Can we talk? How a talking agent can improve human autonomy team performance

Bogg, A., Birrell, S., Bromfield, M. A. & Parkes, A. M.

Original citation & hyperlink:
Bogg, A, Birrell, S, Bromfield, MA & Parkes, AM 2021, 'Can we talk? How a talking agent can improve human autonomy team performance', Theoretical Issues in Ergonomics Science, vol. 22, no. 4, pp. 488-509.
https://dx.doi.org/10.1080/1463922X.2020.1827080

DOI 10.1080/1463922X.2020.1827080
ISSN 1463-922X
ESSN 1464-536X

Publisher: Taylor and Francis

This is an Accepted Manuscript of an article published by Taylor & Francis in Theoretical Issues in Ergonomics Science on 25/11/2020, available online: http://www.tandfonline.com/10.1080/1463922X.2020.1827080

Can We Talk? – How A Talking Agent Can Improve Human Autonomy Team Performance.

Adam Bogg[a]*, Stewart Birrell[a], Michael A. Bromfield[b], Andrew M. Parkes[c]

[a] Institute for Future Transport and Cities, Coventry University, Coventry, United Kingdom

[b]School of Metallurgy and Materials, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[c]School of Art, Design and Architecture, Monash University, Melbourne, Australia

*Adam Bogg, email: bogga@coventry.ac.uk, address: National Transport Design Centre, Swift Road, Off Puma Way, Coventry, CV1 2TT, United Kingdom of Great Britain and Northern Ireland

Biographical notes:

Adam Bogg is a PhD Research Student at Coventry University researching aviation Human Autonomy Teaming communication interfaces. Prior to attending Coventry University Adam was an Aviation Training Manager, specialising in Training Needs Analysis and the development of eLearning solutions, with over 25 years' experience designing and managing the delivery of aviation career training. His career has primarily been involved with military aviation training, serving in the Royal Air Force and Royal New Zealand Air Force and latterly managing contracts for academic training delivered to the Royal New Zealand Air Force and recently the United Arab Emirates Air Force.

ORCID: 0000-0002-8926-1172

LinkedIn: https://www.linkedin.com/in/adam-bogg-3833aa27/

Affiliation: Institute for Future Transport and Cities, Coventry University, Coventry, United Kingdom of Great Britain and Northern Ireland

Stewart A Birrell is a Professor of Human Factors for Future Transport within the National Transport Design Centre (ntdc) at Coventry University. He received his PhD in Ergonomics from Loughborough University, UK in 2007, and first-class degree in Sport Science in 2002. Stewart has spent the previous 15 years working within the transportation sector within industry and academia, with expertise ranging from driver behaviour and distraction, multimodal warnings, user state monitoring and information requirements – all underpinned by the design of in-vehicle information systems, and their evaluation using simulators, virtual reality (VR) and field operational trials. Currently, he applies innovative Human Factors Engineering methodologies to enable real-world and virtual evaluation of user interaction with Connected and Autonomous Vehicle (CAV), Electric Vehicle (EV) and Urban Air Mobility (UAM) technologies and services. Professor Birrell has over 100 journal and conference papers, book sections and articles published in his field to date, and is an Editor of the internationally renowned, Q1/4* journal IEEE Transactions on Intelligent Transportation Systems.

ORCID: 0000-0001-8778-4087

Affiliation: Institute for Future Transport and Cities, Coventry University, Coventry, United Kingdom of Great Britain and Northern Ireland

W:https://pureportal.coventry.ac.uk/en/persons/stewart-birrell

In:www.linkedin.com/in/stewart-birrell-62092433

Sc:https://www.scopus.com/authid/detail.uri?authorId=56723781000

Tw:www.twitter.com/DrSBirrell

Michael A. Bromfield is a Senior Lecturer/Associate Professor in Aerospace and a Flight Safety Researcher within the School of Metallurgy and Materials, the University of Birmingham. A Chartered Ergonomist/Human Factors Specialist and Chartered Engineer, he was awarded his Ph.D. from the Department of Mechanical and Aeronautical Engineering at Brunel University, London, UK, in 2012. He is a former technologist apprentice with Westland Helicopters, a trained flight test engineer and a current private pilot. He specialises in the areas of flying qualities, aerospace systems engineering and human factors enabling insight into complex 'human in the loop' systems. His particular interests are Loss of Control In Flight (LOC-I) and future aerospace vehicle design. He is a Member of the Chartered Institute of Ergonomics and Human Factors and a Fellow of the Royal Aeronautical Society. He plays an active role in a number of National, European and International technical and safety committees including NASA, AIAA, EASA, UKFSC and UKVLN and has over 30 publications.

ORCID: 0000-0003-1426-7446

LinkedIn: https://www.linkedin.com/in/michael-bromfield-56627934/

Affiliation: School of Metallurgy and Materials, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom of Great Britain and Northern Ireland

Andrew M. Parkes has a background in Psychology and Human Factors and is active in areas of virtual and augmented reality, transport design and safety. He has been Vice President of the Forum of European Road Safety Research Institutes (FERSI) and Honorary Professor of Life Sciences at Heriot-Watt University, Edinburgh and is currently Adjunct Professor (Research) at Monash University (Faculty of Art, Design and Architecture), Australia. He has

held research positions at Birmingham, Loughborough, Leeds and Coventry Universities and the Transport Research Laboratory in the UK. Interests over his career have expanded from accident causation and investigation, through to a much wider view of the efficiency, acceptability and safety of transport systems and the needs of future cities. He has published over 230 journal articles, book contributions and sponsored reports in the areas of driver behaviour and performance.

ORCID: 0000-0002-3097-0644

LinkedIn: https://www.linkedin.com/in/andrew-parkes-5521a527/

Affiliation: School of Art, Design and Architecture, Monash University, Melbourne, Australia.

Abstract

High levels of automation in future aviation technologies such as Unmanned Aircraft Systems could lead to human operators losing essential Situation Awareness and becoming 'out-of-the-loop'. Research into Human Autonomy Teaming proposes that improved communication between the human and autonomous agents of a system can address this problem. However, knowledge around the effect of automation audio communication is lacking in the literature and we propose audio-voice conversation would provide the optimum form of communication. In this study we evaluated the impact that providing a conversational interface to a synthetic teammate had on the performance, Situation Awareness and perception of teaming of the human teammate. Twenty-four participants conducted experimental trials on a computer-based task adapted from a Levels Of Automation test method developed by Endsley and Kaber (1999). The results show that synthetic voice communication had a significant positive effect on human performance and perception of teaming. Also demonstrated was that teaming structure has an effect on how that performance increases, with participants in higher Levels Of Automation where the automation provides decision making advice demonstrating a habit of consistently following voice provided advice, even when that advice results in the participant adopting new behaviours and taking more risks.

Keywords

Human Autonomy Teaming, Conversational Interface, SA

Introduction

Near future aviation technology, such as the FAA NextGen Unmanned Aircraft Systems (UAS) Traffic Management (UTM), will be designed to be highly automated, with UAS expected to work autonomously in a co-operative manner (Chakrabarty et al. 2019), communicating with a distributed network of highly automated control systems.

The expectation from research into human automation interaction is that the human operators of these highly automated aviation systems, particularly those where the automation conducts most of the processing and decision making, will be able to demonstrate improved performance. However, that performance gain can often come at the cost of a reduction or even loss of human Situation Awareness (Endsley 1995) frequently described as the operator becoming "out-of-the-loop". This degradation or loss of Situation Awareness (SA) then has its own negative knock on effects with operators misinterpreting displayed information, misunderstanding automation activities and decision making, and ultimately can lead to poor human decision making (Kaber and Endsley 1997). It can even result in human's missing that the automation is not working (Endsley 2017).

The solution proposed by researchers (eg Hoc 2001, Klein et al 2004, Johnson et al 2011, Demir, McNeese and Cooke 2017) is to fundamentally change the relationship between the human operator and the automation towards one of teaming in which the human and automation are actively engaged in cooperating and coordinating activities and communicating their respective cognition. The aim is to make the intent, reasoning and predicted future states of the automation transparent to the human operator (Chen et al 2016) and for the automation to overtly provide key awareness and cognitive data that will assist the human with their SA building and ultimately overcome the human "out-of-the-loop" issue. Clear and effective communication is key to achieving this effect (it has been demonstrated repeatedly that improved human team performance is directly dependent upon effective communication eg Marlow et al 2018), with Battiste et al (2018) identifying effective communication as a key tenet for Human Autonomy Teaming (HAT).

Much of the recent HAT research has focused on improving communication between human and synthetic agents using graphical based interaction to pass the data critical to building team SA. Chen et al (2018) describes three HAT projects where the remote synthetic agent provides information on its current location, configuration and intentions using a graphical HCI, as do Schaefer et al (2017) in a similar ground-based transport system. However, neither system address a key requirement identified by de Visser, Pak and Shaw (2018: 1410) that "human–autonomy interaction should emulate the rich interactions of relationships between people and should adopt human–human models as their initial standards". A key human-human communication model is of course audio-voice conversation, with the added benefit that adding an audio-voice to a system will encourage operators to perceived as more human like (Waytz, Heafner and Epley 2014), and that increase in anthropomorphic characteristics would improve a users' trust in it (Przegalinska et al, 2019), all of which would improve the overall likely acceptance and uptake of the desired teaming relationship.

The logical conclusion to all this advice would be to follow McNeese et al (2017), and when designing a HAT implement a conversational interface as a primary form of communication between human and automation. However, even the McNeese et al (2017) conversational interface was visual, with the conversation taking place using a "text chat" application. This emphasis on visual methods of communication potentially misses the opportunities and gains that could achieved through use of audio-verbal communication as additional channel in which to pass specific SA observations, conclusions and predictions and that can be used to direct attention to display details that will assist the human build their own SA. Furthermore, both multiple-resource theory (Wickens 2008) and the multicomponent model (Baddeley 2010) lead us to suggestion that greater cognitive timesharing (for the operator) could be achieved by using audio-voice based communication as an additional channel for information exchange.

Research Hypotheses

Along with other HAT researchers, we surmise that the implementation of a conversational interface will be useful, even essential, to address the "out-of-the-loop" loss of SA and would have a potentially significant positive impact on improving system safety. This will be particularly the case in the new UTM systems where much of the traffic control processing and decision making is likely to be conducted out of sight of the human operator. In fact, the expectations of multiple resource theory lead us to theorise that an audio-voice conversational interface could offer an opportunity to exchange cognitive outputs (conclusions, decisions, recommendations and intentions) aimed at improving the capability and safety behaviour of a HAT at little or no cost in terms reduced cognition or increased workload.

It is hypothesised that the expected improvement in SA would manifest directly as an improvement in the human team-member's SA, which in turn would lead to an increase in task performance without a change in perceived workload; therefore, we will present our hypothesis to be tested in these terms. For the hypotheses we will identify a HAT as a team consisting of at least one human and one synthetic teammate (aka autonomous agent). However, as it is possible to take a near continuous measure of performance, but only possible to take infrequent discrete samples of SA, we will give priority to the performance hypothesis:

> *Hypothesis 1: Human operators in a HAT will demonstrate improved task performance when the synthetic teammate communicates using a combination of audio-voice and graphics over when the synthetic teammate communicates using graphics alone.*
>
> *Hypothesis 2: In a Human Autonomy Team, the human operators will demonstrate improved SA when the synthetic teammate communicates using a combination of audio-voice and graphics over when the synthetic teammate communicates using graphics alone.*

Hypothesis 3: Human operators in a HAT will register reduced subjective workload when the synthetic teammate communicates using a combination of audio-voice and graphics over when the synthetic teammate communicates using graphics alone..

Furthermore, as an added benefit we also expect that the addition of an anthropomorphic voice will have a positive impact on the attitude of the human in the HAT and will assist that human accept the automation as a synthetic teammate and not just a machine:

Hypothesis 4: Voice communication from a synthetic teammate improves the human's subjective perception of teaming in comparison to when communicating with the synthetic teammate through graphics alone.

Materials and Methods

Aim

The primary aim of the study is to evaluate the overall effect of providing an autonomous agent in the form of a synthetic teammate with a voice communication capability. However, human voice communication is extremely complex and varied, and the aim, style and structure of the messages that are communicated can have their own effect on performance and SA, as demonstrated by Demir, McNeese and Cooke (2016). Thus, it would be beneficial if, for this study, it were possible to impose a reasonable standard and structure to the communication (for consistency of participant experience), and also attempt to measure whether some groups or types of communication message have a greater or lesser impact on performance, SA and the perception of teaming.

Form and Structure of Voice Communications

On the surface it would seem logical that, in order for the human participant to be able to suspend dis-belief and engage with the synthetic agent as a team-mate, the voice communication should be as natural as possible. Supporting this view, Murphy et al (2012: 49) observe that "Natural language is the most normal and intuitive way for humans to instruct autonomous systems", and Battiste et al

(2018: 491) remarked that the Pilot participants of their research into Human Autonomy Teaming recommended giving the automation “a better voice interface that uses natural language”.

However, conversely, the experience of the Aviation industry is that natural language is in fact not a reliable and safe format to ensure a successful outcome from the communication (exchange of knowledge). Natural language is full of colloquialisms, slang phrases, poor pronunciation, accelerated speech, and non-standard phraseology all of which can cause serious problems in communication (UK Civil Aviation Authority 2014). The UK Civil Aviation Authority (CAA) in their CAP 737 recommend the use of a more limited and standardised version of voice communication identified as “Plain Language” (UK Civil Aviation Authority 2014). The CAA also identify that the delivery of the communication should be standardised, emphasizes the importance of a stead speech rate, even in situations of high stress, and using clear, short and simple messages.

For this study the synthetic teammate messages will follow this advice of using succinct natural language by keeping messages short, to the point, and presented in a timely manner. Where possible, messages will be designed to convey an average of two items of information (eg message subject and subject activity), although there may be occasions where more information is exchanged. Furthermore, following the advice from Chen at al (2018) the communication will aim to enhance transparency by providing evidence of cognition or a rational for decisions. The rate of communication will also be designed to deliver approximately one message per minute (to coincide with the measurement sample rate, see Methodology below) although this could vary depending upon participant behaviour.

For this research project four simple generic categories of communication message will be used (Table 1) that have been derived from two examples of previous human-human and human-automation communication studies (Bowers et al 1998, and Sorensen and Stanton 2016) and from discussions on

types of aviation Air Navigation Services communication given in the three UK CAA manuals (CAP 413, CAP 774, CAP 1430).

Communication Category	Description
Acknowledgement	Confirm the undertaking or completion of an activity, either physical or cognitive, or the receipt of a piece of knowledge. Used so that the communication recipient can gain SA of the actions and/or knowledge state of the communicator.
Advice	Provide a recommendation for action, either cognitive, physical or communication, if possible with justification, that does not have to be complied with. Usually, but not necessarily, requires an acknowledgement and response of intention.
Information	Provide general information to build Situation Awareness that directs attention but does not necessarily require a response. Similar in scope to the Information communication of Air Navigation Services (CAP 413)
Warning	Provide active information to build SA that directs attention but does not stipulate a response action or even acknowledgement (e.g. "Warning, Fuel Levels Low")

Table 1: Categories of Voice Communication to be used by Synthetic Team-mate

Human Autonomy Team Structure Selection

Early research into Levels Of Automation (LOA) by Endsley and Kaber (1999) provided us with definitions of task-divisions for a range of teaming structures in which the automation or synthetic teammate became increasingly complex and empowered. Reviewing these LOA, it is possible to see that the structure of the team, specifically the complexity of the synthetic agent, will in turn will dictate the style and range of communication implemented. For example, we would not expect that a synthetic teammate at a low LOA would pass complex decision support information such as *advice*; if it were to it would by default have behaved as a medium level "Decision Support" LOA. Thus, implementing a range of teaming structures or LOAs, could provide us the opportunity to establish realistic and engaging teaming conditions in which the range of communication is carefully varied and would allow us to determine whether there is any variance in effect of different types or categories of communication…

Research Conditions

The design of the research was to conduct a Within-Groups evaluation of conversational communication across three conditions: a baseline control condition where the synthetic agent has No audio-voice capability (labelled the "None" condition); the primary condition where the synthetic agent has an audio-voice capability ("Voice" condition), and additional intermediary condition where the synthetic agent communicates via a text messaging window ("Text" condition). The introduction of the Text condition allowed us to determine whether any change in performance and SA was a result of the data being exchanged rather than the mode of delivery.

Furthermore, as it was determined that the LOA would likely affect the voice communication and thus the impact of voice use, it was determined to set LOA as a Between-Groups factor. For this study, four LOA were chosen from the Endsley and Kaber (1999) ten level model that were similar in description to teaming structures commonly found in previous HAT research, as per Table 2:

Team Structure	Structure Definition	Communication Category
Manual Control (LOA 1)	The human has full task control, undertaking all selections, decisions and activities, with the synthetic teammate monitoring and providing basic task failure warnings.	Warning
Playbook (LOA 3)	The human and synthetic teammate share the task, with the human delegating part or full tasks to the synthetic teammate that the synthetic teammate then carries out in a proscribed manner (eg Schaefer et al 2017, Calhoun et al 2018 although their examples are complex). The synthetic provides task warnings as before, but also provides (low) workload warnings and acknowledges receipt of work.	Warning, Acknowledgement, Advice (Warning Solution)
Decision Support (LOA 5)	The synthetic teammate provides pre-processing assistance for human decision making and then the human selects options for task delegation to the synthetic teammate that the synthetic teammate will then carry out (eg CRM tools in	Warning, Acknowledgement, Advice (Warning Solution) Advice (Strategy Selection)

	Battiste et al 2018). The synthetic provides the same messages as before; however, in addition the synthetic provides suggestion for the remaining human activity.	
Supervision (LOA 9)	The synthetic teammate undertakes all tasks and the human monitors the synthetic teammate behaviour, task selection and activity and interjects, "taking-over" when they feel it necessary to do so. The synthetic teammates communication shifts towards providing reports on what and why it is doing what it is.	Information (others given if LOA dropped to LOA 5 or lower by participant)

Table 2: Structures and Behaviours of Human Autonomy Teams to be evaluated

Experimental Design

An experimental methodology was adapted from an Endsley and Kaber (1999) study into LOA, as it provides a highly detailed description of a computer-based experimental apparatus (a system called Multitask) that could be modified and used to evaluate the interactions between humans and synthetic teammates engaged in different teaming structures. The methodology was reviewed for ethical integrity and approved by the Coventry University Ethics board (ethics project reference P86210), and all participants were asked to sign an Informed Consent Form in order to take part in the study.

The task is a simplified and schematic simulation of an air traffic radar (Figure 1) in which a human operator interacts with a synthetic teammate to achieve a dynamic safety critical task. Participants were presented with a continuous flow of square targets that slowly transited from the outside of a radial display towards a central safety zone or deadline.

Targets were presented in three sizes (small, medium and large) and in three colours (red, blue and green) with the size and colour combination indicating the relative importance and risk of each target. The participants were required to clear the targets before they reached the deadline or collided with

each other by either clicking on a target (if in the LOA 1 Manual Control team structure), or by assigning the synthetic teammate to clear it (all other team structures). Targets being cleared were reduced in size until they reached a minimum size and were then removed from the radar screen. The maximum rate of reduction in all team structures was set at one size reduction per second. The size of the target determined how long it would take to clear: small targets took 2s, medium targets 4s and large targets 6s. Only one target could be processed at a time, and a maximum of five targets were present on the screen at any time.

Participants were given reward points for successfully clearing targets and were given penalty points if a target crossed the deadline (impacted with the deadline) or collided with another target before being cleared. The reward and penalty scores for each target was calculated based upon the colour (red, blue or green) and starting size of the target (large, medium, small) as per Table 2 below:

	Colour					
	Red		Blue		Green	
Target Size	Reward	Penalty	Reward	Penalty	Reward	Penalty
Large	20-30	90-100	10-20	80-90	0-10	70-80
Medium	60-70	60-70	50-60	50-60	40-50	40-50
Small	90-100	20-30	80-90	10-20	70-80	0-10

Table 3: Reward and Penalty Points of Targets (reproduced from Endsley & Kaber 1999)

The participants were presented with three goals, two of which were safety orientated and a third that was performance achievement orientated:

- Prevent targets reaching (impacting) the deadline;
- Prevent targets colliding with each other;
- Maximise the score achieved.

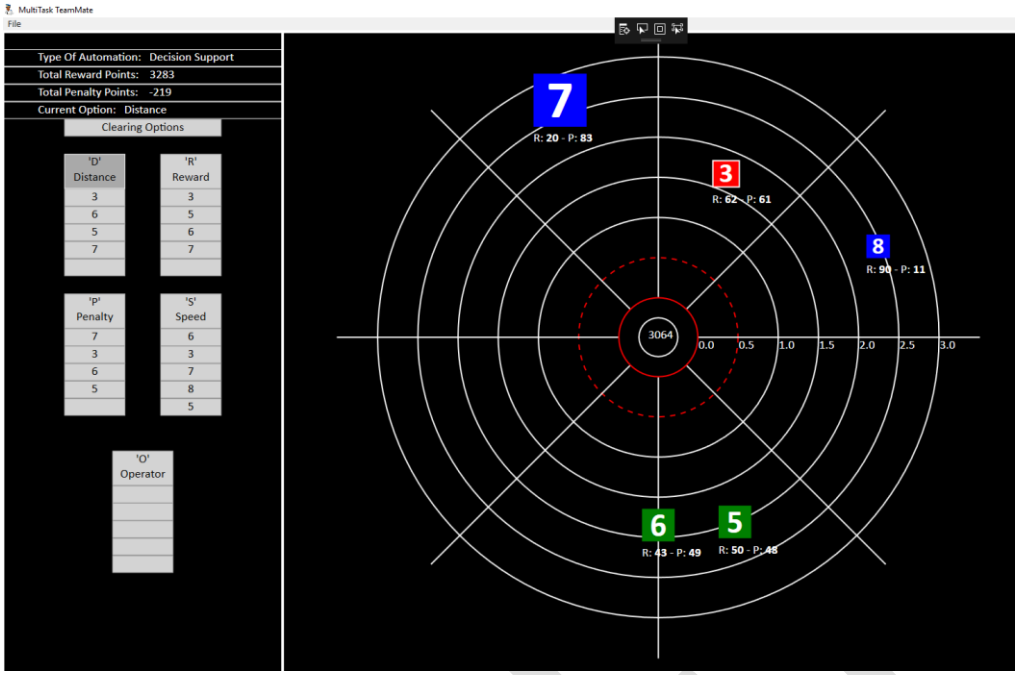

Figure 1: Multitask Team-Mate Display in Decision Support Mode (developed from Endsley & Kaber 1999)

To complete the task the HAT had to a) identify a sequence for clearing targets, b) select the current priority target, c) process the target. Which member of the team the tasks were allocated to was dependent upon the LOA of the team:

- Manual Control (LOA1) – all activities were conducted by the human.
- Playbook (LOA3) – the human carried out a) and b) and the synthetic teammate then completed c).
- Decision Support (LOA5) – the synthetic teammate generated options for a), which the human then selected, and then the synthetic teammate conducted b) and c).
- Supervision (LOA9) – the synthetic teammate carried out all 3 tasks, with the human given the option to over-ride and carry out a).

Equipment

The computer-based simulator[1] was constructed in a Windows Presentation Foundation using C# as the programming language. The performance speed of the simulator was controlled and standardised programmatically across platforms through use of programmed real-time clock timers, and the screen

[1] Source code is available open source from authors.

size was designed to be dynamically scaled up from a minimum size screen of 1024 x 768 pixels (the original 1999 experimental apparatus). The result was a simulator that could be implemented with identical performance rates on any appropriate Windows based desktop computer.

All the trials were conducted using a 23-inch graphical display at 1920 x 1080 resolution. Participants viewing distance was approximately 80 ± 10cm from the screen (depending upon individual posture), with the information graphical display occupying an area 38cm x 24cm, thus giving a vertical viewing angle of 15^{o}-19^{o} and horizontal viewing angle of 24^{o}-30^{o}. Audio voice was provided through a software voice synthesiser allowing the programmatic control of the cadence, tone and emotion of the synthesiser to meet the CRM ideals of a regulated, calm and clear speech. As the experiment was conducted in the UK, the voice synthesiser was set to English (UK) using the MS Hazel (female) set, with a relative speed of 10 (neutral). The audio voice was provided through the built-in mono computer speaker placed on the desktop next to the visual display, with a volume set at 62%.

In the Text communication condition, the text sentence was identical to the voice sentence; the computer simply generated the audio voice sentence but sent it to a text display rather than the voice synthesiser. Text would appear in the text display area for the same duration as it would take the voice synthesiser to speak the audio sentence, thus ensuring data was presented to the participant for equal durations in both Text and Voice conditions.

Participants

Twenty-four voluntary participants aged between 22 and 54 (Mean 35.04, SD 10.243, 9 female and 15 male) took part in the study. Participants were volunteers taken from Coventry University staff (administrative and academic) or students from a wide range of disciplines (eg Social Sciences to Engineering), recruited through a general advertising campaign at the University and were not provided with any financial incentive to participate. Participant education levels ranged from High School graduate to Post-Doctorate. A selection criterion of no previous air traffic control experience

was applied (however no volunteer met this criterion). No participants had visual or auditory impairments or disabilities.

The participants were randomly assigned into four groups of six (with no bias for gender or age), with each group assigned a separate LOA (for measurement of Between-Groups). By chance rather than intent LOA 1, 3 and 9 had 2 female and 4 male, and LOA 5 had 3 female and 3 male. Each participant conducted three Within-Groups trials at their allocated LOA, each trial with a different communication condition (None, Text and Voice). A Latin Square sequence was used to vary the order of the communication conditions for each participant within each LOA. Each trial lasted 10 minutes, in which participants were required to clear an average of 140 targets. The task requirements and target production and movement rates were consistent between all LOAs and communication conditions. Each trial was structured to include three automatic SAGAT interrupts and concluded with two post-activity surveys: a NASA-TLX survey on workload; and a bespoke survey on perception of teaming (Figure 2)

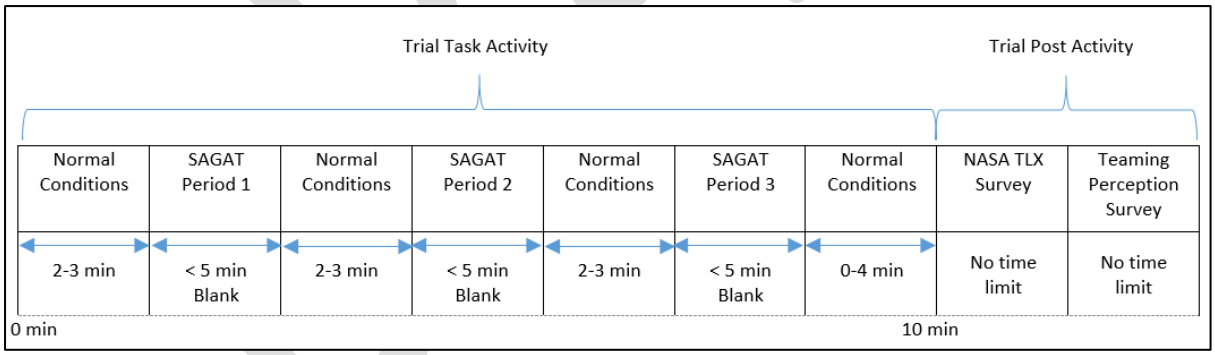

Figure 2: Trial Schedule Including SAGAT Freezes (developed from Endsley & Kaber 1999)

Prior to engagement in the study, participants were instructed on the task goals and the expected activities of the synthetic teammate and, if appropriate, any strategies that the synthetic teammate might use to achieve the goals. Participants were then provided with five minutes of induction training, which included exposure to the SAGAT interrupts and the subjective teaming perception survey. After that they were given 10 minutes of further practice before conducting the three 10-minute long trials. Participants took a 5-minute rest break between each trial.

After the three trails were complete the participants were given a short verbal debrief and asked to give a short written subjective qualitative statement to answer three questions: “Which of the three configurations did you prefer and why?”; “In which of the three configurations did you trust the automation the most & why?”; “What information on the screen did you focus on and why?”.

Results

Data Sampling

In the original LOA study Endsley and Kaber had taken a sample of each performance measure every minute of the simulation giving them 20 samples per trial for each participant. Following the original study design, we determined to gather 10 samples of each measure in each 10-minute trial (and thus gather 60 samples per measure per condition for each LOA subgroup of 6 participants). This sample size gave a sensitivity effect size of 0.23 and an a-priori power of 0.8. It is accepted that taking 10 samples per participant does not overcome the issue of low numbers of participants; however, as this study was to be the pilot study in a much larger research programme, this potential issue was accepted with the expectation that further studies with an increased number of participants would follow to explore any findings in greater detail. The results of these subsequent studies will be published in due course.

Thus, as per the original Endsley and Kaber (1999) study, task performance data was sampled continuously and processed to give an episodic measure for each minute. Three measures of performance were taken for all LOA: the average number of targets cleared per minute (as a measure of success); average processing per minute (activity or work-rate); and average rewards acquired per minute (goal success).

In addition to the three primary performance measures, six selection behaviour measures were also taken to provide further detail on the actions that likely led to a performance. Two of the behaviour

measures provided an indication of the participant's management of risk: proximity to the deadline and proximity to the other targets when cleared. The remaining four measures sampled selection patterns by evaluating the relative (ranked) reward value, penalty value, size and distance from the centre of each cleared target in comparison to the other targets on the screen at the same time (eg if the target was the largest on the screen they would be ranked 1; if the smallest of 5 they would be ranked 5). These measures were also averaged to give an episodic measure for each minute, again providing 10 samples per measure for each participant condition.

It is important at this point to differentiate between the "proximity to deadline" measure from the "relative distance to deadline" measure. The first measures the physical distance (in pixels) from the centre, the second shows whether the target was further in or out compared to the other targets when at that physical distance.

SAGAT questions were taken from the original 1999 Endsley and Kaber study and consisted of asking about target characteristics (size, colour, reward value, penalty value) and target activity (distance, speed and likely impact). Perception of Teaming was measured using a bespoke post-trial survey that asked participants to provide a subjective rating of five characteristics of teaming identified as essential characteristics for synthetic teammates. Perception of Workload was measured using the NASA-TLX survey without pairwise comparison.

All hypotheses data sets tested for normality with the Shapiro-Wilk test, with the majority of sets (57%) successfully evaluated as normally distributed. The distribution of normal/non-normal was not consistent across measures with some measures having a mix of normal/non-normal sets (for example in the measure Distance Rank at LOA1 the None data set failed the test p=.006, but the Text p=.287 and Voice p=.438 data sets passed). As the data sets were relatively large (240 samples per measure overall, splitting into 60 samples per measure for each LOA) advice from Minitab (2015) on

determining whether to use Parametric or Non-Parametric tests and Pallant (2007:204) that "With large enough samples sizes (eg 30+), the violation of this assumption should not cause major problems", was heeded and data was analysed through SPSS for difference using a parametric Repeated Measures mixed ANOVA with the communication conditions *(None, Text, Voice*) as the Within-Groups factor and the LOA (*Manual, Playbook, Decision Support and Supervision)* as the Between-Groups factor. In addition, for Hypothesis 1 (audio-voice improves performance), further Within-Groups Repeated Measure one-way ANOVA were conducted for the behaviour measures for each discrete LOA to determine likely behavioural causes for any differences in performance found in the primary hypothesis tests. For all ANOVA Within-Groups Data sets were tested for Sphericity using Mauchly's test and if failed a Greenhouse-Geisser correction was applied. A pairwise comparison (Bonferroni adjusted) was undertaken for both mixed and one-way ANOVAs.

Confirmation of Adherence With Previous Research

Evaluating the performance Between-Groups on LOA across all conditions provided results consistent with those found by Endsley and Kaber (1999). There was significant variance between LOA for all performance measures: targets cleared ($F(3, 236) = 6.086$, $p = .001$), reward score ($F(3, 236) = 3.163$, $p = .025$), and targets processed ($F(3,236) = 65.955$, $p < .0005$), with the best performance consistently at LOA3, followed by LOA9, with LOA5 only slightly higher than LOA1 (see Figure 3). The results provide confidence that we had correctly interpreted and appropriately modelled the Endsley and Kaber (1999) experimental apparatus.

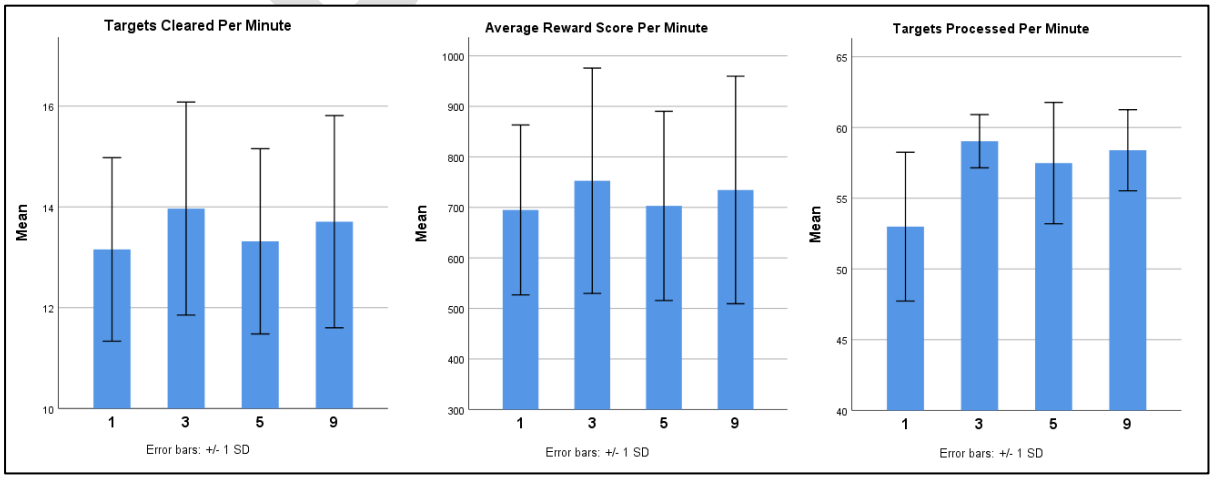

Figure 3: Clearance Performance Measures by LOA (taken across all Voice Conditions)

Hypothesis 1 - Improved Performance with Synthetic Audio-Voice Communication.

Significant difference was found for the within-subjects communication conditions for two of the three performance measures: Targets Cleared ($F(2, 472)=4.140$, $p=.016$); and Targets Processed ($F(1.816, 428.491)=5.817$, $p=.004$). The pairwise comparisons show that for Targets Cleared the primary variance is between the Voice condition (m=13.74, SD=2.074) and None condition (m=13.25, SD=1959, $p=.022$) and for Targets Processed is also between the Voice condition (m=57.54, SD=4.427) and None condition (m=56.53, SD=4.714, $p=.010$). The profile plots (Figure 4) indicate a general rise in performance from None to Text to Voice for all measures, supporting the Hypotheses.

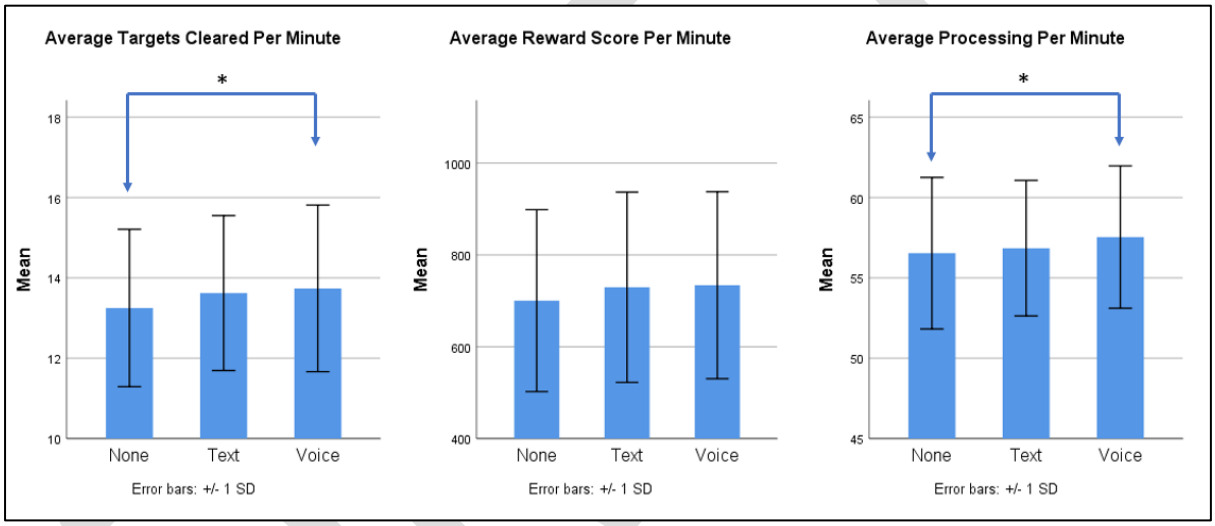

Figure 4: Performance Measures by Voice Condition (taken across all LOA). Statistically significant effects are indicated by asterisks

However, of interest, detailed profile plots showing both Within-Groups (communication) and Between-Groups (LOA) conditions show that the variation in performance by communication is inconsistent between LOA, with Voice generally providing the greatest improvement in LOA1 and LOA5 (apart from target processing where text provided the greatest improvement), but Text providing the greatest improvement in LOA3, and at LOA9 the communication actually marginally degrading performance for both targets cleared and reward score (Figure 5).

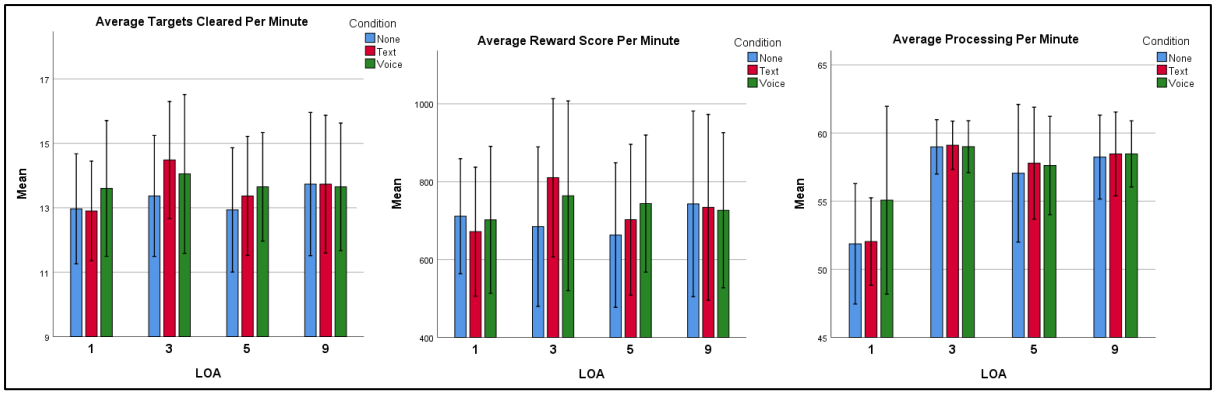

Figure 5: Performance Measure by Communication and LOA conditions

To explore this inconsistency further, the performance measures were separated and examined at the individual LOA. Furthermore, additional risk management and behaviour measures were also analysed to assist identify likely causes for the differences in variation for each LOA.

LOA 1 - Manual Control.

The ANOVA for Manual Control showed significant variance in the performance measure of target processing (F(1.466, 86.497)=6.953, p=.004), with the pairwise comparison identifying the significant variance as between Voice (m=55.07, SD=6.894) and None (m=51.87, SD=4.421, p=.024), with participants in Voice clearing more targets per minute than in None and Text. There was also significant variance in the risk management measure of proximity to deadline (F(1.334, 78.691)=4.493, p=.27), the pairwise comparison showing that the variance was between the Voice (m=185.87,SD=26.74) and Text condition (m=169.28, SD=31.47, p=.042). The graphical plot (Figure 6) shows that in targets were cleared physically closer to the deadline in the None and Text conditions than in the Voice condition.

The behaviour measures penalty rank (F(2, 118)=4.321, p=.015), size rank (F(2,118)=3.448, p=.035), and distance rank (F(1.778, 104.927)=6.290, p=.004) all showed significant variance. Pairwise comparisons identified that the primary difference in penalty rank was between Voice (m=3.311, SD=.434) and None (m=3.130, SD=.419, p=.028), and in distance rank was between Text (m=1.765, SD=.330) and None (m=1.987, SD=.445, p=.011); however, there was no significant pairwise variance in size rank. The profile plots (Fig 6) indicate that participants under the Voice condition selected

larger targets (size rank) with greater penalty scores (penalty rank) what were closer than others to the centre (distance rank).

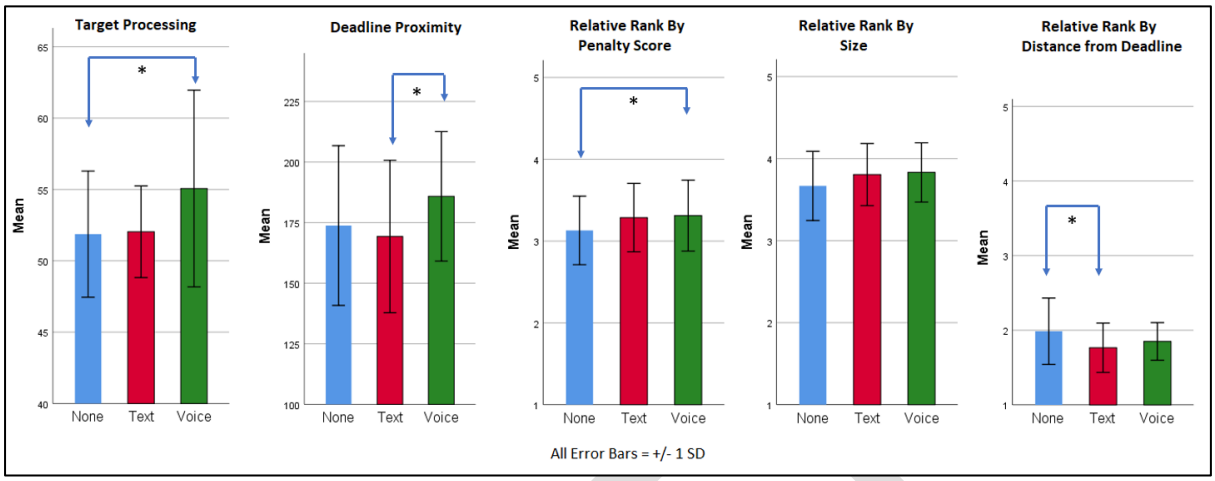

Figure 6: Hypothesis 1 Measures with Significant Variance for LOA 1. Statistically significant effects are indicated by asterisks

LOA3 - Playbook. The ANOVA for the performance measures shows significant variance in the targets cleared ($F(2,118)=4.555$, $p=.012$), with pairwise variance between Text (m=14.48, SD=1.818) and None (m=13.37, SD=1.877, $p=.010$), and in average reward score ($F(2,188)=5.242$, $p=.007$), with the pairwise variance again between Text (m=810.13, SD=203.26) and None (m=684.52, SD=204.95, $p=.008$). In addition, significant variance was found for deadline proximity ($F(2, 118)=6.338$, $p=.002$), with pairwise variance between Text (m=183.60, SD=33.97) and None (m=172.15, SD=25.77, $p=.001$), and for target proximity ($F(2, 118)=4.232$, $p=.017$), with pairwise variance between Text (m=149.71, SD=37.10) and None (m=137.48, SD=36.25, $p=.013$). The profile plots (Figure 7) for the significant measures show that in both Text and Voice, the participants increased the number of targets cleared and total their reward score and tended to clear targets further away from the centre and further away from each other.

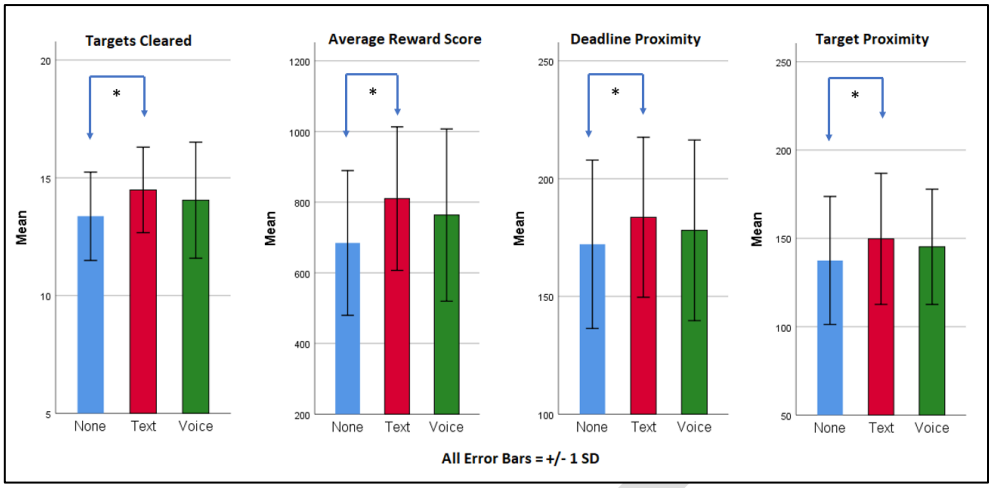

Figure 7: Hypothesis 1 Measures with Significant Variance for LOA 3. Statistically significant effects are indicated by asterisks

LOA 5 – Decision Support. In contrast to the two lower LOA, the ANOVA for Decision Support did not show any significant variance in any of the performance and risk management measures but did show significant variance for all four of the target selection behaviour measures, although the profile plots do show both targets cleared and reward score rising from None to Text and again from Text to Voice. Reward rank ($F(2,118)=6.111$, $p=.003$) demonstrated significant pairwise variance between Voice (m=3.347, SD=.402) and None (m=3.118, SD=.285, $p=.002$), but size rank ($F(2,118)=3.779$, $p=.026$) and penalty rank ($F(2,118)=3.769$, $p=.026$) did not register any pairwise variance. The profile plots of the behaviour measures (Figure 8) indicate that in both the Text and Voice condition (with the greatest effect in Voice) participants tended to select targets that were smaller, had greater reward value and lower penalty value. Finally the distance rank ($F(2,118)=3.670$), $p=.028$) with pairwise variance Voice (m=1.611, SD=.447) and None (m=1.419, SD=.310, $p=.032$) indicates that participants would select targets that were relatively further away from the centre.

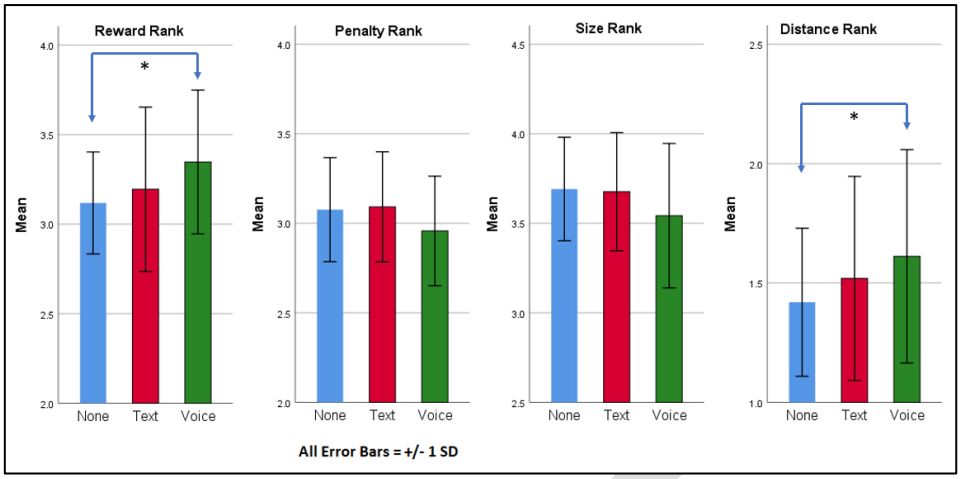

Figure 8: Hypothesis 1 Measures with Significant Variance for LOA 5. Statistically significant effects are indicated by asterisks

The primary difference in the scope of communication for LOA5 (Decision Support) was that the synthetic teammate would provide clearance strategy selection advice in addition to the deadline and target proximity warnings and work receipt acknowledgements of the lower LOA. That advice would, in occasional circumstances where all targets were well clear of the deadline, recommend adopting a strategy to prioritise the clearing of high reward (smaller) targets in order to increase the participants score. It appears that the addition of this advice was sufficient to influence the participants to take a riskier but potentially more rewarding target selection strategy.

LOA9 - Supervision Control. In Supervisory Control the synthetic teammate carries out all tasks following its fixed algorithms, with the consequence that if left undisturbed (ie the human participant did not take over) all measures would remain consistent between conditions and LOA. The ANOVA of LOA9 data did not identify any significant variance for any of the target clearance performance measures, nor for any of the target selection measures. Thus, it would appear that in Supervision Control the participants largely left the synthetic teammate to carry out all tasks without interference, irrespective of whether the synthetic was talking or not.

Hypothesis 2 - Improved SA With Synthetic Audio-Voice Communication.

The mixed ANOVA gave no significant variance in overall Situation Awareness (SA) between the conditions for communication ($F(1.737, 118.110)=.843$, $p=.419$) nor between Levels of Automation (LOA) conditions ($F(3,68)=.534$, $p=.660$). No variance was found for SA Level 1 (perception) nor SA Level 3 (projection), although the profile plots (Figure 9) suggest a potential marginal drop in SA1 and rise in SA3 between the None and Voice conditions. Significant variance was only found for SA Level 2 (comprehension) questions ($F(1.809, 123.045)=3.558$, $p=.031$) although there was no pairwise comparison difference of significance between any of the three communication conditions. Thus, with only one of three SA Levels showing significant variance, and that variance countering the hypothesis, the hypothesis was not supported.

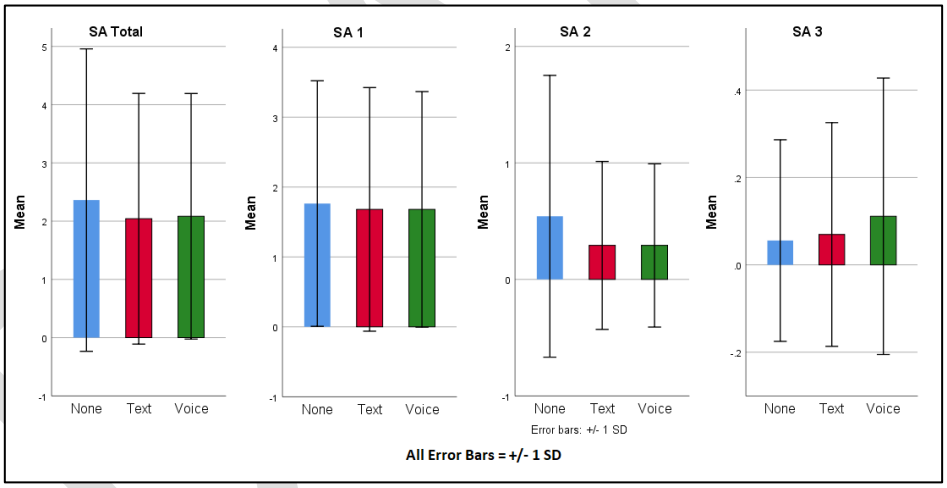

Figure 9: Profile Plots for Hypothesis 2 Measures

Potentially of more interest than the lack of variance in SA is how few results were obtained from the SAGAT questions. The SAGAT questions were generally answered very poorly in all conditions, with 70% of questions in each SAGAT interrupt not answered at all, and of those answered only 28% were correct. Most participants were only able to recall 2-4 bits of SA data (m=2.16, SD=2.290) across all conditions, scoring an average of 8%. This observation leads us to dis-regard the SA results and actually consider the hypotheses not reliably tested rather than not proven.

Hypothesis 3 – Reduced Workload With Synthetic Audio-Voice Communication.

As with the SA results, the mixed ANOVA of the NASA TLX survey also showed no significant variance for communication conditions ($F(2, 40)=.572$, $p=.569$) or LOA ($F(3,20)=2.660$, $p=.076$). Therefore, the hypothesis is not supported. The profile plots for the data sets are provided in Figure 10.

Hypothesis 4 – Improved Perception Of Teaming With Audio-Voice Communication.

The mixed ANOVA of the Perception of Teaming survey results indicates significant variance between the communications conditions ($F(2,40)=14.058$, $p<.0005$), with the pairwise comparison identifying the significant variance as that between the Voice condition (m=19.035, SD=3.451) and the None (m=14.597, SD=4.143, $p=.001$) and Text (m=14.215, SD=4.476, $p<.0005$) but no significant variance between LOA ($F(2,20)=2.949$, $p=.058$) although there is a significant variation with the pairwise comparison between LOA9 (m=18.327,SD=4.410) and LOA3 (m=13.867; SD=4.616, $p=.048$). Thus, the hypothesis is supported.

The profile plot for communication (Fig 10) indicates that adding a text communication capability had no effect on perception of teaming but adding an audio-voice increased the perception scores by approximately a third. Interestingly, whilst the mixed ANOVA shows no significant variance across all LOA, the profile plot of data by LOA does show that greatest effect was in the LOA9 Supervision, the condition where the synthetic was the most “chatty”. Furthermore, in the post-trial interview, the majority of the participants (20) expressed a subjective preference for the Voice condition over the other two conditions. One participant noted that the synthetic team-mate talking was valuable because “it allowed you to hear its thought process”, a direct reference to the hypothesised effect of providing valuable SA data on synthetic teammate cognition.

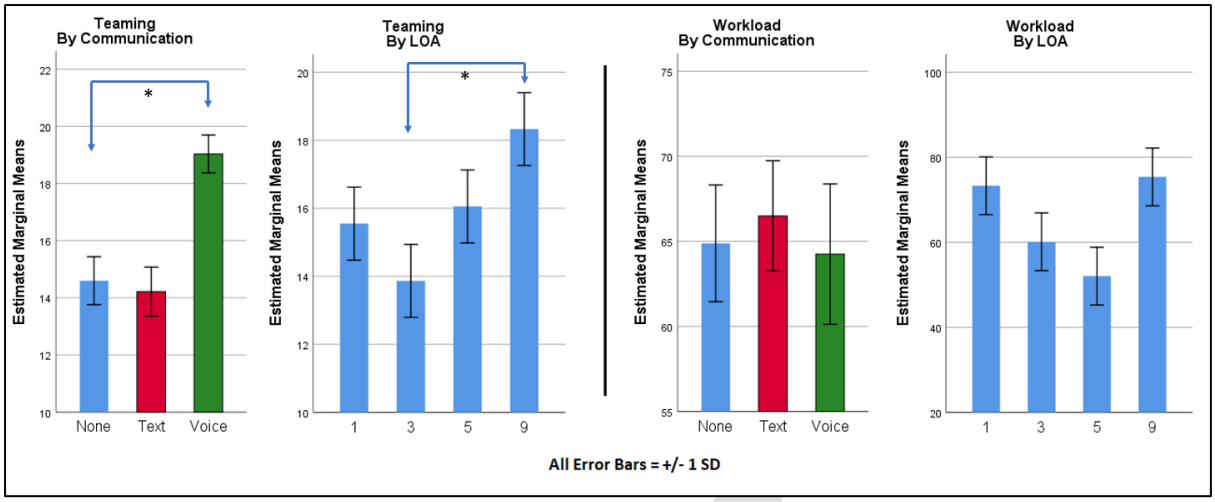

Figure 10 : Profile Plots of Teaming and Workload. Statistically significant effects are shown by asterisks

Discussion

The aim of the study was to provide an experimental scenario in which a human and synthetic team-mate could work together in four different and popular teaming structures or Levels Of Automation (LOA) and evaluate the impact of voice communication from the synthetic teammate on performance, Situation Awareness (SA), perception of teaming and workload. Participant performance could be improved by processing and clearing more targets as possible and gaining a high score. This performance could be further improved by prioritising the clearance of smaller high reward targets that were quicker to clear (thus increasing score and rate of target clearance).

Overall, the results directly support Hypotheses 1 and 4 in that human performance and perception of teaming are positively influenced by providing the synthetic teammate with a conversational communication channel that is used to deliver key SA data to the human. These findings are consistent with those of Chen et al (2016) who observed that improving agent transparency by the deliberate provision of SA data (abet in a graphical form) lead to an increase in operator performance. However, unlike Chen et al (2016) who reported an increase in SA with an increase in transparency, the statistical analysis of our results for SA and workload did not provide any evidence of significant variance to directly support Hypothesis 2 and 3.

Thus, contrary to Endsley's (1995, 40) expectations of a positive correlation between SA and performance "Good SA can therefore be viewed as a factor that will increase the probability of good performance", the results show an increase in performance without an apparent commensurate rise in SA. However, some researchers have been critical of the efficacy of the SAGAT methods (eg Salmon et al 2009) and other attempts to measure the SA of participants in similar dynamic simulations (eg Lo et al 2016 researching Train Traffic Controller) have had similar issues with poor SAGAT scores and in fact even registered a negative correlation between measured SA and task performance. This effect warrants further investigation and our future research will specifically examine this unusual phenomenon.

The pairwise comparisons of the results demonstrate that the significant variance in both performance and perception of teaming occurred between the audio-voice condition and the baseline control condition with no communication, with the text condition giving intermediary results. This indicates that it is not just the message, the extra data, that improves performance, it is also the mode of communication.

However, just as Endsley and Kaber (1999) observed that the participant performance varied with the LOA (and between LOA), so our results show that the effect on performance of adding the voice also varied with the LOA. The results show similarities in behaviour at LOA1 and LOA3; however, they show behaviours in LOA5 and LOA9 are quite dissimilar from each other and from the LOA1 and LOA3.

In LOA1 Manual Control and LOA3 Playbook, the human participant was given an active and leading role, particularly with respect to decision making. In both LOA the participant would both form the decision options (possible sequences of target selection, priorities for clearing targets) and then take the decision. When the synthetic communication was added, either as Text or Voice, the reaction from the participant appears to have been to attempt to work faster (increase in targets processed in

LOA1) or harder (increase in targets cleared in LOA3) in order to increase the distance from the centre at which the targets were cleared (increase in deadline proximity in both LOA1). In both LOA (significant in LOA1, marginal in LOA3) behaviour was consistent with reducing risk by prioritising larger (increase in size rank measure), less rewarding targets (increase in penalty rank measure).

Overall the performance changes at LOA1 and LOA3 indicate that participants actively followed the two safety goals given to the participants of "prevent targets hitting the deadline" or "prevent targets hitting each other" and tried to achieve the third goal "maximise your score" by simply attempting to increase the total number of targets cleared, rather than attempting to increase score through the deliberate prioritisation of high reward targets.

Conversely, in LOA5 Decision Support, where the synthetic teammate took an active part in the decision making and provided advice and recommendations on which clearance strategy to use, the introduction of the voice induced a very different change in behaviour. In the voice condition in LOA5 instead of prioritising the clearance of large high penalty targets, participants prioritised the clearance of smaller high reward targets (show by increase in reward rank and decrease in size rank measures).

The primary difference between the audio voice communications of LOA5 Decision Support compared to the lower LOA was that in LOA5 Decision Support the synthetic teammate provided advice on which clearance strategy to take as well as providing the same risk warnings of the lower LOA. The new advice would occasionally include a recommendation to adopt a more "risky" target selection strategy of selecting targets for clearance based upon their reward value over distance from deadline or each other. Reviewing participant reactions after the synthetic communications shows that participants more often than not heeded the advice from the synthetic (60.2% of all advice was actioned); however, the take up rate of the risk-tolerant advice was strongly affected by the mode of

communication; 53% of risk-taking suggestions in the Text condition were implemented, increasing to 75% in the Voice condition.

This shows that unlike at LOA 1 and LOA 3 where the voice had a positive effect on output, but little effect on changing behaviour away from a "safe" approach, the influence of the voice at LOA 5 was much more profound, encouraging the taking of risks. Simply put, the voice communication appears to have influenced participants more at the higher LOA, compared to the lower LOA. The addition of the voice at the higher LOA appears to have increased participant's trust in the synthetic (identified by their willingness to accept advice), in much the same way that adding a voice to an autonomous car can improve participants trust in the vehicle (Waytz, Heafner and Epley, 2014); however, this effect appears to be absent at the lower LOA.

Finally, a third type of behaviour can be seen under LOA9 Supervision, where the lack of variance in performance, risk or behaviour message across the conditions indicates that irrespective of whether there were messages delivered or not, the participants appeared to largely leave the synthetic teammate alone to take all decisions and carry out all actions.

Recommendation for Further Research

This study was a pilot study for an extended programme of research with a relatively low number of participants per subgroup (6), which must be taken into consideration when reviewing our results and discussion. Future research will need to address by ensuring that there are larger numbers of participants.

This study was primarily focused on evaluating the impact of the synthetic teammate talking to the human operator. The impact of the human operator responding to that communication or initiating audio-voice communication of their own has yet to be evaluated and needs to be researched to

complete our understanding of how a conversational interface can affect SA and performance. Furthermore, research needs to be conducted on the underperformance of the SA me

Conclusions

The original intent of the study was to be able to determine whether the provision of an audio-voice speech capability to a synthetic teammate in a Human Autonomy Team (HAT) could improve the Situational Awareness (SA) of the participant, which would in turn give rise to an improvement in performance. The anticipated improvement in performance was found with significant (positive) variance in two of the three performance measures. However, the measured SA of the participants did not appear to vary, which would suggest that the introduction of the voice did not directly improve the measurable SA in our somewhat novice participant group even though performance was improved. This finding was contrary to our expectation but on reflection was considered more likely a consequence of generic issues with the use of freeze-probe SA measurement techniques rather than an actual incidence of systemic poor SA. With uncertainty over the reliability of the SA measures taken, but acknowledging the expectation of a positive correlation, it is assessed that However, if we are to suppose there is a positive relationship between SA and performance (after all, if not the case, why all the research on how to improve SA?) such that good SA leads to good performance, the improvement in performance observed with the introduction of audio-voice communication can be considered to provide a secondary indicator of a commensurate improvement in SA. Thus, overall, our view is that our results demonstrate that providing an audio-voice communication capability to the synthetic agent improves performance, could help negate loss of SA and as a secondary benefit, promotes the subjective perception of teaming in the human operator.

Also, of value, or of potential concern depending upon your viewpoint, the study has demonstrated that using the synthetic audio-voice capability to deliver advice and recommendations can have a significant impact on the fundamental behaviour of the human operator. In our more detailed analysis

of the effects of introducing the audio-voice communication at different Levels Of Automation (LOA) the influence on the participant behaviour increases with LOA, with the fulcrum of influence and change in behaviour appearing to be the provision of decision-making advice at LOA5 Decision Support. At this LOA5, operators are strongly influenced by the synthetic voice, and we would anticipate they would likely accept synthetic advice given, even if that could result in unsafe behaviour. This could be of great value when the HAT is in a situation where interjection is needed to combat a loss of SA in the human operator, or where the human needs advice to choose between two equal-risk options ('least bad choice situations'); however, care must be taken with the presentation of high-risk options as the results show that operators would be willing to take options they would not normally consider.

However, despite this warning we can conclude from the study results that the addition of a synthetic voice providing task essential information (Warnings), teaming information (Acknowledgements), decision making information (Advice) and teammate activity information (Information) as appropriate to the teaming structure will be of benefit to human operators of complex systems almost irrespective of LOA. This in turn should assist combat the loss of SA or out of the loop situations when in HAT with high levels of automation.

References

Baddeley, A. (2010) 'Working Memory'. *Current Biology* 20 (4), R136-R140. available from <https://www.sciencedirect.com/science/article/pii/S0960982209021332>

Battiste, V., Lachter, J., Brandt, S., Alvarez, A., Strybel, T. Z., and Vu, K. L. (2018) 'Human-Automation Teaming: Lessons Learned and Future Directions'. in *Human Interface and the Management of Information. Information in Applications and Services. HIMI 2018. Lecture Notes in Computer Science, Vol 10905*. ed. by Yamamoto, S. and Mori, H. : Springer, Cham, 479-493

Bowers, C. A., Jentsch, F., Salas, E., Braun, C. C. (1998) 'Analyzing Communication Sequences for Team Training Needs Assessment' *Human Factors* 40 (4), 672-679. available at <doi:10.1518/001872098779649265>

Calhoun, G. L., H. A. Ruff, K. J. Behymer, and E. M. Frost. 2018. "Human-Autonomy Teaming Interface Design Considerations for Multi-Unmanned Vehicle Control." Theoretical Issues in Ergonomics Science 19: 321–352. doi:10.1080/1463922X.2017.1315751

Chakrabarty, A., Ippolito, C., Baculi, J., Krishnakumar, K., Hening, S. 2019 "Vehicle to Vehicle (V2V) communication for Collision avoidance for Multi-copters flying in UTM-TCL4." Paper presented at the AIAA Scitech 2019 Forum, San Diego, January 7-11. doi:10.2514/6.2019-0690

Chen, J. Y. C., Barnes, M. J., Selkowitz, A. R., and Stowers, K. (2016) 'Effects of Agent Transparency on Human-Autonomy Teaming Effectiveness' in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. held 9-12 October 2016 at Hotel Intercontinental Budapest. Red Hook New York: Curran Associates Inc, 001838-001843 available from < https://ieeexplore.ieee.org/abstract/document/7844505 >

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., and Barnes, M. (2018) 'Situation Awareness-Based Agent Transparency and Human-Autonomy Teaming Effectiveness'. *Theoretical Issues in Ergonomics Science* 19 (3), 259. available from <http://www.tandfonline.com/doi/abs/10.1080/1463922X.2017.1315750>

Demir, M., McNeese, N. J., and Cooke, N. J. (2016) 'Team Communication Behaviors of the Human-Automation Teaming'. (eds.) *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. held 21-25 March 2016 at San Diego, California: IEEE. available from <https://ieeexplore.ieee.org/document/7497782>

Demir, M., McNeese, N. J., and Cooke, N. J. (2017) 'Team Situation Awareness within the Context of Human-Autonomy Teaming'. *Cognitive Systems Research* 46, 3-12. available from <https://www.sciencedirect.com/science/article/pii/S1389041716301747>

de Visser, E., Pak, R., Shaw, T. H. (2018) ‘From ‘automation’ to ‘autonomy’: the importance of trust repair in human-machine interaction’. *Ergonomics* 61 (10), 1409-1427. available from <https://www.tandfonline.com/doi/full/10.1080/00140139.2018.1457725>

Endsley, M. R. (1995) 'Toward a Theory of Situation Awareness in Dynamic Systems'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37 (1), 32-64. available from <http://www.ingentaconnect.com/content/hfes/hf/1995/00000037/00000001/art00004>

Endsley, M. R. (2017) ‘From Here to Autonomy: Lessons Learned From Human-Automation Research’. *Human Factors* 59 (1), 5-25. available <https://journals.sagepub.com/doi/10.1177/0018720816681350>

Endsley, M. R. and Kaber, D. B. (1999) 'Level of Automation Effects on Performance, Situation Awareness and Workload in a Dynamic Control Task'. *Ergonomics* 42 (3), 462-492. available from <http://www.tandfonline.com/doi/abs/10.1080/001401399185595>

Kaber, D. B., and Endsley, M. R. (1997) ‘Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety’. Process Safety Progress 16 (3), 126 – 131. available from <https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/prs.680160304>

Hoc, J. (2001) 'Towards a Cognitive Approach to Human–machine Cooperation in Dynamic Situations'. *International Journal of Human - Computer Studies* 54 (4), 509-540. available from <https://www.sciencedirect.com/science/article/pii/S1071581900904543>

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Hoffman, R. R., Jonker, C., van Riemsdijk, B., Sierhuls, M. (2011) ‘Beyond Cooperative Robotics: The Central Role of Interdependence in Coactive Design’. *IEEE Intelligent Systems* 26 (3), 81-88. available from <https://ieeexplore.ieee.org/abstract/document/5898449>

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. (2004) 'Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity'. *IEEE Intelligent Systems* 19 (6), 91-95. available from <http://ieeexplore.ieee.org/document/1363742>

Lo, J. C., Sehic, E., Brookhuis, K. A., and Meijer, S. A. 2016 “Explicit Or Implicit Situation Awareness? Measuring the Situation Awareness of Train Traffic Controllers”. Transportation Research. Part F: Traffic Psychology and Behaviour 43: 325-338 doi:10.1016/j.trf.2016.09.006

Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., Salas, E. (2018) ‘Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance’. *Organizational Behavior and Human Decision Process* 144, 145-170. available at <https://www.sciencedirect.com/science/article/pii/S074959781630125X>

McNeese, N. J., Demir, M., Cooke, N. J., and Myers, C. (2017) 'Teaming with a Synthetic Teammate: Insights into Human-Autonomy Teaming'. *Human Factors: The Journal of Human Factors and Ergonomics Society* 60 (2), 262-273. available from <https://journals.sagepub.com/doi/full/10.1177/0018720817743223>

Murphy, R., Shields, J., Schmorrow, D., Appleby, B., Howe, A., Israel, K., Livanos, A., McCarthy, J., Mooney, R., Nathman, J., Parker, K., Tenney, R., and Woods, D. (2012) *The Role of Autonomy in DoD Systems*: Defence Science Board. available from <https://fas.org/irp/agency/dod/dsb/autonomy.pdf>

Pallant J. 2007 “SPSS survival manual, a step by step guide to data analysis using SPSS for windows. 3 ed.” Sydney: McGraw Hill

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., Mazurek, G. (2019) ‘In bot we trust: A new methodology of chatbot performance measures’. *Business Horizons* 62, 785-797. available from < https://www.sciencedirect.com/science/article/pii/S000768131930117X>

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., and Young, M. 2009 “Measuring Situation Awareness in Complex Systems: Comparison of Measures Study.” International Journal of Industrial Ergonomics 39 (3): 490-500. doi: 10.1016/j.ergon.2008.10.010

Schaefer, K. E., Straub, E. R., Chen, J. Y. C., Putney, J., and Evans, A. W. (2017) 'Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human-Agent Teams'.

Cognitive Systems Research 46, 26-39. available from <https://www.sciencedirect.com/science/article/pii/S1389041716301802>

Sorensen, L. J., Stanton, N. A. (2016) 'Keeping it together: The role of transactional situation awareness in team performance' *International Journal of Industrial Ergonomics* 53, 267-273. available at https://www.sciencedirect.com/science/article/pii/S0169814116300117

The Minitab Blog 2015, Choosing Between a Nonparametric Test and a Parametric Test [online] available from < https://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test

UK Civil Aviation Authority (2014), *CAP 737: Flightcrew Human Factors Handbook*. Gatwick Airport South: UK CAA. available from <http://publicapps.caa.co.uk/modalapplication.aspx?catid=1&pagetype=65&appid=11&mode=detail&id=6480>

Waytz, A., Heafner, J., Epley, N. (2014) 'The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle'. *Journal of Experimental Social Psychology* 53, 113-117. available from < https://www.sciencedirect.com/science/article/pii/S0022103114000067>

Wickens, C. D. (2008) 'Multiple Resources and Mental Workload'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50 (3), 449-455. available from <https://journals.sagepub.com/doi/abs/10.1518/001872008X288394>