features later in the analysis, was not normally distributed [18]

Table 3: Partial Correlations of f-Info-Source with Situation Awareness (SA) controlling for OSIVQ Spatial-imagery score. These statistically significant correlations are shown along with their two-tailed p thresholds.

	SA (Controlling for Spatial-imagery)	
	Spearman's *rho*	*p*
f-Info-Source Chat	0.72	<.01
f-Info-Source SeeTrack	-0.58	<.05

5.4 Multimodal Interface Performance

In terms of concept accuracy, the conversational agent was able to understand the user's request and answer 86.4% of all user queries appropriately (not counting user inputs with misspellings or out-of-domain requests). Where it failed to answer correctly or replied that it did not know, the lack of appropriate response was due to the parser not covering the wording of the response. Including misspellings in the concept-accuracy calculation reduced it to 84.0% and adding in the out of domain queries reduced it further to 78.8%.

Overall positive usability (SUS) ratings were given by the expert operators to the combined chat and original UI (mean 76.6/100, Table 1). User Satisfaction score has a median of 3/5 and focuses on evaluating the conversational part of the multimodal system. PARADISE style [49] evaluation was performed to understand which objective metrics contribute to the response variable of user satisfaction score. This was done by analysing the weights derived through multi-variable linear regression. A model derived in this manner showed variables with positive weights that include the number of user/system turns indicative of the user's amenability to longer chat interactions (coefficient=0.2 for both). This somewhat goes against task-based dialogues evaluations, which have deemed dialogue length to negatively affect User Satisfaction (as reflected by a negative coefficient [50]). 'Mean time to answer' is negatively correlated to user satisfaction in our model (coefficient=-0.3) and SA is positively correlated (coefficient=0.3). Perhaps unsurprisingly, these results indicate that the users liked interactions where they got enough information as to be able to give a (correct) answer quickly. Finally, the number of system words per turn is positively correlated (coefficient=0.2) and there is some confirmation for this in the qualitative comments (see Section 5.5).

Whilst this model gives us some insight into the aspects of dialogue that contribute to high user satisfaction the model does not cover a large proportion of the variance ($R^2 = 0.36$). Prior studies have shown that Task Success is a large contributor to this variance [50]. As discussed in the introduction, some types of interactive systems can be easily evaluated in terms of Task Success, such as finding a restaurant that matches the user's criteria. However, defining Task Success for interactive systems for remote autonomy, such as the one described here, is not straightforward. Task success could relate to the performance of the operator in terms of good decision making or time on task [9] or it could relate to mission success of the autonomous systems. To confound the problem, the definition of Task Success will vary from mission to mission, again not necessarily known at the point of development or the point of deployment. Task Success could also be in terms of learning gain as with intelligent tutoring systems [19] or traditional Information Retrieval measures such as F-score. SA could be viewed as a type of learning gain, but it is clear that Task Success is multi-dimensional and, therefore, further work is needed to model it more accurately.

5.5 Qualitative Feedback

In the post-task questionnaire, we asked two open questions: 1) *"Tell us what you liked about MIRIAM"* and 2) *"Tell us what you didn't like about MIRIAM"*. An inductive, thematic analysis was done using grounded theory with open coding [10, 46]. There was a single coder (an author). Five themes were identified.

Theme 1: Suggestions for extra data coverage: Two participants were missing one particular aspect of mission information and requested that it be added. e.g. *"Some elements of the mission were left out (spiral down to survey altitude)"* [P7]. We have since added this to the data stream.

Theme 2: Conversational interaction and wish for more detail: Three participants expressed a wish for the conversational agent to be able to go into more detail when needed and for the system to accept a wider range of questions to access any given fact. e.g. *"She does not understand all questions, and it can be hard to think of a different way to phrase a question once it is in someone's mind."* [P5] and *"[MIRIAM] can give status updates but does not know what these mean... when I asked her what a general electric fault is ... she was unable to clarify in the way a person might be able to."* [P2]. Definitions of general terms and explanations of behaviour have been added to subsequent versions of the system [20].

Theme 3: Notifications (number and succinctness): Five participants were positive e.g. *"I liked the fault warnings and notifications about reached / completed objectives."* [P3], *"[I liked the] Rapidity... clear small snippets of information "* [P7], and *"This helped to get a good situation awareness throughout the mission (especially when combined with the [SeeTrack] Interface)"* [P5]. Here, P5 was also appreciating the synergy of combining SeeTrack with a conversational agent. Two participants thought there was sometimes too much information e.g. P4 liked the fault alerts but felt there may have been too many updates in general although this may have been due to the format of the task: *"Good alerting of faults."*. Then later: *"Some of the fault alerts were too verbose and some status updates felt too frequent (this may be primarily due to the speed up of the simulation)."* [P4]. We see here that the balance between enough and too much information can be delicate.

In Themes 2 and 3, we can see that the amount of detail provided needs to be balanced with a need to keep the output succinct. We can see that users expect the conversational agent to be both concise and possess knowledge, which can be queried in depth via several context sensitive sub-dialogues.

Theme 4: Wish for added multimodalty: P8 desired greater integration between the map display and chat requesting that the GPS information for the next objective given in chat be linked to the map. P1 suggested "*perhaps sound would be good for critical alarms/faults in addition to text*". Speech in and out has been added to subsequent versions of the system.

Theme 5: Ease of use and supporting less experienced operators: Eight participants expressed their appreciation of the ease and/or speed with which information could be extracted through dialogue interaction, e.g. *"Easy to use. Responsive."* [P6], and *"It makes it easy to get contextual information, and information that is not easy to get unless you are an expert on the SeeTrack/Neptune UI"* [P2]. A sub-theme here was using chat to confirm information noticed in SeeTrack: "*Has a useful purpose I think for the operator to confirm what they are seeing and I think the operator would feel more comfortable and assertive on making operational decisions*" [P12].

While we interpret Theme 3 as demonstrating that expert users see the benefit to situation awareness provided by the notifications and alerts in particular, Theme 5 also shows that our expert users see the potential for conversational agents to enable less experienced operators to successfully query the status of missions and to support operators of any level of experience in making more confident decisions by providing additional confirmation of mission facts over and above information gleaned from graphic and tabular displays.

6 DISCUSSION AND FUTURE WORK

In this section, we will discuss cross-cutting themes in our results and possible directions for future work. The overall positive usability (SUS) ratings given by the expert operators to the combined chat and original UI (76.6/100 in Table 1) are consistent with their qualitative feedback about the chat interface. We interpret the User Satisfaction along with the SUS ratings and positive comments as approval for developing the C2 UI into a multimodal UI with chat. The positive comments about ease of use and usefulness of fault alerts (Themes 3 and 5) confirm that overall the augmentation with chat was well received. Detailed comments were highly informative as to where the expert operators would like to see improvements.

Qualitative comments revealed a desire in some of these expert users for the conversational agent to have access to more detailed knowledge about aspects of the missions and this has sign-posted future improvements. In addition, a major theme in their comments was the ease of use the chat interface offers and its potential for making mission data accessible to less experienced operators. Indeed, there may be a novelty factor at play with regards to the chat interface and we plan to do long term studies to explore this and confirm our findings.

We are working on extending the conversational agent to work with multi-vehicle missions so as to further exploit the benefits of chat when operators face more challenges in maintaining situation awareness. In addition, extending the interface to be able to explain causal reasoning [4, 20] behind the actions of agents operating within the autonomy framework is a key future direction [26]. We also plan for versions that further embrace multimodal interactivity e.g. (*input*) clicking on the vehicle in question on the map display to indicate the object of a chat query or (*output*) highlighting a survey area on the map when under discussion in chat. Input could take into account the user's state, as well as the system state as in [34], for example enabling intelligent alerting to mitigate cognitive overload and customised views and behaviours depending on category or training level of the user [11, 17].

It has been observed informally by the authors that operators of autonomous systems make use of generic human-human chat to coordinate their activities when multiple vehicles are teamed on a mission. The human-human chat can contain utterances that are highly salient for SA, however, cognitive overload due to chat has also been observed in military training scenarios [11]. A combined human-human-robot chat in a system, which would allow automated monitoring and filtering of the chat stream to perhaps provide a salient chat digest is a goal we hope to pursue.

Our prototype multimodal interface encompassing natural language interaction augmenting an existing C2 interface has been developed in the context of the Neptune autonomy framework, which facilitates the tasking and monitoring of autonomous vehicles through a map-based and tabular C2 interface. The system populates a database with mission data and frequent, periodic vehicle status updates. The conversational agent then can pro-actively output alerts and notifications as well as answer natural language queries from users as a mission progresses. The contribution of such a system is its ability to interact on-the-fly with dynamic mission information through interfacing with the Neptune API and natural language processing using a custom-built NLP. The system is rule-based as was in-line with funders' requirements for regularised expressions that conform with standard operating procedures. Future work could look at a hybrid system that allows for the use of machine learning for interaction enabling the system to cope with an even broader range of unseen situations in a robust manner [33], whilst still conforming to interaction rules and regulations.

7 CONCLUSIONS

To help operators of autonomous systems maintain good situation awareness, we have developed a multimodal interface encompassing a natural language text chat interface to augment existing command and control (C2) interfaces. In a mixed-methods observational study, we evaluated the new system with expert operators of the original C2 interface. We gathered objective measures: situation awareness, the frequency with which users sourced information from the original UI or from chat, and dialogue system performance. We measured users' cognitive styles in case these had an effect on interaction with this combination of visual and verbal interfaces. We also collected usability ratings, user satisfaction ratings and qualitative comments for feedback on the system.

We found that our participants (expert users of the original interface) when asked to extract information from the augmented interface used chat statistically significantly more than the original map and table interface components. Using partial correlation to control for Spatial-imagery cognitive style, a correlation analysis exposed strong and statistically significant correlations showing the more that subjects used chat, the greater their situation awareness. In addition, the combined interface was rated highly for usability. In short, our study provides evidence that combining an autonomy control interface with natural language interaction supports situation awareness in operators of autonomous systems and is deemed effective by users.

ACKNOWLEDGMENTS

We gratefully thank and acknowledge our funders: UK MOD Dstl ACC101939; RAEng/ Leverhulme Trust (Hastie/LTSRF1617/13/37); EPSRC EP/R026173/1 ORCA Hub.

REFERENCES

[1] Julie A Adams. 2007. Unmanned vehicle situation awareness: A path forward. In *Human systems integration symposium.* 31–89.

[2] Marilyn Jager Adams, Yvette J Tenney, and Richard W Pew. 1995. Situation awareness and the cognitive management of complex systems. *Human factors* 37, 1 (1995), 85–104.

[3] Ann Austin and Jose Abdelnour Nocera. 2015. So, Who Exactly IS The HCI Professional?. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2732906, 1037–1042.

[4] Vaishak Belle. 2017. Logic meets Probability: Towards Explainable AI Systems for Uncertain Worlds. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 5116–5120.

[5] 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.

[6] Olessia Blazhenkova, Maria Kozhevnikov, and Michael A. Motes. 2006. Object-spatial imagery: a new self-report imagery questionnaire. *Applied Cognitive Psychology* 20, 2 (2006), 239–263.

[7] Antoine Bordes and Jason Weston. 2016. Learning End-to-End Goal-Oriented Dialog. *arXiv preprint* (2016). arXiv:1605.07683 http://arxiv.org/abs/1605.07683

[8] John Brooke. 1996. SUS-A quick and dirty usability scale. In *Usability evaluation in industry*. Taylor Francis, London, Chapter 21, 189–194.

[9] Jing Cheng, Xiao Hu, and P Bryan Heidorn. 2010. New Measures for the Evaluation of Interactive Information Retrieval Systems : Normalized Task Completion Time and Normalized User Effectiveness. In *Proceedings of ASSIST*. 1–9.

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

[11] M. L. Cummings. 2004. The Need for Command and Control Instant Message Adaptive Interfaces: Lessons Learned from Tactical Tomahawk Human-in-the-Loop Simulations. *Cyberpsychology and behaviour* 7, 6 (2004).

[12] Francis T. Durso and Scott D. Gronlund. 1999. Situation awareness. *Handbook of applied cognition* (1999), 283–314.

[13] Mica R. Endsley. 1988. Design and Evaluation for Situation Awareness Enhancement. *Proceedings of the Human Factors Society Annual Meeting* 32, 2 (1988), 97–101.

[14] Mica R. Endsley. 1995. Measurement of situation awareness in dynamic systems. *Human factors* 37, 1 (1995), 65–84.

[15] Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human factors* 37, 1 (1995), 32–64.

[16] Mica R. Endsley and Esin O. Kiris. 1995. The out-of-the-loop performance problem and level of control in automation. *Human factors* 37, 2 (1995), 381–394.

[17] Yu-Hong Feng, Teck-Hou Teng, and Ah-Hwee Tan. 2009. Modelling situation awareness for Context-aware Decision Support. *Expert Systems with Applications* 36, 1 (2009), 455–463.

[18] Andy Field. 2009. *Discovering Statistics Using SPSS* (3rd ed.). Sage, London.

[19] Kate Forbes-Riley and Diane J. Litman. 2006. Modelling User Satisfaction and Student Learning in a Spoken Dialogue Tutoring System with Generic, Tutoring, and User Affect Parameters. In *Proceedings of NAACL*. 264–271. https://doi.org/10.3115/1220835.1220869

[20] Francisco J Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explain Yourself: A Natural Language Interface for Scrutable Autonomous Robots. In *Proceedings of the HRI Workshop on Explainable Robotic Systems, HRI'18*.

[21] Erhan Selcuk Haciomeroglu. 2016. Object-spatial Visualization and Verbal Cognitive Styles, and Their Relation to Cognitive Abilities and Mathematical Performance. *Educational Sciences: Theory and Practice* 16, 3 (2016), 987–1003.

[22] Erhan Selcuk Haciomeroglu and Mark LaVenia. 2017. Object-Spatial Imagery and Verbal Cognitive Styles in High School Students. *Perceptual and Motor Skills* (2017), 0031512517698555.

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

[24] Helen Hastie, Marie-Aude Aufaure, Panos Alexopoulos, Hugues Bouchard, Catherine Breslin, Heriberto Cuayáhuitl, Nina Dethlefs, Milica Gašić, James Henderson, Oliver Lemon, et al. 2014. The PARLANCE mobile application for interactive search in English and Mandarin. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 260.

[25] Helen Hastie, Francisco Javier Chiyah Garcia, David A. Robb, Pedro Patron, and Atanas Laskov. 2017. 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

[26] Helen Hastie, Francisco J. Chiyah Garcia, David A. Robb, Pedro Patron, and Atanas Laskov. 2018. MIRIAM: A Multimodal Interface for Explaining the Reasoning Behind Actions of Remote Autonomous Systems. In *Proceedings of the ACM 2018 International Conference on Multimodal Interaction (ICMI âĂŹ18),*. ACM, New York, NY, USA.

[27] Helen Hastie, Xingkun Liu, and Pedro Patron. 2017. Trust triggers for multimodal command and control interfaces. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI'17*. ACM, 261–268. https://doi.org/10.1145/3136755.3136764

[28] 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*.

[29] Brian Kidwell, Gloria L. Calhoun, Heath A. Ruff, and Raja Parasuraman. 2012. Adaptable and Adaptive Automation for Supervisory Control of Multiple Autonomous Vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1 (2012), 428–432.

[30] Maria Kozhevnikov, Olesay Blazhenkova, and Michael Becker. 2010. Trade-off in object versus spatial visualization abilities:Restriction in the development of visual-processing resources. *Psychonomic Bulletin and Review* 17, 1 (2010), 29–35.

[31] 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, New York, NY, 225–255.

[32] 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.

[33] Oliver Lemon and Olivier Pietquin. 2012. *Data-driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*. Springer. https://doi.org/10.1007/978-1-4614-4803-7

[34] Ivan Maza, Fernando Caballero, R. Molina, N. Pena, and A. Ollero. 2010. Multimodal interface technologies for UAV ground control stations A Comparative Analysis. *Journal of Intelligent and Robotic Systems* 57 (2010), 371–391. https://doi.org/10.1007/s10846-009-9351-9

[35] Emily W. Medina, Sunny Fugate, LorRaine Duffy, Dennis Magsombol, Omar Amezcua, Gary Rogers, and Marion G. Ceruti. 2007. Next-Generation Tactical-Situation-Assessment Technology (TSAT): Chat. In *Web Engineering*, Luciano Baresi, Piero Fraternali, and Geert-Jan Houben (Eds.). Springer Berlin Heidelberg.

[36] 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.

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

[38] 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. *Proceedings of the AUVSI's Unmanned Systems Europe, La Spezia, Italy* (2009).

[39] Stephen Rayner and Richard Riding. 1997. Towards a categorisation of cognitive styles and learning styles. *Educational Psychology* 17, 1-2 (1997), 5–27.

[40] David A. Robb, Stefano Padilla, Thomas S. Methven, Britta Kalkreuter, and Mike J. Chantler. 2017. Image-based Emotion Feedback: How Does the Crowd Feel? And Why?. In *Proceedings of the 2017 ACM Conference on Designing Interactive Systems, DIS'17*. ACM, 451–463. https://doi.org/10.1145/3064663.3064665

[41] Eugene Sadler-Smith. 2001. The relationship between learning style and cognitive style. *Personality and individual differences* 30, 4 (2001), 609–616.

[42] Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. *CoRR* abs/1611.06216 (2016). arXiv:1611.06216 http://arxiv.org/abs/1611.06216

[43] Kip Smith and Peter A Hancock. 1995. Situation awareness is adaptive, externally directed consciousness. *Human Factors* 37, 1 (1995), 137–148.

[44] Neville A. Stanton, Paul M. Salmon, Guy H. Walker, and Daniel Jenkins. 2009. Genotype and phenotype schemata as models of situation awareness in dynamic command and control teams. *International Journal of Industrial Ergonomics* 39, 3 (2009), 480–489.

[45] Manfred Stede and David Schlangen. 2004. Information-Seeking Chat: Dialogue Management by Topic Structure. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue*. 117–-124. https://pdfs.semanticscholar.org/a629/abf9b8a896968b6f9db0baf234f8401ccb90.pdf

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

[47] 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.

[48] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[49] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: a framework for evaluating spoken dialogue agents. In *Association for Computational Linguistics*. 271–280.

[50] 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 ICSLP 2002*. 269–272.

[51] Christopher D Wickens. 1996. Situation awareness: impact of automation and display technology. *Situation awareness: limitations and enhancement in the aviation environment* (1996), k2.