Coventry University

Coventry University

DOCTOR OF PHILOSOPHY

Improving Human Autonomy Teaming Efficacy Through a Voice Communication Interface

Bogg, Adam

Award date:
2022

Awarding institution:
Coventry University

[Link to publication](#)

Improving Human Autonomy Teaming Efficacy Through a Voice Communication Interface

By

Adam Bogg

PhD

August 2021

Improving Human Autonomy Teaming Efficacy Through a Voice Communication Interface

Adam Bogg

A thesis submitted in partial fulfilment of the University's

requirements for the Degree of Doctor of Philosophy

August 2021

Coventry University

Certificate of Ethical Approval

Applicant:

Adam Bogg

Project Title:

Development of Operator Situational Awareness For Autonomous Unmanned Aerial Vehicle Systems

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:

20 March 2018

Project Reference Number:

P69106

Ethics Certificate – P86210

Certificate of Ethical Approval

Applicant:

Adam Bogg

Project Title:

Effect of Automation Voice Communication on Situation Awareness of Human Operators of Autonomous UAS Management Systems

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Medium Risk

Date of approval:

26 February 2019

Project Reference Number:

P86210

Ethics Certificate – P93432

Ethics Certificate – P109048

Coventry University

Certificate of Ethical Approval

Applicant: Adam Bogg

Project Title: Unmanned Traffic Management System - Conversational
                                     Interface

This is to certify that the above named applicant has completed the Coventry University Ethical
Approval process and their project has been confirmed and approved as Medium Risk

Date of approval: 14 Oct 2020

Project Reference Number: P109048

Candidates Declaration Form – Section 3

Section 3 Submission Declaration

	Yes	No
Have materials contained in your thesis/submission been used for any other submission for an academic award?	☐	☒

If you have answered Yes to <u>above</u> please state award and awarding body and list the material:

	Agree	Disagree
To the best of my knowledge, there are no health reasons that will prevent me from undertaking and completing this assessment and I will ensure to notify my Director of Studies and the Doctoral College if there is any change to these circumstances	☒	☐

	Yes	No
Ethical Declaration: I declare that my research has full University Ethical approval and evidence of this has been included within my thesis/submission. Please also insert ethics reference number below Project Reference: P69106, P86210, P93432, P109048	☒	☐

Freedom of Information:

Freedom of Information Act 2000 (FOIA) ensures access to any information held by Coventry University, including theses, unless an exception or exceptional circumstances apply.

In the interest of scholarship, theses of the University are normally made freely available online in CURVE, the Institutions Repository, immediately on deposit. You may wish to restrict access to your thesis for a period of three years. Reasons for restricting access to the electronic thesis should be derived from exemptions under FOIA. (Please also refer to the University Regulations Section 8.12.5)

Do you wish to restrict access to thesis/submission: No

If <u>Yes</u> please specify reason for restriction:

Does any organisation, other than Coventry University, have an interest in the Intellectual Property Rights to your work?
                    No

If <u>Yes</u> please specify Organisation:

Please specify the nature of their interest:

Candidates Signature:		**Date:** **29 June 2021**

Declaration of Authorship

Work presented in this thesis has been published in a peer reviewed journal, and presented at an international conference. The content of that work has contributed towards the thesis in the Pilot Study (Chapter 4), the Introduction (Chapter 1) and Methodology (Chapter 3).
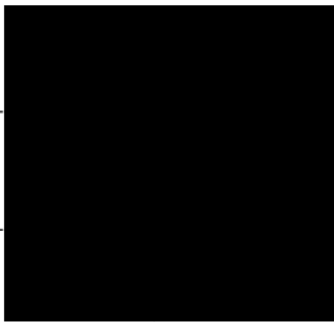
I declare that the two papers detailed below are original works and that I was the primary and lead author, with my supervisors as my co-authors providing contributions through study supervision and editorial direction.

Name: Adam Bogg	Signature:
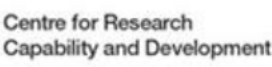
Paper 1:   Bogg, A., Birrell, S., Bromfield, M.A., and Parkes, A.M. (2020) 'Can we talk? How a talking agent can improve human autonomy team performance'. *Theoretical Issues in Ergonomics Science*, 22 (4) 488-509. available from <doi.org/10.1080/1463922X.2020.1827080>

Paper 2:   Bogg, A., Parkes, A., and Bromfield, M. (2020) 'Can we talk?–the impact of conversational interfaces on human autonomy teaming perception, performance and situation awareness'. in Ahram, T., Karwowski, W., Vergnano, A., Leali, F., and Taiar, R. (eds.) *International Conference on Intelligent Human Systems Integration* held 19-21 February 2020 at Modena, Italy. Cham: Springer, 938-944. available from <doi.org/10.1007/978-3-030-39512-4_143>

I agree that Adam Bogg was the primary and lead author of both articles with our contributions from study supervision and editorial direction.

Co-author Signatures	
Name: Andrew Parkes	
Name: Stewart Birrell	
Name: Mike Bromfield	

Declaration of Authorship

Library Declaration and Deposit Agreement

	Coventry University	Doctoral College | Centre for Research Capability and Development	**Library Declaration and Deposit Agreement**

Section 1: Candidate Information			
PGR ID:	**Forename:**		**Family Name:**
8321851	Adam		Bogg

Section 2: Research Details	
Faculty/URC:	Faculty of Engineering, Environment and Computing
Award:	**PhD**
Thesis Title:	**Improving Human Autonomy Teaming Efficacy Through a Voice Communication Interface**

Freedom of Information:	
Freedom of Information Act 2000 (FOIA) ensures access to any information held by Coventry University, including theses, unless an exception or exceptional circumstances apply. In the interest of scholarship, theses of the University are normally made freely available online in the Institutions Repository, immediately on deposit. You may wish to restrict access to your thesis for a period of up to five years. Reasons for restricting access to the electronic thesis should be derived from exemptions under FOIA. (Please also refer to the University Regulations Section 8.11.10)	
Do you wish to restrict access to thesis/submission:	No
Please note: If your thesis includes your publications in the appendix, please ensure you seek approval from the publisher first, and include their approval with this form. If they have not given approval, they will need to be removed from the version of your thesis made available in the Institutional Repository.	
If Yes please specify reason for restriction:	
Length of restriction:	
Does any organisation, other than Coventry University, have an interest in the Intellectual Property Rights to your work?	No
If Yes please specify Organisation:	
Please specify the nature of their interest:	
Signature:	Date: 21 January 2022

For Doctoral College and Centre for Research Capability and Development use	
Date Final Thesis Submitted	
Date of Thesis release to Library	

Abstract

Continuous advancements in information technology have significantly changed the scale of use and scope of application of automation, to the point where future applied technologies, such as Unmanned Aircraft Vehicle Traffic Management systems, will be critically dependent upon integral intelligent automation capability. Whilst the implementation of intelligent automation generally results in an increase in system and operator performance and capability, research has shown that the high levels of automation can lead to operators losing essential environmental or system Situation Awareness (SA) and becoming 'out-of-the-loop' and at risk of missing or making safety critical errors.

Research into Human Autonomy Teaming (HAT) proposes that improved communication between the human and automation of a system can address this problem. HAT proposes changing the relationship between the human and automation away from a human using a machine to one of a human teaming with an autonomous synthetic agent. Typically, experimental platforms for research into HAT communication use graphical or text-based communication channels, with the result that knowledge is lacking in the literature about the effect of implementing speech or audio-voice communication on human team-member SA, performance and perception of teaming.

This research project proposes that speech conversation delivered through the audio-voice medium would provide the optimum form of communication for a HAT. Audio-voice communication would offer many human factors advantages, including ease of use, focused exchange of SA information, cognitive timesharing, and anthropomorphic appeal.

The aim of this research project is to investigate the effect the use of audio-voice intra-team communication has on the SA, performance, and subjective perception of teaming of the human team-member. The objective of the research is to provide consolidated advice and guidance on

the use of audio-voice based communication that can be used for future systems interface design.

The project initiated with a review of research literature on SA that exposed a potential issue in core SA theory and resulted in the determination to utilise decision-making results as a secondary measure of SA to support evidence gathered using the SAGAT methodology. This was followed by a qualitative review of key factors identified as likely to impact on the implementation and structure of audio-voice communication such as direction of communication, teaming structure, evidence of cognition or reasoning, and automation degradation. The outcome of the initial review was a proposed framework for designing and implementing HAT messages drawing upon the aviation best practice human-human teaming model of Crew Resource Management (CRM).

To conduct the investigation into the effect of audio-voice communication, a series of three experimental studies were conducted during which the key factors identified were systematically varied using an aviation-based traffic control simulator. The three experiments consisted of:

1. A pilot study to evaluate the effect of synthetic agent communication as either audio or text messages against a range of teaming structures based upon four Levels Of Automation (LOA);

2. A study into the specific effect of the human speaking to the synthetic agent;

3. A final study examining the impact of increasing the synthetic agent message content to include evidence of cognition and reasoning.

The findings of all three studies were essentially positive, provided measured results and evidence that synthetic agent voice communication can have a significant positive effect on human SA, performance, decision-making and perception of teaming, improving human trust of the synthetic agent and influencing human decision-making. Importantly, it was found that

whilst providing the autonomy with a human voice can encourage humans to adopt synthetic agent recommendations, the additional provision of reasoning information embedded within the synthetic voice messages ensured human trust was optimised, preventing over-trust and unsafe behaviour.

However, the studies also provided unexpected negative findings on the use of SAGAT, with the method consistently providing apparently unusually low absolute measures and values of SA that largely contradict the positive performance and decision-making findings. Practical and theoretical analysis of these findings raise questions about the validity of the SAGAT methodology and, by extension, the SA theory that it is based upon. The same analysis supports using measures of decision-making as a more accurate and extensive approach to evaluating SA. It is recommended that future research continues to examine this approach, especially in the context of complex systems of multiple human and synthetic actors.

The conclusion of the research project is that cumulatively, the results of the studies indicate that there are significant positive advantages to providing a HAT with an audio-voice communication interface that lever off the best practice methodologies of human-human team systems such as CRM, and that audio-voice communication can improve overall systems safety by assisting the human team-member build SA and make more balanced and considered decisions.

Acknowledgements

My most sincere thanks and gratitude got first and foremost to my supervisory team: Prof Andrew Parkes for his belief in me, his tenacity in getting me accepted to the programme and his astute advice on the reason and meaning of a PhD; Dr Dayle Richards for helping me "see the gap" and setting the scope of the project; Dr Mike Bromfield for his support and advice on aviation technology, human factors and presentation preparation; Dr Thomas Statheros for his technical insight and advice on study design; but, most of all, to Prof Stewart Birrell who is without doubt a superlative PhD supervisor and researcher who not only provided outstanding professional advice and guidance, but also was an exceptional mentor providing much needed support and guidance during the extraordinary days of 2020 and 2021. And thanks for encouraging me to submit that paper!

I also wish to thank all my colleagues and fellow researchers at the ntdc, IFTC and Coventry University, be they undergrad, post-grad or post-doc who provided such excellent support and an unusual willingness to be participants in the simulator studies. Worthy of particular mention is Dr Allison Duncan who provided much sage advice on the life of a researcher and who was always willing to be impressed by my results and Mr Lee Mansell for introducing me to the world of UAV detection.

And of course, finally and most importantly, I wish to dedicate this whole thesis and effort to my wife Gill who has put up with four years of enthusiasm and "tech-speak" and who has apparently an endless supply of patience. You gave me time and my dream and for that I am eternally grateful.

Table of Contents

Contents

Glossary of Acronyms

Acronym	Term
AF	Air France
AP	Auto-Pilot
BEA	Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile
CAA	(United Kingdom) Civil Aviation Authority
CAP	Civil Aviation Publication
CRM	Crew Resource Management
C-UAS	Counter Unmanned Aircraft Systems
DSA	Distributed Situation Awareness
DV	Dependent Variable
FAA	(United States) Federal Aviation Administration
HAT	Human Autonomy Team
IV	Independent Variable
LOA	Level Of Automation
MS	MicroSoft
NASA	National Aeronautics and Space Agency
QASA	Quantitative Analysis of Situation Awareness
RO	Research Objective
SA	Situation Awareness
SAGAT	Situation Awareness Global Assessment Technique
SART	Situation Awareness Rating Technique
SA-SWORD	Situation Awareness Subjective Workload Dominance
SPAM	Situation Present Assessment Method
TSA	Team Situation Awareness
UAS	Unmanned Aircraft Systems
UAV	Unmanned Aviation Vehicle
UTM	Unmanned aircraft system Traffic Management
WOO	Wizard Of Oz

Chapter 1 - Introduction

1.1   Background and Rational

Near future aviation systems and technology, such as the Federal Aviation Authority (FAA) NextGen Unmanned Aircraft Systems (UAS) Traffic Management (UTM) (FAA 2020), will be designed to be highly automated, with both the UTM and client UAS expected to work near autonomously in a co-operative manner (Chakrabarty et al. 2019). Human's will still interact with these autonomous systems, generally serving as strategic supervisory controllers responsible for overseeing the performance of the autonomy, normally with the additional role of intervening and providing direction in situations that the autonomy software cannot handle (Endsley 2017).

The positive expectation from research into human automation interaction is that implementing these highly automated aviation systems, particularly those where the automation conducts a large proportion of the processing and decision making, is likely to expand system performance, reduce human workload, and mitigate risk to humans (Demir et al. 2019b). However, the transfer of processing and decision making from the human to the automation is also increasing the potential for human operators of these complex systems to lose system and environmental Situation Awareness (SA). This loss of SA can lead to operators become dissociated from the internal workings of the systems, an effect identified as becoming "out-of-the-loop" (Kaber and Endsley 1997). The loss of SA then has its own negative knock-on effects increasing the probability of the human operators misinterpreting displays leading to a lack of comprehension of the meaning of the displayed information and ultimately leads to poor decision making in process and systems control (Kaber and Endsley 1997). If the system then suffers failure and the "out-of-the-loop" operator may not realise that the automation is acting incorrectly or understand why (Endsley 2017), the consequences can be unexpected, catastrophic or even fatal.

The sensor failure that caused automation degradation in the fatal AF 447 crash on 31 May 2009 is a good example of how poor transparency on automation in unusual circumstances can lead to a loss of SA and ultimately trigger a terminal chain of events. On AF 447 the autopilot disconnected after a pilot static tube froze over; however, the autopilot provided no warning nor reason for the disconnect. The post incident report explains that "the crew, at this time, did not know why the AP [auto-pilot] had disconnected and the new situation that had suddenly arisen clearly surprised the pilots". The report explains that key situation diagnostic information was not directly supplied and that "the philosophy of both the manufacturer and the operator is for the crew to look for additional information necessary to understand the problem and take action" (BEA 2012: 174). The pilots on AF 447 took 10 seconds to identify the fault likely to have caused the autopilot to disconnect, but even after that successful diagnosis, when trying to gather in more information from the aircrafts' displays (the Electronic Centralised Aircraft Monitor or ECAM) "the crew was unable to identify any logical link between the symptoms perceived and these ECAM messages". It would thus appear that when the autopilot disconnected the pilots instantly lost SA of the aircraft systems, were thrown out-of-the-loop, and catastrophically, were unable to regain and rebuild effective SA.

The solution proposed by researchers (eg Hoc 2001, Miller and Parasuraman 2003) to address the loss of SA about the system and automation is to fundamentally change the relationship between the human operator and the automation. Whilst originally used to argue for complex Levels of Automation (LOA) (eg Miller and Parasuraman 2007) the concept has more recently been to move the human relationship with automation away from one where computers are used as a tool (Demir et al. 2019b) towards one of teaming in which the human and automation work collaboratively, The human and automation work together to successfully achieve common goals (Demir, McNeese and Cooke 2018) and share work and information about their decision making to facilitate shared understanding (Chen et al. 2018) to build SA. The

automation is designed to appear to a human observer to be a recognisable agent capable of autonomous behaviour; effectively the automation becomes an autonomous synthetic agent.

In the solution the human and autonomous synthetic agent form a team, a Human Autonomy Team (HAT) and work together, collaborating and coordinating to accomplish system and task goals (Strybel et al. 2017). Like a human-human team, the team members work together to achieve shared mission goals, and its effectiveness is grounded in team situation awareness, team cognition and teamwork (Tokadlı and Dorneich 2019). The expectation is that the focus on improving sharing and improving transparency of the autonomous synthetic agent processing not only improves human situation knowledge but also provides an improved but balanced trust in the synthetic agent with the added benefit of addressing the misuse, disuse and abuse issues identified by Parasuraman and Riley (1997). This expectation has been confirmed through research that has demonstrated that improve system transparency has a positive impact on the development of the human operator situation awareness and trust in the system (eg Selkowitz, Lakhmani and Chen 2017, Chen et al. 2018, Guznov et al. 2020).

Clear and effective communication between all team members, human and synthetic, is crucial for building effective SA (Demir et al. 2017) in the team and of the team and is the foundation to achieve other team processes such as coordination and cooperation (Wynne and Lyons 2018). It has been demonstrated that effective communication is a positively and significantly related to team performance (Marlow et al. 2018), and other researchers (eg Battiste et al. 2018) have identified effective communication as a key requirement for HAT. Indeed, many HAT researchers have for some time been actively engaged in studying how improving communication between the human and synthetic agents of a system, can improve the SA of the human and generally improve the overall efficacy and efficiency, and most importantly, improve the critical safety of the total system (eg Chen et al. 2014, Demir, McNeese and Cooke 2016, Eriksson and Stanton 2017, Guznov et al. 2020).

A key, and arguably the primary human-human communication method is of course audio-voice conversation. Hypothetically, replicating such a highly human centric method in a HAT will not only provide a method to achieve the data exchange requirements of communication, but giving it a human voice would also help operators to anthropomorphise the synthetic agent (Waytz, Heafner and Epley 2014) that anthropomorphisation increasing the operators trust in the autonomy (Przegalinska et al. 2019) and improve the overall teaming effect desired of HAT.

Thus, it would appear logical that the ideal solution to address the loss of SA, HAT communication requirements, and create the perception of teaming is to implement an auditory speech or conversational interface. However, to date much of the HAT research into team communication has been conducted with the communication achieved through the visual channel, via graphical interfaces. In a meta study into experimental HAT research covering some 76 studies, O'Neill et al. (2020) identified 10 articles that discussed communication achieved using visual representation, 39 articles where communication was achieved using text-based chat and just two where the audio-vocal channel was used.

This emphasis on visual methods of communication potentially misses the opportunities and gains that could achieved through use of audio-verbal communication as an additional channel in which to pass specific SA observations, conclusions and predictions. Alternatively, it could be used to direct attention to details in the graphical display that will assist the human build their own SA. It would also provide an increased opportunity to anthropomorphise the synthetic agent and thus improve trust and engagement. It would take advantages of the opportunities for improved cognitive capability and thus performance arising from improved cognitive timesharing (for the human) as proposed in multiple-resource theory (Wickens 2008) and the multicomponent model (Baddeley 2010).

1.2   Research Gap, Questions and Objectives

The low incidence of research into use of audio-voice offers a research gap and research opportunity. The research gap, as identified within this thesis and confirmed by O'Neill et al. (2020), is of a requirement for research on the effects of implementing speech communication between the human and synthetic agent of a HAT through an audio-voice channel. The research opportunity is to seek to determine if the use of the audio-voice communication to transfer SA data between team members would improve the SA of the human. The research would also determine whether that improvement in SA would result in the human improving the quality and safety of their decision-making and thus improve the quality and safety of task completion (ie quality of their performance). The opportunity is to provide empirical evidence for the possible benefits of implementing an audio-voice communication capability in an aviation HAT setting

Furthermore, the identified research gap also provides an opportunity to contribute further to fundamental research on the implementation of HAT by evaluating whether the implementation of an audio-voice communication channel would achieve some of the human-human teaming and anthropomorphic requirements identified by researchers (eg Groom and Nass 2007, de Visser et al. 2012, Wynne and Lyons 2018) as necessary for the human-autonomy teaming effect to exist.

The gap and two opportunities provide the underpinning Research Question:

- *"Would providing a highly automated synthetic agent of a Human Autonomy Team with a audio-voice conversational interface have a significant impact upon the situation awareness and task performance of the human, and facilitate the creation of a teaming relationship between the human and synthetic agent?"*

Reflecting upon the context of the loss of SA and "out-of-the-loop" problem discussed in the background, and on the possible solution of HAT, five more specific research questions derived

5

from this key question are proposed, each question addressing a factor that the researcher considers might affect the outcome of implementing an audio voice channel:

RQ 1. *Would the presence of an audio-voice conversational capability facilitate the creation of a human-autonomy team, and provide the safety enhancing benefits of teaming that address human "out-of-the-loop" safety concerns? Specifically:*

   a. *Would the presentation of information as audio-voice messages affect the situation awareness of the human operator?*

   b. *Would the presence of audio-voice capability affect the output or task completion performance of the human operator?*

   c. *Would the presentation of information as audio-voice messages affect the decision-making behaviour of the human operator?*

   d. *Would the presence of audio-voice capability affect the workload of the human operator?*

   e. *Would the presence of audio-voice capability facilitate the acceptance, by the human, of the automation as a teammate?*

RQ 2. *Could the teaming structure or Level Of Autonomation (LOA) of the system lead to changes in communication between team-members that would influence any change in teaming perception, performance, situation awareness and workload?*

RQ 3. *Would making the communications more transparent and anthropomorphic by including evidence of reasoning such as explanations for deduction and decisions change the human response to the communication?*

RQ 4. *Would the implementation of audio conversational capability prepare the human to successfully and safety take-over tasks during periods of automation degradation?*

To answer the detailed research questions, it is intended to conduct a programme of research that includes a series of experimental human factors studies. The objectives of the research programme are as follows:

RO 1. Conduct a review of research literature on human situation awareness, human autonomy teaming and the implementation of audio-voice interfaces between human and synthetic agent to increase researcher knowledge and identify key experimental design factors and constraints to be accounted for in subsequent studies.

RO 2. Prepare a design framework for generating functional task-orientated synthetic audio-voice communication messages based upon aviation best practice.

RO 3. Select a methodology for evaluating and measuring human operator SA, performance and teaming.

RO 4. Design and build simulated aviation applications complete with a synthetic agent capable or reacting to real-time events and speech synthesis and voice recognition.

RO 5. Evaluate the impact of audio-voice communication from a synthetic agent on human SA, performance and perception of teaming when conducting a dynamic task.

RO 6. Evaluate the impact of audio-voice communication from a human to a synthetic agent on the SA, performance and perception of teaming when conducting a dynamic task.

RO 7. Provide scientific evidence-based knowledge on the possible benefits of using an auditory speech communication interface between the human and autonomous agent in an aviation HAT setting

1.3   Research Aim & Hypotheses

The aim of this research programme is to experimentally evaluate whether implementing an audio-voice speech communication capability in a system that included human operators and autonomous synthetic agents would help solve some of the issues associated with operators becoming "out-of-the-loop". It was also to assess whether the implementation of the auditory

communication could contribute significantly towards creating the teaming effect required for HAT.

1.3.2 General Research Hypotheses

From the primary research questions three general hypotheses emerge:

Hypothesis 1:   In a Human Autonomy Team (HAT), the human operators will demonstrate improved Situation Awareness (SA) when the human and autonomous synthetic agent communicate using a combination of audio-voice and graphics over when the human and autonomous agent communicate using graphics alone.

Hypothesis 2:   Human operators of a HAT will demonstrate improved task performance when the human and autonomous synthetic agent communicate using a combination of audio-voice and graphics over when the human and autonomous agent communicate using graphics alone.

Hypothesis 3:   Using a combination of anthropomorphic audio-voice communication and graphical communication between the human and autonomous synthetic agent will improve the overall teaming and facilitate the creation of a HAT in comparison to when the human and autonomous agent communicate using graphics alone.

1.3.3 Research Project Plan

The research approach adopted is to carry out a mixed methods analysis. Quantitative methods will be used to observe and evaluate the effect that the presence of audio-voice communication has on human situation awareness and performance. Qualitative questionnaires and interviews will be used to collect data on the participants experiences and perceptions of the autonomous synthetic agent as a team-member.

The research aims to contribute towards scientific knowledge and flight safety by demonstrating and providing empirical evidence of the possible benefits of using an auditory speech

communication interface between the human and autonomous synthetic agent in an aviation HAT setting.

1.3.3.1 Thesis Structure

The thesis is developed in three primary structures or stages closely linked to the chapters of the thesis: a Theoretical Proposal conducted as a Literature Review and a Methodology; a series of Empirical Studies; and an Evaluation and Analysis of the cumulative results, observations and deductions arising from the Empirical Studies. The relationship between the Research Stage, Research Objectives and the Thesis Chapters is provided below in Table 1.1.

Table 1.1   Relationship between Research Stages, Objectives and Thesis Structure.

Research Stage	Research Objective	Thesis Structure
Theoretical Proposal	RO1. Conduct a review of research literature on human situation awareness, human autonomy teaming and the implementation of audio-voice interfaces between human and synthetic agent. This review will increase researcher knowledge and identify key experimental design factors and constraints to be accounted for in subsequent studies.	Chapter 2 Literature Review
	RO2. Prepare a design framework for generating functional task-orientated synthetic audio-voice communication messages based upon aviation best practice. RO3. Select methodologies for evaluating and measuring human operator SA, performance and teaming.	Chapter 3 Methodology
Empirical Studies	RO4. Design and build simulated aviation applications complete with a synthetic agent capable or reacting to real-time events and speech synthesis and voice recognition.	Chapter 3 Methodology
	RO5 Evaluate the impact of audio-voice communication from a synthetic agent on human SA, decision-making behaviour, task performance and perception of teaming when conducting a dynamic task.	Chapter 4 Pilot Study Chapter 5 "Shisa Kanko" Study Chapter 6 UTM Study
	RO6 Evaluate the impact of audio-voice communication from a human to a synthetic agent on the SA, decision-making behaviour, task performance and perception of teaming when conducting a dynamic task.	Chapter 5 "Shisa Kanko" Study
Evaluation and Analysis	RO7 Provide scientific evidence-based knowledge on the possible benefits of using an auditory speech communication interface between the human and autonomous agent in an aviation HAT setting.	Chapter 6 UTM Study Chapter 7 Discussion Chapter 8 Conclusion

1.3.3.2 Theoretical Proposal

The primary aim of the theoretical proposal is twofold: to conduct a review of the existing literature on SA and human interaction with autonomy; and to then use the key elements of information gathered to design a series of three experimental studies that will answer the research questions. The first stage of the research is covered in the Chapter 2 Literature Review and Chapter 3 Methodology as follows:

Chapter 2 – Literature Review. The literature review will focus on examining scientific knowledge from theoretical and experimental research previously conducted on how human SA is linked to and effects human performance and decision-making, and how that SA is affected both positively and negatively by operating within in a team of agents, those agents either human or synthetic. The literature review will also examine how the division of labour and the working relationship between a human and highly automated systems can affect human SA and performance; and what teaming and communication mitigations can be implemented to improve that human SA and performance. The intent of the review is to provide broad spectrum background knowledge of the key topics of the research, but also to inform the decision-making on the scope of experimentation, the factors that will need to be incorporated or addressed by the experimentation (eg what content should be included in communication messages to improve SA) and the opportunities to provide new knowledge and contributions to science.

Chapter 3 – Methodology. The methodology will then apply the knowledge found in the literature review to inform the design of the empirical studies, discussing the key factors that either limit the research or need to be limited to ensure that the experimental approach proposed provides reliable and valid results. The key factors considered include communication message use (type), message structure, and teaming structures and communication interactions. A detailed description of the two experimental apparatuses (two simulators) will

be given. Details will be provided of the five independent variables identified (Presence of Voice, Team Structure, Detail of Message, Operator Speech and Automation Degradation) and the experimental conditions for each independent variable. The four Dependent variables (Performance, SA, Perception of Teaming, and Workload) will also be identified and discussed, and the methods available to measure those dependent variables discussed and selected from for implementation in the studies. Finally, some of the overarching administrative elements of the studies will be discussed, such as the audience selection and ethics adherence.

1.3.3.3 Empirical Studies

The second stage of the research is covered in Chapters 4, 5 and 6 which will provide details of the three major studies conducted to evaluate the effect of audio-voice communication on a human operator of a highly automated system, in specific configurations of work and task distribution and specific implementations of an audio-voice communication capability. Three studies will be required to evaluate all the conditions necessary to determine effect of the five independent variables.

Prior to conducting the studies two simulators of abstracted aviation traffic management tasks will be produced, complete with a synthetic agent. Designed, coded and built specifically for this thesis by the candidate, these simulators will dynamically produce and transmit audio-voice messages that meet the key requirements for transferring SA and enhancing teaming identified in the Literature Review (Chapter 2). To save replication, the function and use of these simulators will be described in the Methodology (Chapter 3) rather than each study chapter.

The first of the three studies will be a Pilot Study, this will provide a general identification of the effect of implementing a synthetic agent voice capability in four different teaming structures that are designed to emulate common teaming interactions. The pilot study will only feature speech from the synthetic agent to the human. The second study will expand the scope of communication to include speech from the human to the synthetic agent. The third study will

be an applied study simulating a Counter UAV system, with the synthetic agent providing decision initiation and decision support.

1.3.3.4 Evaluation and Analysis

The final stage of the research and thesis will be the analysis of the results of the three studies collectively to provide overarching themes and conclusions on the efficacy of implementing a synthetic audio-voice communication capability between a synthetic agent and a human operator. The discussion will summarise key conclusions, observations and deductions emerging from the three studies in concert. This final stage of the research will aim to provide the final contribution (and meet the final objective) of summary design advice for determining whether to and how to implement a conversational interface and further research that could be conducted to expand upon that knowledge.

Chapter 2 – Literature Review

2.1 Aim

In order to prepare to conduct research into the effect of implementing audio-voice communication in a Human Autonomy Team (HAT), a Literature Review of historical and current research into the key subjects of the underpinning research question and outcomes identified in the Introduction (Chapter 1) was conducted. This chapter provides a report on the findings and conclusions made from that Literature Review.

The aim of the review was to provide context to the research area and aid the researcher objectively understand the original problem that led to the research questions and the likely causes of that problem. That knowledge was then used to identify and propose likely solutions and research necessary to test those solutions. The review examined theoretical and applied research conducted on human interaction with highly automated systems, particularly research that focused on understanding human-automation interaction and performance issues. The intended outcome of the review was to determine how this thesis project could conduct research that would provide scientific knowledge that would contribute towards addressing the root cause problem of human loss of Situation Awareness (SA) and generally contribute towards the improvement of aviation flight safety.

This chapter is broken down into three primary sections: a discussion on human SA; a discussion on general human interaction with automation; and a final shorter discussion on the specific case of human automation interaction through speech communication.

In the first section, the report discusses previous theoretical and practical research into the subject of human SA. The discussion on SA is primarily an examination of the founding theories of SA that remain in common use and how they have developed. Topics of discussion include what SA is; and how it is developed, maintained, and adapted to meet the ongoing demands of the operations that the human is undertaking. The scope of the review is broad, examining how

SA is developed in the individual, in teams of humans, and in teams of humans and non-humans or synthetic agents (automation or autonomy).

The second part of the literature review examines historical and recent research into human interaction with highly automated or autonomous systems, focusing on theoretical analysis and proposals on the division of labour and work relationship between the human and automation of the system. This section of the review starts by identifying fundamental issues identified with human and system performance in complex highly automated systems. It then examines theoretical proposals for interaction protocols to address those issues, starting with early proposals and developments on Levels of Automation (LOA) and then moving onto the more recent concept of Human Autonomy Teaming (HAT).

Finally, the third part of the review will examine existing research into the implementation of speech or conversational interfaces in autonomy.

An examination of literature on the measurement of SA, performance, teaming or workload will not be covered in this literature review but will instead be covered in brief in the Methodology (see Chapter 3.3 and Chapter 3.4.4) portion of this thesis. Where possible, attention will be paid to SA research conducted in the context of aviation.

2.2 Part 1 – Situation Awareness

2.2.1 Individual Situation Awareness

Situation Awareness (SA) is loosely and generically identified as "knowing what is going on around you" or "having the big picture" (Jones 2015:98) and has been variously described as a mental construct (Salmon et al. 2008, Dekker 2015), an abstraction that exists in our minds (Billings 1995) and a dynamic mental model or schemata of the current situation or system (Endsley 1995a). Since the mid-1980s the topic of SA has attracted a large amount of interest and has generated both theoretical and experimental research and debate (Endsley 2000, Edgar et al. 2017), with researchers either advocating the concept (eg Endsley 1995a, Stanton et al. 2017) or opposing it (eg Billings 1995, Dekker 2015).

The concept has been identified as likely originating as a military aircrew expression or "jargon" (Taylor 1990, Salmon et al. 2008) particularly in the context of "losing situation awareness". This observation of the source of the term is important as it helps answer critiques made of the scientific credibility of the topic (eg Billings 1995, Dekker 2015) by identifying that the utility and value is not necessarily in cognitive psychology, but rather in engineering, ergonomics and human factors applications in industry – particularly the aviation industry where pilots identify that a loss of SA can significantly increase the chance of them making a critical performance error (Endsley 2004).

Since the provision of the early definitions and discussion on the nature of SA, the use of the term SA has increased dramatically (Patrick and Morgan 2010, Stanton et al. 2017) and SA pure or applied has become one of the most popular safety-related research concepts (Salmon and Stanton 2013). Whilst initial research and conferences may have been centred on the aviation industry (eg the International Conference on Experimental Analysis and Measurement of SA held in 1995 in Daytona Beach, Florida) research into the topic has become more widespread, with many researchers investigating the construct in a range of different domains such as sport

(Murray et al. 2017), chemical manufacturing (Naderpour, Lu and Zhang 2016) and autonomous vehicles (McAree, Aitken and Veres 2017).

Early research focused on providing theories and models to define and explain how an individual achieved SA (Salmon et al. 2008), whereas more recent research has tended to examine the creation of SA in complex systems with teams of human and artificial agents (eg Grimm et al. 2018, Guznov et al. 2020). This change from primarily theoretical to applied research would suggest that a theoretical model of SA has been universally accepted; however, this does not appear to be the case with periodic outbreaks of robust debate on quite basic aspects of the theory (eg Special Issue of Human Factors Journal Volume 37 March 1995; and Special Issue of Journal of Cognitive Engineering and Decision-making, Volume 9, March 2015).

In 2017, Stanton et al. conducted a review of the most cited articles that provided either a definition of SA or a detailed discussion of the nature of SA. The top four cited articles, all of which in 2017 had over 100 citations, were:

- Endsley (1995a) "Towards a Theory of Situation Awareness in Dynamic Systems" (2253 citations)

- Sarter and Woods (1991) "Situation Awareness: A Critical but Ill-Defined Phenomenon" (273 citations)

- Taylor (1990) "Situational Awareness Rating Technique (SART): The Development Of A Tool For Aircrew Systems Design" (272 citations)

- Smith and Hancock (1995) "Situation Awareness Is Adaptive, Externally Directed Consciousness" (204 citations)

This literature review will focus on the three most popular theoretical models identified by Stanton et al. (2017); those of Endsley (1995a), Sarter and Woods (1991), and Smith and Hancock (1995). Taylor's (1990) article is primarily a description of how an SA evaluation methodology

was created and is thus more fitting for discussion in the methodology chapter of this thesis. The remaining articles of the Stanton et al. (2017) list will not be considered.

2.2.1.1 Endsley's Three Level Model of SA

The most well-known and, according to Stanton et al. (2017), in 2017 the most frequently cited theory of SA is the model proposed by Endsley (1995a). This model continues to be extensively referenced in modern research on SA in a wide spectrum of application as demonstrated in more recent articles on Automation (Demir, McNeese and Cooke 2016, Chen et al. 2018), Training (Lehtonen et al. 2017) and Medicine (González-Martínez, Bangerter and Le Van 2017).

Endsley (1995a) describes SA as the resultant dynamic mental model of the situation, or situation model. The creation of that situation model is driven by the goals and experiences of the individual and observations on the conditions that the individual finds themselves in. Endsley (1995a:36) provides a definition of SA as:

> "the Perception of the 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"

This description is deconstructed to produce a three-level model of SA:

- Level 1 SA: Perception of elements in the environment: "The first step in achieving SA is to perceive the status, attributes, and dynamics of relevant elements in the environment" (Endsley, Bolté and Jones 2003:14)

- Level 2 SA: Comprehension of the current situation: "The second step in achieving good SA is understanding what the data and cues perceived mean in relation to relevant goals and objectives. Comprehension (Level 2 SA) is based on a synthesis of disjointed Level 1 elements, and a comparison of that information to one's goals" (Endsley, Bolté and Jones 2003:17).

- Level 3 SA: Projection of future status: "Once the person knows what the elements are and what they mean in relation to the current goal, it is the ability to predict what those elements will do in the future (at least in the short term) that constitutes Level 3 SA" (Endsley, Bolté and Jones 2003:18)

The elements of the definition are the physical things or activities that an individual is gathering data on such as physical location, orientation, aircraft systems states and so on. The 3 levels are cumulative: comprehension is more than just the awareness of the elements perceived but is understanding of those elements in context of the goals of the individual; and projection, or anticipation, is of what the future states or actions of the elements being perceived are likely to be. The projection is based upon knowledge of the elements (perception) and comprehension of the situation (Endsley 1995a). The most frequently repeated of Endsley's models of SA is the first diagram of the 1995 article which illustrates the role of SA in the context of *decision-making*, and demonstrates the dynamic, temporal nature of SA (Figure 2.1).

Figure 2.1: Model of Situation Awareness in Dynamic Decision-making (reproduced from Endsley 1995a)

In the model Endsley separates SA from decision-making, explaining (Endsley 1995a: 36) that "SA is explicitly recognised as a construct separate from decision-making and performance" and "a person who has perfect SA may still make the wrong decision". Endsley identifies that there are key factors that affect the ability of an individual to create SA and divides them into either individual factors, such as goals, experience and information processing mechanisms, or task and system factors, such as interface design, complexity and automation (Figure 2.1).

Perhaps of greatest value, not only does Endsley provide a definition of SA and guidance, in the three levels, for deconstructing the complexity of expected SA, the model is also used to generate a SA measurement tool: the Situation Awareness Global Assessment Technique or SAGAT (Endsley 1988, Jones and Kaber 2005). This tool attempts to sample the theorised mental model of SA by asking individuals fact-based questions about their environment and activities,

thus providing a measure of the accuracy of their observations, deductions and anticipations, and by extension, providing a measure of their SA.

2.2.1.2 Criticisms of The Endsley Model

Appropriately, since the original publication, Endsley has shifted some of the details of the theories to reflect continual research in cognitive psychology (Klein 2015). However, despite the popularity of the theory, Endsley notes that the model has received some criticism prompting the preparation of at least one clarification paper (Endsley 2000) and more recently, to the publication of a special edition on the Journal of Cognitive Engineering and Decision-making (Vol 9, Issue 1, March 2015) which appeared to be primarily occupied with debating the validity of the Endsley model. Two of the criticisms and debates that are most pertinent to this research project are discussed below.

2.2.1.2.1 Process vs Product?

The issue most commonly raised is that in the 1995 article, Endsley differentiates between Situation Awareness as a state of knowledge, and *Situation Assessment* as the process of achieving SA (Endsley 1995a). In another publication Endsley (1995b: 20) again states "situation awareness as defined here is a state of knowledge about a dynamic environment. This is different than the processes used to achieve that knowledge". However, as observed by Patrick and James (2004: 62), "it is difficult, if not impossible, for psychology to separate cognitive processes from their associated products, particularly in complex, dynamic performance situations".

This separation of cognitive product from cognitive process led to some confusion and robust debate. Salmon et al. (2008) draw attention to the apparent contradiction of Endsley's article in which Endsley asserts that SA is a state yet at the same time provides a definition of SA that is a description of a process. Sarter and Woods (1991) are quite clear that they regard Endsley's

model as a description of the process of situation assessment, stating Endsley in the 1988 paper "describes this process as involving three different levels" (Sarter and Woods 1991: 50).

Endsley did address the issue (Endsley 2015: 11) observing that "some people have used this statement to claim that the Endsley 1995 Model only addresses SA as a state and not as a process". Endsley (2015) explains that the original 1995 differentiation was done purely for terminology at the request of the editors to allow discussion and debate over the two facets of SA; the process and the product. Endsley, Bolté and Jones (2003: 20) explains that the two are practicably inseparable; "A person's SA will in turn have an effect on what information is searched out and attended to, with product affecting process in a circular fashion". More recently Endsley (2015: 11), has identified that the scope of SA includes the processes needed to generate the product, explaining "it is inaccurate to claim that this model does not address SA processes or that it does not show the two as being clearly intertwined". Thus, the deduction from the Endsley (2015) article is that SA consists of both product and process, but for consistency of conversation and discussion the term is generally applied directly to product and only obliquely to the process.

Unfortunately, this initial separation has a legacy, with some researchers still prone to consider SA as only a product, for example Naderpour, Lu and Zhang (2016: 147) observe "It introduces SA as a state of knowledge in the human mind".

2.2.1.2.2 Linear Information Processing?

A second point of debate is over whether the Endsley decision-making model is a linear Information-Processing model, and thus, whether it is necessary to go through each of the 3 Levels of the SA process to successfully gain SA. To illustrate the point that Endsley's 1995 definition is of a linear model, both Klein (2015) and Stanton, Salmon and Walker (2015) comment on the language Endsley used to describe how SA is achieved in her earlier articles ("the first step", "based on") and Klein (2015) notes that the arrows and layering of the decision-

making diagram show a sequential arrangement. Stanton et al (2010: 31) also observe that the model appears to be a "standard information-processing model found in many text of cognitive psychology [..], with the exception that the words 'comprehension' and 'projection' are written in the middle box in place of 'working memory' or 'global workspace'"

Not all observations on the use of a linear model are negative. Stanton et al. (2010) suggest that there are advantages to using a standard Information-Processing model to introduce a concept as complex as SA. Patrick and Morgan (2010) also found value in the logical sequence, proposing to present SA as a task with a Hierarchical Task Analysis that decomposes the Task "1.1 To Achieve and Maintain SA" into three Sub-Tasks that must be achieved in sequence: "1.1.1 To Perceive Elements of Current Situation"; "1.1.2 To Comprehend Current Situation" and "1.1.3 To Project Future Status of System".

Endsley (2015) offers the clarification that the 1995 definition is not of linear stages of SA, but is describing three ascending levels of SA and explains that SA can be constructed both "top-down" and "bottom-up". However, even this determination that it is levels not stages attracts opposition with Sorensen, Stanton and Banks (2010) noting that experts with SA are unable to (declaratively) divide their SA into the three levels in a meaningful way.

2.2.1.2.3 Impact of Endsley Model For Research Project

The key outcome from this occasionally robust exchange of argument and counterargument is that some of the more rigid (and potentially unnecessary) Endsley definitions over what is and is not SA have been relaxed, allowing the model to move away from strict adherence to the cognitive psychology model of information processing theory. Endsley's (2015) later clarifications that the model is simultaneously top-down and bottom-up, is not linear but is cyclical and that the levels are hierarchical and not sequential allow the model to be viewed in relation to more recent cognitive psychology theories such as the dual processing theory (eg Evans 2006) and the multi-component model (eg Baddeley 2010). Ultimately, the deduction

from Endsley's robust defence in 2015 is that the term encapsulates both product and process. Thus, logically for completeness a measurement of both could be used to assess the quality of an individual's SA. A discussion on the methodology for measuring SA is provided in the Methodology (Chapter 3.3).

Perhaps most importantly, this expanded (and logical) viewpoint that the process of developing SA is as critical as the product and that the process is a multi-layered feedback loop indicates that when preparing systems to support SA it is not sufficient to simply identify what SA information to present. The cognitive processes that an individual is required or likely to go through must also be considered, including the decision-making activities that result from the SA, as these processes will all in turn feedback and effect the SA built and the scope of the ongoing situation assessment. Furthermore, if attempting to design a system to deliberately push SA information, the timing and mode of the presentation of that information must be considered as that presentation could have a disruptive effect on the already ongoing SA processes that could theoretically result in diminished SA.

This deduction on the criticality of the timing and conduct of pushed SA information leads to the consideration that advantage could be gained from cognitive timesharing suggested by multiple-resource theory (Wickens 2008) and the multicomponent model (Baddeley 2010) in that the building of SA could be positively supported by presenting essential SA information simultaneously through multiple channels. This in turn provides some theoretical support for the primary hypothesis of this research project that the audio-voice channel could present an ideal medium for the automation (synthetic agent) to deliver selected and essential SA information without disrupting ongoing SA building processes using visual cues.

2.2.1.3 Sarter and Woods Definition of SA

Sarter and Woods (1991) do not attempt to provide a discrete diagrammatic model, but instead provide a debate over some of the key factors, pre-requisites and cognitive requirements of SA

in order to offer a definition of SA. Like Endsley, Sarter and Woods identify that SA is created

through the process of making situation assessments; however, they place a much greater

emphasis on the temporal requirements for SA. They identify that SA is not only dependent

upon current situation assessments, but on the continuous integration of current assessments

with previous assessments and pre-existing and context derived mental models. The

information for the mental model and the assessment comes from knowledge that is either

currently available or can be activated. This leads to their definition of SA as (Sarter and Woods

1991:55)

> "all knowledge that is accessible and can be integrated into a coherent picture, when
>
> required, to assess and cope with a situation"

Thus, Sarter and Woods suggest that SA can only be achieved if the results of previous situation

assessments are recalled and integrated. If a situation assessment is missed or inaccurate, SA

immediately degrades and there is a temporal cost to restabilising optimal SA, during which

period we could expect SA dependent cognition such as decision-making and performance to be

sub-optimal. Recently, in an experimental study Strayer et al. (2016) found evidence for this

temporal penalty in for re-establishing SA. They found that drivers who had been subject to a

distracting high cognitive workload task, once that task was removed, would take up to 27

seconds to recover their reaction-time capability back to their pre-distraction optimal levels.

2.2.1.4 Smith and Hancock Perceptual Cycle Model

In contrast to Endsley (1995a), Smith and Hancock (1995) do not consider SA to be the product

of cognition, but rather view SA as the cognition process that leads to a state of knowledge.

They do not regard SA as a construct, but rather as a competence (Smith and Hancock 1995:141)

> "SA is the competence that directs the agents sampling of factors in the environment
>
> that determine how the agent can come to know what it must do"

Smith and Hancock (1995) propose an ecological model of SA that theorises that SA emerges from the constant and cyclic interaction between the agent and its environment. They argue that for SA to exist, an individual must be directing their consciousness externally towards achieving a goal that lies in the task environment rather than just the individual's mind; "until an external goal and criteria for achieving it are specified, examination of greater or lesser degrees of SA or even loss of SA remains impossible" (Smith and Hancock 1995:139). They reiterate that even in a team situation, as the team are part of the environment then SA can only be achieved through goal directed inter-action "to assess SA it is not sufficient for you to think you understand the actions of another. Those actions must be assessed in terms of explicitly stated goals" (Smith and Hancock 1995:140)

Smith and Hancock relate their proposal to Neisser's (1976) Perception Cycle (Figure 2.2) to provide an explanation for the cyclic interaction and how that relates to SA (Smith and Hancock 1995:141):

> "the environment informs the agent, modifying its knowledge. Knowledge directs the agent's activity in the environment. That activity samples and perhaps anticipates or alters the environment, which in turn informs the agent. The informed directed sampling and/or anticipation capture the essence of behavior characteristic of SA".

Figure 2.2: Ecological Model of Situation Awareness (reproduced from Smith and Hancock 1995)

Whilst Smith and Hancock (1995) acknowledge the place of knowledge in their model, unlike Endsley (1995a) they actually argue against the use of mental models, suggesting that [in 1995] it is too ill-defined a concept. However, they base their model on Neisser's (1976) perception cycle which is completely dependent upon the use of schema; Neisser explaining that mental schemata "direct perception activity and are modified as it occurs" (Neisser 1976: 14).

Stanton et al. (2009) and more recently Edgar et al. (2017) address this disconnect by relating the Smith and Hancock model to the two types of schemata proposed by Neisser; genotype and phenotype schemata. Stanton et al. (2009) explain that Genotype Schema are generic long-term storage schema and are used, with current environmental and local context to dynamically create a Phenotype Schema of the current situation. Edgar et al. (2017), propose a cyclical, ecological model of SA (Figure 2.3) where experience and knowledge, stored in Genotype Schemata is combined with information of the current situation to develop a live and dynamic Phenotype Schemata that is then used as the basis for decision-making and general behaviour (Figure 2.3). The Phenotype Schemata is constantly modified by information received in a process that appears to be very similar to the process of creating situation models proposed by

Endsley (1995a). Edgar et al. (2017) even acknowledge that the Phenotype schema shares some of the features of mental models, thus drawing a connection between the Smith and Hancock (1995) ecological model and the Endsley (1995a) three level model.

Figure 2.3: Model of Actual Situation Awareness (reproduced from Edgar et al. 2017)

2.2.1.5 Common Themes In Models Of Individual Situation Awareness

Despite the robust debate evident in the 2015 special edition of the Journal of Cognitive Engineering and Decision-making, there are key common themes to be found between all the models.

2.2.1.5.1 Goal Driven Activity

Most of the models identify that SA is directly linked to the goals of the individual; the exception being Sarter and Woods (1991) who never actually mention goals but do discuss that SA is directed and in context. Endsley (2000: 15) is very clear "Goals are central to the development of SA" and includes goals as one of her key Individual Factors.

Smith and Hancock (1995: 140) are equally direct: "to stake a claim to SA, an agent must be seeking information and taking action in pursuit of an externally specified goal". They are very specific that the goals must be stated as an external goal, an interaction with the environment, effectively measurable by achievement and not an internal reflection. They acknowledge that individuals have their own innate personal goals that will influence behaviour and SA built, but do not specifically discuss these goals and their impact on SA. This is perhaps a missing factor for discussion, as it would seem plausible that an internal reflection or internal emotion based upon a natural goal (eg personal safety) could significantly affect the development of SA and could even lead to the generation of a new "externally specified" goal.

Irrespective of the debates over the models of SA and whether it is a product or process (or both) the review of the literature does lead to the conclusion that an analysis of goals, both internal and explicit, is essential to be able to build of a model for how SA is developed over time. The reasoning and problem-solving activities that contribute towards the development of SA are all goal focused; they rely upon reactive or predictive and above all subjective measurements against a chosen arbitrary standard, in this case a goal. Thus, the identification of goals for a task is essential if we are to be able to measure or direct the creation of SA associated with that task.

2.2.1.5.2 Mental Models and Schema

Most models identify that SA is achieved by reviewing the current situation assessment in the context of active mental models of devices or anticipated states that are, in part, derived from long-term mental schema. Sarter and Woods (1991: 49) are clear "mental models are one of the prerequisites for achieving situation awareness". Endsley (1995a: 41 and 48) provides diagrams showing the cyclic relationship between mental models, schema and SA.

Whilst Smith and Hancock avoid the use of mental models, the later work on their ecological model by Stanton et al. (2009) and Edgar et al. (2017) does introduce and discuss the roll of

cognitive schema in the creation and maintenance of SA, with Edgar et al. (2017) observing that their proposed phenotype schemata share some of the features of mental models and identify that Endsley (2000) uses both interchangeably. Thus, Edgar et al. (2017) demonstrate that it is possible to draw a connection between the ecological model with its genotype and phenotype schema and the Endsley three level model with its schema and mental models. Thus, whilst not all authors agree on the use of the term mental models, all do agree on the role of schema in generating and defining SA.

2.2.1.5.3 Temporal Nature of Situation Awareness

All the articles and models detail the temporal nature of SA. Whilst Smith and Hancock (1995) are overt and identify that SA is cyclical, Endsley (1995a) is not so clear providing pictorial models that show a cyclical relationship between SA and decision-making (Figure 2.1) and latterly arguing that the model is not linear (Endsley 2004, Endsley 2015).

However, on common ground both Endsley (1995a) and Smith and Hancock (1995) propose that SA is dynamic, being constantly updated. Sarter and Woods (1991) provide the strongest argument for the temporal nature of SA, arguing that not only does SA require constant updates through continuous assessments of the situation, it also takes time to build and requires those current and past assessments to be compared and integrated to be accurate, a dimension of SA they identify as temporal awareness. They explain (Sarter and Woods 1991: 49) "temporal awareness requires the diagnosis of problems that are caused or influenced by precursors in the past as well as the prognosis and prevention of potential future problems based on the analysis of currently available data". Klein (2000: 47) provides further clarity, explaining that "there exist situations in which the projection is backward into the past. For example, a medical diagnostician or a troubleshooter attempts to explain a sequence of events in order to understand the present more fully".

The common themes provide key requirements for consideration when determining how to evaluate the SA of the participant. The identification that SA is generated to address goals provides the key to establishing what to measure SA against; knowing the likely goals of the participants provides a clue as to what SA questions to ask, but equally, asking the participants about their goals allows the evaluation of the validity of any SA measure. The identification that SA is temporal suggests that to gain the best assessment of SA it is not appropriate to attempt to measure SA through a single post-exercise sample such as through SART (Taylor 1990) or SA-SWORD (Vidulich and Hughes 1991); rather multiple dynamically gathered samples are needed such as gathered by methods like SAGAT (Endsley 1988) or SPAM (Durso and Dattel 2004).

The observation that SA is not just product but also process indicates that it is not sufficient to simply sample or measure the resultant "knowledge" or phenotype schema through sampling, but it is also necessary to measure the frequency and informational scope of the situation assessments being undertaken to build the SA knowledge. A more detailed discussion on SA measurement techniques will be conducted in the Methodology (Chapter 3).

Finally, the association between SA and schema and mental models provides thoughts on participant audience selection; if the study is to provide an unbiased measure of "average" SA it is essential that participants do not have pre-existing task or skill schema that could make SA building "easier" or bias the scope and structure of the SA created significantly away from that created by other unskilled individuals.

2.2.1.6 Conclusions on Individual Situation Awareness

As Stanton, Chambers and Piggott (2001) observe, there are elements of value to be found in all of the models of SA. The citation evidence provided by Stanton et al. (2017) shows that the Endsely's three level model of SA is the most popular and most utilised model (and referenced) of SA. According to Salmon et al. (2008) the utility (and popularity) of Endsley's model is in part

due to its simplicity and the identification of the three hierarchical levels of SA, which allow the construct to be measured easily, and in part due to the provision of a comprehensive list of individual, task and system factors. Rousseau, Tremblay and Breton (2004) regarded the model as the most extensive SA model [in 2004] available, and Millot (2015) observed that it is one of the few definitions that has an associated measurement method (SAGAT). Thus, it would seem the Endsley model of SA is the logical singular model of SA to be used.

However, as identified earlier, many researchers have observed that there are apparent contradictions in the Endsley model. Endsley's (2015) clarifications that the model is both top-down and bottom-up, is not linear and that the levels are hierarchical not sequential appear to draw the model ever more closely to the ecological perpetual cycle model proposed by Smith and Hancock (1995). Both models propose SA is developed over time through interaction with the environment, is separate from decision-making but drives decision-making, and is dependent upon goals and schema. Nevertheless, one key difference remains: Endsley is clear SA has a product element that can be measured and provides a measurement tool, whereas the ecological model identifies SA as a competence and process that cannot be measured directly and can only be inferred through observations of interactions with the environment (which includes other team-members).

Partly because of its popularity and partly because of the availability of the SA measurement technique SAGAT, the Endsley model of SA will be given primacy for this research project when developing the methodology and hypotheses for the study, with one exception. Endsley's (1995a) argument that SA and decision-making cannot be directly linked seems potentially contrary, as the primary outcome of SA in the Endsley model is the directing of decision-making, as Endsley (1995b: 65) explains "It provides the primary basis for subsequent decision-making". In fact, Endsley's later (2000: 7) argument that "Decisions are formed by SA and SA is formed by decisions", suggests that SA and Decision-making are so intertwined and mutually dependent

that it would be quite difficult to effect one without effecting the other. Furthermore, the cognitive processes described by Endsley (Endsley, Bolté and Jones 2003) for Comprehension and Projection both include forms of decision-making; it is difficult to imagine making an evaluation of perceived data against expected goal cues and then making an estimate of likely future data without making a judgement decision in some form. Thus, whilst it is possible to accept that a singular good decision does not necessarily demonstrate good SA, it appears logical (using Endsley's own arguments on the efficacy of SA) to accept that a large cohort of individuals demonstrating a collective improvement in the quality of decision-making can be taken as evidence of a holistic "whole of cohort" improvement in SA.

This mutual dependency between decision-making and SA is of particular importance for this research project when considering how to measure SA, as whilst it may be difficult to measure a construct or abstraction (Billings 1995), it is not as difficult to observe and "measure" the decisions that are made and enacted. Thus, for this research project, where appropriate, decision-making will be used as a secondary evaluation of SA, the exact method to be identified and discussed in the introduction and methodology for each study.

In addition, as Endsley (2015) identifies that SA consists of both product and process (see Section 2.2.1.2.3 above), and both Endsley (2015) and the supporters of the ecological model (eg Smith and Hancock 1995) identify that interactions with the environment are key to building SA, it would appear that it is not sufficient to regard individual SA as only the knowledge product of "knowing what is going on around you". From the Literature Review it would appear that the process of building SA is as important, if not more important that the temporal product. Thus, when evaluating SA, the evaluation must include an estimate not only of the product, the state of knowledge, but also of the scope and efficiencies of the processes and activities conducted to build SA.

2.2.2 Situation Awareness in Groups

Early research on SA focused on defining and describing what SA meant to the individual. However, as most industrial activity and employment is conducted in teams, as soon as the primary individual models were complete, researchers soon shifted the focus of their attention towards the potentially greater challenge of describing and measuring the SA of teams of agents, especially those where the agents could be synthetic and autonomous. In fact, a good portion of the more robust discussion on SA (eg Endsley 2015, and Stanton, Salmon and Walker 2015) covers the differences between Endsley's Shared Situation Awareness and the Distributed Situation Awareness proposed by Stanton, Salmon, Walker, Sorensen and colleagues (eg Stanton et al. 2006, Sorensen and Stanton 2011, Stanton, Salmon and Walker 2015, Sorensen and Stanton 2016) that discusses how SA is built and distributed in systems consisting of humans, synthetic agents and even artifacts.

2.2.2.1 Team and Shared Situation Awareness

In the initial 1995 Human Factors special edition on SA at least two authors discussed SA in teams. Endsley's (1995a) view was that team SA was a resultant of the combination of the Individual SA; every team member must have the same SA for the team to have SA. Endsley viewed that there would be an overlap in SA requirements, and that communication or sharing of displays and creation of a shared mental model would ensure that the team members established the appropriate level of SA required to conduct their allocated roles in the team. Endsley and Robertson (2000) later identified the overlap of individual SA requirements amongst team members working on common goals as *Shared Situation Awareness*. Endsley and Jones (2001: 3) define Shared SA as "the degree to which team members possess the same SA on shared SA requirements". Thus, they view the SA of a team to consist of the summation of individual SA with each individual having two components in the summation: a personal SA needed for their unique tasks and goals; and a Shared SA developed in common with other members of the team.

Salas et al. (1995) also concluded that team SA was composed of two parts; however, they viewed the parts as two different abstractions: individual SA and team processes. They considered team SA from a process perspective; observing that each member of the team builds individual SA and then through team processes, influences the SA of the other members of the team and in turn has their own SA influenced. Thus, each member of the team only ever has individual SA. Rousseau, Tremblay and Breton (2004) appear to support this individual focus and view team SA as individual SA plus a number of process. Shu and Furuta (2005) also follow this theme and propose that Team Situation Awareness consists of an extension of individual SA to include a mutual awareness of the other members of the team and the team structures and behaviours.

The significant difference between the two schools of thought is that, whereas Salas et al. (1995: 131) identified that "team SA is at least in part the shared understanding of a situation among team members at one point in time" Endsley and Robertson (2000) are more demanding and determine that it is not sufficient for individuals to have a similar understanding; the individuals must share the same mental model to have the same SA of a shared aspect (Endsley and Robertson 2000). This requirement has led later researchers to question the validity of the model, with Stanton et al. (2009) suggesting that it is questionable whether individuals can share the same mental model and SA given that individuals are likely to have different goals, roles, experiences, training, skills, that inevitably lead to variances in personal schema.

Shu and Furuta (2005) were also critical of the concept of shared requirements and observed that an emphasis on team members having the same knowledge (Shared Situation Awareness) could lead to the dangerous situation (on the flight deck) where every team member shares the same common but incorrect SA and likely remain locked within their incorrect SA until an external event occurs to alter it. Ironically, Endsley and Jones (2001: 3) identify the same risk "The most dangerous situation is when both team members share common but incorrect SA"

Bolstad and Endsley (2000) evaluated the Shared SA model with interesting results. They theorised that shared data displays would provide the data requirements for the generation of Shared SA (where every member has the same SA) which would improve team performance. However, results indicated that providing each team member with a full version of all their colleagues displays induced lower overall performance compared to when each team member was only provided with a summary or abstraction of their colleagues displays (Bolstad and Endsley 2000). This does tend to support the alternative view that individuals, providing an individual contribution to a team effort, do not need the same SA (Shared SA) but indeed need their own individual SA that includes an element of SA of and about the team.

2.2.2.2 Ecological Approach to Group or System Situation Awareness

Not all researchers are content with Shared SA, Stanton et al. (2017: 456) observing that it provides less than satisfactory answers to questions over how SA is generated in "highly interdependent but remote complex systems in which awareness resides in perhaps hundreds or even thousands of disparate human and machine entities, some in close proximity and others distributed across the globe".

To illustrate the point, Garbis and Artman (2004) gave the example of the development of Team Situation Awareness between the various members of an underground rail system that they argued was dependent upon the team members negotiating and communicating with each other through or with artefacts that present essential SA information to the team. They argued that artefacts used for storing cognitive data need to be included in the unit of analysis of SA in order to be able to understand team or system SA. Hollnagel and Woods (2005) also argue this point, building upon Hutchins (1995) work on Cognitive Systems and suggesting that it is necessary to view complex systems containing both humans and automation as Joint Cognitive Systems. More recently, Chiappe, Strybel and Vu (2015) attempted to unify the Endsley view of SA with this ecological concept of Joint Distributed Cognition, observing the human habit of

offloading data into the environment as memory stores (eg paper, instruments and displays) and then no longer holding that raw data as SA, but rather holding knowledge of the location of the data as SA. Thus, they theorise that when using a system, individuals have a tendency to build SA of the system and its interface over SA of the environment represented.

In a series of articles (eg Stanton et al. 2006, 2009, 2010, Stanton, Salmon and Walker 2015, Salmon et al. 2010, Stanton 2016) Stanton, Salmon, Walker and others applied the theories of distributed cognition to provide an alternative model of SA in a group or system, identified as *Distributed Situation Awareness* (DSA). The model proposes that SA is distributed across a cognitive system and can be held by human and non-human agents (Stanton et al. 2006). The model views DSA as being held at a system level and more than a simple overlap or summation of individual SA. The model introduces two new concepts, *Compatible Situation Awareness* and *Transactive Situation Awareness* as explained below:

- Compatible Situation Awareness: is the view that system members possess unique SA of the same situation (driven by their individual role and goals) that is compatible with other members of the system and allows them to work with the other members of the system (Stanton et al. 2009). There is no requirement for an overlap of SA as required in Shared Situation Awareness thus allowing members of the system to have very different, non-overlapping roles.

- Transactive Situation Awareness: describes how through intra-team transactions of data, which could be active verbal communication, passive digital communication, or even transfer of processed data into artefacts for storage and later recall, team members effect each other's SA, leading to new observations, deductions and anticipations. The data recipients use the Transactive SA data for their own means, making individual interpretations (Salmon et al. 2010) which they then turn into further transactions.

Thus, in a logical extension of the Smith and Hancock (1995) ecological model whereby interaction with the environment builds individual SA, Stanton et al. (2006) observe that team or systems SA is built through interactions with the team or system.

Perhaps controversially, Stanton and colleagues view the non-human agents to include the general artefacts of a system that can be used to store and process data used for by agents in the system to develop SA; hence the observation "technological artefacts (as well as human operators) have some level of SA (at least in the sense that they are holders of contextually relevant information)" (Stanton et al. 2006: 1290). Endsley has objected to this extension and remains very clear; "Inanimate objects do not have 'awareness' of the situation or of anything else" (Endsley 2015: 26). Endsley considers SA to exist only in the cognition of the human mind (Endsley and Jones 2001) and that technological artefacts can only be repositories for SA data for human decision makers.

The DSA model has been successfully tested and applied in a range of studies from complex military command and control systems (Walker et al. 2009) and to UK Energy Distribution (Salmon et al. 2008) and Marine Pilotage (Sharma and Nazir 2017). Sorensen and Stanton (2016) demonstrated that more effective teams have a higher frequency of communications and a higher number of relevant SA transactions, and go on to recommend that the interactions between human and non-human agents should be the focus for measurement for SA in a system. In an experimental comparison of Shared Situation Awareness and Distributed Situation Awareness (Kitchin and Baber 2016) concluded that when teams are comprised of experts a DSA approach promotes high performance and that if a Shared Situation Awareness approach is employed it will hinder the performance of those same teams.

2.2.2.3 Conclusions on Situation Awareness in Groups

As with individual SA, there are elements of value to be found in all models and discussions of group SA be it Shared Situation Awareness, Team Situation Awareness, or Distributed Situation

Awareness. The value of each model will depend upon the structure, composition, structure and aims of the group.

The Endsley (1995a) model of Shared Situation Awareness appears to attract the most debate and does seem to be the most questionable as it carries the requirement for team members to share the same mental model for shared activities when in practice few teams have individuals carrying out the same activities, and mental models, being dependent upon individual cognitive schema, are unlikely to be duplicated between individuals. Furthermore, in teams deliberately structured with redundancy for safety, as in for example an aircraft Flight Deck with two or more Pilots, having a Shared Situation Awareness can be dangerous if the team share a common but erroneous picture of the situation (Endsley and Jones (2001). Thus, ironically it appears that the greatest value of the concept of Shared Situation Awareness is to provide a warning for how situation awareness in a group can be compromised and biased and turned into potentially dangerous "group think".

Conversely, the more varied concepts of Distributed Situation Awareness (DSA) acknowledge that the synthetic elements of a system can have and react to SA (abet in a pre-programmed manner) and requires that the team members constantly and proactively communicate selected elements of SA data designed to assist other team-members develop a unique SA to achieve their unique goals. Thus, the DSA model offers greater utility for this research project, in particular the concept and requirement for Transactive SA, which appears to directly support the primary hypothesis in that it suggests that communicated SA data is as valuable, if not more valuable, that observed SA data. The DSA concept that systems, in this research the synthetic agent, can also generate, store and then at the right time, transfer SA also provides guidance on how to design the autonomous synthetic agent that will be used for this research.

Finally, the wider discussions from other researchers on the value of SA of the team over shared SA assist determine that the SA transactions need to be focused on providing information to

assist the recipient build their own unique SA rather than assist the team reach SA synchronicity. Furthermore, the information should not only address building SA of the environment but should also address building an SA of the team itself by providing information on the current and planned activities of the other team-members and the reason for conducting those activities.

Thus, combining the knowledge learnt from the individual and team SA research the primary (hypothetical) determination is that for participant SA and performance to be improved the participant and the automation should use the audio-voice communication channel to explicitly pass information that the participant will use to create SA about the achievement of the task (eg progress, risks and opportunities) and SA on the automation (eg current workload, deductions on observations and rational for decisions). This latter requirement to pass reasoning for automation activities should be implemented in the methodology and its likely effect on participant decision-making and trust should be tested in the latter study as an independent variable (Reasoning Transparency).

2.3   Part 2 – Human Interaction with Autonomous Systems

2.3.1   Impact of Automation & Autonomy on Situation Awareness

Taylor (1990), Sarter and Woods (1991), and Endsley (1995a) all initiated their research on SA because of concerns over safety and performance issues resulting in part from unexpected negative effects on performance following the introduction of highly complex and capable automation in aviation flight decks and aircraft systems. The effects on human performance were unusual, varying from an inclination to turn the automation off, to over-confidence and over-reliance in the automation, even when it was performing in error (Parasuraman and Riley 1997).

As early as 1983 Bainbridge had identified that the introduction of automation technologies frequently carried with it unexpected and often ironically negative consequence on human performance. Bainbridge observed that although automation is often introduced to replace or reduce human manual control, it is not possible to totally automate all functions nor find automation solutions for all eventualities (of breakdown). Often, "designers tend to automate everything that leads to an economic benefit and leave the operator to manage the resulting system" (Parasuraman and Riley 1997: 232) through abnormal (predicted or unpredicted) events. However, as Bainbridge observes, it has been known since 1950 that it is "humanly impossible to carry out the basic function of monitoring for unlikely abnormalities" (Bainbridge 1983: 775).

Endsley (1995a) included Automation as a factor effecting SA, and observed that lack of SA on automation can lead to humans either deliberately or accidentally becoming disconnected from the automation and even system processes; in effect becoming "out-of-the-loop". The result of becoming out of the loop is that the operator loses access to key SA data, loses SA and starts to suffer a degradation in performance. The loss of SA is often because the highly automated and often semi-autonomous systems lack transparency and predictability (Shively et al. 2017);

operators do not know what information the automation is using to make a decision and how it is processing that information. The consequences are operator errors, over- and under-trust and reduced usage (Shively et al. 2017). The case of reduced usage refers to situations where automation is implemented to replace a human, but to reduce risk, the human is still engaged as supervisor with the responsibility of "taking-over" in extreme situations where the automation degrades or fails. As Bainbridge (1983: 776) identified some 50 years ago:

> "the job is one of the worst types, it is very boring but very responsible, yet there is no opportunity to acquire or maintain the qualities required to handle the responsibility".

This issue of both becoming out of the loop and suffering from reduced usage has not been solved by improvements in automation, in fact those improvements appear to often make the problem worse and often leads to a situation Endsley (2017: 8) labels the *Automation Conundrum:*

> "The more automation is added to a system, and the more reliable and robust that automation is, the less likely that human operators overseeing the automation will be aware of critical information and able to take over manual control when needed."

2.3.2 Levels Of Automation (LOA)

Various researchers (eg Kaber and Endsley 1997, Parasuraman 2000, Onnasch et al. 2014) proposed that a potential solution to address the human supervisor "out-of-the-loop" loss of SA problem was to avoid attempting to automate as much of a system as possible, but instead to take a more human centric approach to determining the scope of automaton based upon consideration of the capabilities and capacities of both the human and computer. The identified solution is to acknowledge that automation does not have to be all or nothing, but can be deliberately set to one of a number of pre-determined Levels of Automation (LOA) varying between no automation to full automation (Parasuraman, Sheridan and Wickens 2000). The expectation is that systems would be designed to meet a specific description of LOA that has

been selected for its known effect on human and automation, rather than simply automate as much as possible.

LOA are set or determined by identifying functions rather than tasks that could be conducted by either human and computer and using the combination of assignment to human or computer to generate the different levels (eg see Table 2.1 reproduced from Kaber and Endsley 1997 for an example).

Table 2.1: Functional Implementation of Automation to create LOA (Reproduced from Endsley and Kaber 1997)

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

The functions are often generic, cognitive and based around behaviours rather than tasks; for example, Sheridan and Verplank (1978: 8-16), in possibly the oldest proposal for LOA, identified six possible "Behavioral Elements to Characterize Degrees of Automation" (Requests; Gets; Selects; Approves; Starts; Tells) that could be assigned as functions to either a human or computer in a variety of combinations to give 10 discrete combinations or LOA. Sheridan and Verplank (1978), and Endsley and Kaber (1999) were not the only authors to propose their own taxonomy of LOA as an approach to human automation system design. In a review of LOA, Vagia, Transeth and Fjerdingen (2016) identified 12 different models or taxonomies of LOA all with different numbers of levels (from 4 to 10) and different descriptions of those of LOA, giving a total of 19 unique LOA.

Kaber and Endsley (1999) conducted one of the first experimental evaluations of the impact of LOA on SA and task performance of participants at each of 10 different LOA in normal operating conditions and full automation failure. They found that the variance of human performance was not sequential to the LOA, but rather seemed to be dependent upon whether decision-making was fully allocated to one of the two agents (human or automation) even if the other had decision over-ride capability. They observed that the best performance results under normal conditions were in Batch Processing (LOA3) and Action Support (LOA2) where the human generated options (made decisions) and the computer processed those options, followed by Supervisory Control (LOA9), Full Automation (LOA10) and Rigid System (LOA7) where the computer generated the options and processed results. These results can be seen in Figure 2.5 (reproduced from Endsley and Kaber 1997) taken from an experimental study where participants were engaged in completing a simulated Air Traffic Control task that required them to process by "collapsing" targets before the "Collided" with each other or "Expired" when they reached the centre of a radar screen. In the diagrams below the "Collapses" columns demonstrate performance success and "Expirations" and "Collisions" demonstrate performance failure. The worse performance was found in Decision-Support (LOA5) and the baseline condition of Manual Control (LOA1).

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

Figure 2.5: Performance by LOA in Normal Conditions (Reproduced from Endsley and Kaber 1997)

However, these LOA were also some of the worst for performance under abnormal conditions

(where the automation failed, and the human had to take full control).

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

Figure 2.6: Performance by LOA in Automation Failure Conditions (Reproduced from Endsley and Kaber 1997)

Unfortunately, the results for Kaber and Endsley's (1999) analysis of the impact of LOA on SA

were not so clear with no significant variance at SA Level 1 Perception and SA Level 3 Projection,

although they did observe significant variance and SA L2 Comprehension improvement in higher

LOA (L6, L8, L9 and L10) where decision-making and selection were automated.

These results on LOA seem to be relatively universal, with Onnasch et al. (2014: 8) reporting that

in a meta-analysis of 18 experimental studies into Degree of Automation (DOA is akin to LOA) "a

vast majority of these studies indicated a strong positive correlation of DOA and routine

performance" but also observed an inverse correlation between DOA and impaired performance

under abnormal conditions; as DOA increased, so the performance under automation failure

decreased. Furthermore, they also reported that the results on impact of LOA on SA were not

so clear, with five studies reporting loss of SA with DOA, four finding no correlation and 2 finding

a gain in SA with DOA.

The research into LOA has not meet with universal approval. Bradshaw et al. (2013) argue that

the LOA concept encourages the myth that the full scope of machine autonomy can be measured

on a single ordinal scale and does not allow that a system may have different functions at different levels simultaneously (and may in fact have adaptive or variable implementations of autonomy). This argument can be seen with both the Kaber and Endsley (1999) and Parasuraman, Sheridan and Wickens (2000) studies where the intent was to apply the LOA for each of the four function types universally across all tasks in a system without necessarily taking account of the task type and the innate task capabilities of the agents of the system. This same argument of over-generalisation could also be applied to understanding the impact on SA of automating a function. Functions cannot be automated effectively in isolation from an understanding of the task, the goals, and the context (Bradshaw et al. 2013), the safety risks and the impact on SA.

Perhaps the most damning evaluation of LOA comes in a United States Department of Defence (DOD) report on the future implementation of autonomy (Murphy et al. 2012: 23):

> "The milestones and roadmaps based on computer functions needed for some level of autonomy— rather than to achieve a capability through the best combination of human and machine abilities—foster brittle designs resulting in additional manpower, vulnerabilities and lack of adaptability for new missions"

> "The attempt to define autonomy has resulted in a waste of both time and money spent debating and reconciling different terms".

In the view of the DOD the issue stems from trying to apply as a design tool what was originally produced as a tool "to capture what was occurring in a system to make it autonomous" (Murphy et al. 2012: 24) and that using the LOA "as a developmental roadmap misses the need to match capabilities with the dynamic needs of the task or mission and directs programming attention away from critical, but implicit, functions needed for overall system resilience and human trust in the system". In other words, it appears that the DOD consider the value in LOA is in generating

a description of a system that can then be used as a point of reference for reporting on that system structure, and not in its use to attempt to design a system.

Despite these criticisms LOA remains in popular use, particularly in the automotive design industry after the formal creation of the six levels of driving automation (SAE International 2014). Other research into more generic LOA still continues, although much of it is theoretical (eg Onnasch et al. 2014) or does not actually propose a taxonomy of LOA but rather tends to use LOA as a vehicle to generate Within-Groups conditions when studying the variance of a facet of human performance with aspects of automation, such as the variance of human SA performance under automation degradation at LOA (Li et al. 2014) or impact of human cognitive abilities on performance when provided varying levels of information by automation (Jipp and Ackerman 2016).

It is this more recent habit of using LOA to provide conditions for testing human performance when interacting with automation (eg Chen et al. 2019) that is of utility to this research. The approach of artificially setting a range of specific team working structures allows a comparison of the effect of automation under different forms of interaction, the form of interaction effectively determined (or at least influenced) by the separation of work between the human and automation envisaged at a LOA. For example, it would be possible to run an experimental study using a simulator where setting the system at a range of specific LOA would allow the evaluation of a specific limited set of human-automation work interactions in isolation such as the human leading and synthetic supervising (LOA1), the human and synthetic sharing a job (LOA3), the synthetic acting as an advisor (LOA5/6), or the synthetic working and the human supervising (LOA9). This approach is used for the pilot study of the research programme.

2.3.3 Technology Advances Leading to Human Autonomy Teaming

As the research into LOA was being conducted, advancements in the capability and availability of information technology were simultaneously changing the way and extent by which

automation was being used. Automation was being used for functions of a system that at one time could only be performed by humans, especially complex cognitive functions, and even to implement functions that humans could not perform as accurately or reliably (Parasuraman 2000). Hoc (2001: 510) highlights the impact of this increase in computing intelligence, observing that "Machines are now able to make decisions and to implement them autonomously, being more than 'assistants'' to humans" and discusses the "new conception of the human-machine relation".

Miller and Parasuraman (2003, 2007) also accept Hoc's proposed change in perception, identifying that there must be "an approach to human-automation relationships". They suggest that systems should be constructed with neither human nor automation to be exclusively in charge of tasks, but instead tasks should be adaptively allocated through human delegation of tasks to the automation, delegation happening at the design stage or dynamically. They suggest that the human should interact with the automation in a way that mimics or levers off human teaming behaviours "Human operators need to be able to delegate tasks to automation, and receive feedback on their performance, in much the same way that delegation is performed in successful human-human teams" (Miller and Parasuraman 2003: 462). The consequence is a move away from the fixed and finite LOA to a more dynamic and almost infinite range of system structures, allowing either the human to undertake all actions, through combined work, to the automation undertaking all work (and potentially the human being called upon in exception when the automation encounters a problem it cannot deal with).

Other authors have also provided theoretical recommendations for changes to how humans should interact with complex automated or autonomous systems. Klein et al. (2004) develop the requirement of human and automation to be able to enter into an agreement or *Compact* to work towards shared goals and to support team coordination. They provide a list of 10 requirements for intelligent agents to be effective team members that centre on engaging in

negotiation over goals, task allocation and cooperation, and on the transparency of the automation. However, they also identify that agent autonomy must be controllable, directable and bounded, which appears to pose an oxymoron (constrained autonomy).

Lately, authors have emphasized the concept of the team-like cooperation between humans and machines (Johnson et al. 2011) and the more formal term of Human Autonomy Teaming emerged to describe the expectation of a relationship between the Humans and the Autonomous Agents in a system. Demir, McNeese and Cooke (2018a: 303) observe that "Traditionally, teaming consisted of only human-human teams, but advancements in robotics and advanced machine learning have led the way for Human–Autonomy Teaming",

2.3.4 Human Autonomy Teaming

The term Human Autonomy Teaming (HAT) has now come to represent a field of research that aims to understand how humans and automated systems can work together to optimise performance (Russell 2018). The general concept of Human Autonomy Teaming is that the human can no longer be regarded as the sole intelligent agent of the system who uses the automation as a machine or tool. In the HAT the automation is regarded as an entity with agency and autonomy (implementing Hoc's 2001: 510 quote "now, agents are not only human but also machines"), effectively becoming an autonomous synthetic agent.

The important and central requirement of HAT is that the human and the synthetic agent are to be considered as partners in a team. Goodman et al. (2017) require team member interdependence, the sharing of goals and the dynamic allocation of roles. Demir, Cooke and Amazeen (2018: 498) add to this definition by explaining that a HAT is "a team of humans working with intelligent autonomy by effectively coordinating and communicating with each other". In the HAT, both of the agents, the human and autonomy, should be able to cooperate, coordinate, and know when to interrupt an ongoing task and solve counterfactual "what if" questions (Battiste et al. 2018).

These teaming characteristics give some of the basic capabilities the autonomous synthetic agent must demonstrate to become a teammate. To be a synthetic teammate the autonomy must be able to communication and coordinate with team members (Demir et al. 2015), especially as communication in teams is an essential precursor and enabler for other team processes such as coordination and cooperation (Wynne and Lyons 2018). O'Neill et al. (2020) identified communication and transparency as critical to HAT performance; the high levels of transparency were considered likely beneficial to HAT as it clarifies the reasoning and decision-making used by the autonomous synthetic agent.

Whilst most definitions focus on specific characteristics of teaming, de Visser, Pak and Shaw (2018) simply recommend that the new human-autonomy interactions need to emulate the rich interactions of relationships between humans, and that human-human teaming models should be adopted for implementing HAT. Whilst not as overt in their conclusions, other researchers also appear to acknowledge the parallels between the two forms of teaming and identify that the requirements for HAT can be extrapolated from research into human-human teaming. For example, McNeese et al. (2017) discuss how effective human-human teaming depends upon good coordination and proactive communication and ask the question "Can HAT function adequately when the autonomous teammate lacks some of these subtle coordination behaviors exhibited by humans?" (McNeese et al. 2017: 2).

Battiste et al. (2018) proposed and tested modelling of HAT on existing human-human teaming paradigms, using the core tenets of the aviation human-human teamwork system of Crew Resource Management (CRM) as a model upon which to design and implement a HAT interface for their "Electronic Flight Bag". The resultant system gained positive feedback from the participant pilots of the study. From this and other studies Shively et al. (2017) and Battiste et al. (2018) concluded that CRM provided three key requirements for successfully implementing Human Autonomy Teaming:

- Operator Directed Execution. A single member of the team, the Operations Director needs to have responsibility for the overall performance of the team and thus has the final choice or direction on whether and what actions to be taken. The single lead member may delegate actions and responsibility to other members of a crew but retains overall authority and ability to reclaim control over actions. For example, in an aviation Flight Crew on any aircraft, the Captain has will delegate responsibility for the actions of the Cabin Crew to the lead Cabin Crew member or Purser for the majority of the flight but can recall command and take direct control in the event of an emergency. In the case of a Human Autonomy System, Battiste et al. (2018) determined that the Human Operator should occupy the position of Operations Director.

- Bi-Directional Communication. All members of the team are expected to explicitly communicate all actions and intentions, as required to achieve the tenet of transparency. All team members are expected to actively exchange information in both directions (to and from other team-members) to improve intra-team SA using a shared and unambiguous language or "phraseology".

- Transparency. There is a requirement for all members of the team to be able to understand the cognitive behaviours of the other members; to be aware of team-members, their observations on the situation, their planning, and their decisions and intentions. Battiste et al. identify that this is particularly important in the case of highly automated systems; the Human Operator must be able to develop mental models of the functioning of the automation.

Interestingly, many other HAT authors identify transparency as a core requirement for the synthetic agent to prevent the out-of-the-loop situation HAT is meant to prevent. For example, Furukawa, Nakatani and Inagaki (2004) demonstrated that human operators developed better SA of system performance and were more capable of taking-over from automation when the humans were provided details of the intention of the automation.

Continuing this line of research, in a series of studies from 2014 to the present, Chen, Lakhami, Selkowitz, Wright, Barnes, Stowers and others (eg Chen et al. 2014, Chen et al. 2016, Selkowitz, Lakhmani and Chen 2017, Schaefer et al. 2017, Wright, Chen and Barnes 2018, Chen et al. 2018) have investigated how improving transparency of agent planning and activity could improve HAT effectiveness, in particular the human's SA and trust in the system. Chen et al. (2014) observed that an operator's attitude and their trust in an agent guided their decision as to whether or not to use automation. They proposed that improved SA would result in creation of a balanced or "calibrated" level of trust (with elements of both positive and negative trust) that would inform the operator on whether or not to use the automation. They then conducted three studies in parallel, using various computer-based interfaces constructed using an Ecological Interface Design to provide information abstraction to meet all three of the Endsley (1995a) SA levels.

In 2018, at the end of the three projects, Chen et al. (2018) observed that transparency leads to the human agents being better able to comprehend the reasoning of the agent and better able to predict the agents future state, and that operators had greatest trust when the agent provided all three levels of SA information. However, the relationship between transparency, trust and complacency appeared to be more complex, with complacency seemingly affected by the human's familiarity with the task and environment. Finally, the projects suggest that transparency on the part of the synthetic agent benefits the human's decision-making and thus the overall human–agent team performance.

Demir, McNeese, and Cooke and others also conducted a series of empirical studies (eg Demir, McNeese, and Cooke 2016, Demir, McNeese, and Cooke 2017, McNeese et al. 2017, Grimm et al. 2018) to attempt to generally evaluate HAT and experimental identify key requirements necessary to generate the effect of HAT. They evaluated a range of effects, primarily centred on intra-team communications and how to make the autonomous agent be accepted as a teammate by the human members of the team. The experiments variously evaluated the team

performance, integration and SA of a HAT against the same (effects) in a Human-Human Team. The follow key observations on communication behaviour are taken from the series of studies:

- Human-Like Delivery. For best performance the autonomous agent should communication in a manner similar to a human; "even if human team members are properly communicating with a synthetic teammate, the errors and lack of human-like behavior on the part of the latter can still result in a negative team performance" (Demir, McNeese and Cooke 2016: 1);

- Pro-Active "Pushing" of Data. For best team SA the autonomous agent must anticipate other team behaviours and information needs and proactively "push" information to human team-members, as humans would not compensate for poor proactive communication by querying the automation more. This pushing of information was needed to improve Team Situation Awareness (Demir, McNeese and Cooke 2017) and to improve the intra-team coordination (McNeese et al. 2017);

- Communication Flexibility During Autonomy Failure. Teams displaying flexibility in communication, in the form of teammates anticipating communication requirements, performed better in abnormal situations when the system suffered from partial or full automation failure (Grimm et al. 2018). This indicates that all members of the HAT must be proactive in their communication and prepare for situations where any member of the team suffers a loss of capability, irrespective of whether human or autonomous agent.

2.3.4.1 Human Requirements for Human Autonomy Teaming

Increasingly, researchers are emphasising that for humans to accept a synthetic agent as a teammate, that agent must firstly be recognisable as both teammate and as human, as Groom and Nass (2007: 485) observe:

"When researchers ask the question, 'How do we make robots better teammates in human teams?' what they are really asking is, 'How do we make robots better human teammates?'"

The deduction is that simply operating a system that has an implementation of automation sufficiently complex to be considered semi-autonomous is not sufficient to generate HAT. As Battiste et al. (2018: 480) identify "People working with automation, even when that automation has a certain level of autonomy, does not equate to human autonomy teaming".

For HAT to occur, the human must be able to recognise the automation as a distinct entity, an autonomous synthetic agent, that fulfils unique roles on the team that would otherwise have to be filled by a human (O'Neill et al. 2020). Then, the human must perceive the synthetic agent as intent on assisting the human and sharing goals and mental models (Wynne and Lyons 2018). Groom and Nass (2007: 486) issuing the warning that "if a robot is not identified as a teammate, it may be treated as a tool", and O'Neill et al. (2020: 4) "If they are not recognised by humans as team members, there is no HAT". To be recognised as a teammate the autonomous synthetic agent needs to be engineered and adapted to integrate into human teams: the "autonomy must be able to work alongside human counterparts and carry out the fundamentals of teamwork and taskwork" (Cooke, Demir and Huang 2020: 134).

Wynne and Lyons (2018) build upon the basic agency requirements, proposing that the more anthropomorphic the synthetic agent is, the stronger will be the perception (in the human) that it is a teammate. This is because the more human-like it is the more that humans trust it, even to the point where humans will trust computer systems more than other humans (Przegalinska et al. 2019).

In fact, trust has long been identified as a core factor in determining the approach human's taking towards using the system that the synthetic agent represents. A definition of trust often quoted in literature discussing trust in automation is "the attitude [of the human] that an agent

will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See 2004: 51)

Poor reliability or difficulty of operation can lead to under-trust and misuse. Conversely, improved reliability and accuracy can lead to over-trust and abdication with humans not paying appropriate attention to the systems (Parasuraman and Riley, 1997). Rapid degradation in trust can then occur when the automation, previously considered to perform perfectly, makes errors (Madhavan and Wiegmann 2007) with the result that the human operators stop using the system. Then, once that trust is lost it can be very hard to re-establish it (Hoffman et al. 2013).

The solution to the trust issue may lie, ironically, in the core idea behind HAT that the relationship between the human and synthetic agents needs to change to resemble that found between members of a human team. It has been identified that people can be more forgiving of trust breeches from humans than from machines (Hoffman et al. 2013); their trust is more balanced or calibrated as they do not expect their human partners to be perfect (Madhavan and Wiegmann 2007). To make the human trust in automation more calibrated it is necessary to create a situation where the human's trust in the automation matches the actual capability of the automation (Khastgir et al. 2018).

There is in fact some evidence to support that making an autonomous synthetic agent more human-like will improve trust that will help create the HAT relationship. In an experimental study Waytz, Heafner and Epley (2014) identified that giving an autonomous driver anthropomorphic characteristics significantly increased the participant's trust in the autonomy. They suggested that the anthropomorphisation of the autonomy supported the perception that it had improved mental capacity and hypothesised that this attribution of mind to a machine could result in the creation of a machine to which users might entrust their lives. This finding of Waytz, Heafner and Epley is of particular relevance as, like the proposed approach in this project, they achieved their anthropomorphisation by providing the automation with a voice rather than

through a graphical or physical imitation of a human face or body. Thus, previous research has demonstrated that the addition of a speech capability to a synthetic may be able to provide appropriate anthropomorphic cues to make the synthetic appear sufficiently human to improve trust and convince the human operator that the automation is a synthetic agent.

Thus, the hypothesis that presents itself for this project is that making the automation behave more human-like by speaking will increase but also calibrate trust closer to that expected to be found between two humans. This calibration of trust will then in turn positively affect the relationship between the human and automation, turning it towards the desired HAT effect. A default metric for calibration could be the match between the human reliance of a system and the automation's actual level of reliability (Rusnock, Miller and Bindewald 2017). If the actual use and actual reliance are similar, then calibration of trust can be said to have been achieved.

2.3.5 Conclusion on Human Interaction with Autonomous Systems

The growth of interest in SA can be traced back to the widespread introduction of automation into the systems on the flight deck of aircraft in the 80's and 90's (Endsley 1995a, Taylor 1991), with Kaber and Endsley (1997) describing how humans working with complex automation often became "Out-Of-The-Loop", a phenomenon labelled by aircrew to describe not knowing what was happening to the aircraft. Much of the SA problem stems from the fact that modern automated systems are so complex and capable it is not possible for the human to see what data the system is processing and how it is doing so, nor keep up with that processing, and that the implementation of automation tends to remove the human from working and places them as supervisors of automation working. Researchers quickly identified this issue gave rise to an emerging group of negative human factors for operators working with systems with high levels of automation, such as lack of trust in automation, unexpected high levels of workload during safety critical periods, unexpected periods of reduced human SA, and skill fade in tasks now conducted by automation.

Early research into mitigations to counter this loss of SA focused on deliberately setting or constraining automation functionality at a specific Level of Automation (LOA) to ensure that the human had a specific function within the system; however, other researchers and major funding organisations (eg US DOD) rejected this solution, identifying that it was founded upon the concept that automation was a replacement for human labour rather than a new technology allowing the development of new system capabilities (eg Murphy et al. 2012). The alternative solution identified by researchers was to require a fundamental change in perception, relationship and thus interaction between the human and automation. The automation should no longer be viewed as just a complex machine or system but rather as an autonomous synthetic agent, and the human should team with this agent to achieve mutually identified goals.

This approach identified as Human Autonomy Teaming (HAT) has proven to be popular and apparently enabling. Research has not just been on the value of HAT to solve the out-of-the-loop problem (eg Chen et al. 2014 evaluating how HAT could improve transparency and trust), but also on the requirements to enable or create a HAT. Examining the research, the overall conclusion is that active and deliberate communication holds the key to implementing HAT; through it collaboration and coordination of team effort can be achieved, and essential SA information can be exchanged. This view is directly supported by applied researchers like Strybel et al. (2017) and theoretical researchers like Tokadlı and Dorneich (2019).

The conclusions drawn on the value of communication, like that found on team SA, support the hypothesis that audio-voice communication offers an opportunity to not only directly address the concerns over loss of SA when interacting with opaque automation (by ensuring the automation delivers SA knowledge that prevents or mitigates loss of SA) but also presents an opportunity to provide an anthropomorphic appeal that will enhance trust and increase the likelihood of the human regarding the automation as a fellow team member; ie the

communication can effectively amplifying the perception of HAT and encouraging human teaming behaviours.

A good proportion of the available research literature into implementing HAT shows that many researchers are indeed focusing on improving intra-team communication in order to enhance transparency, operator SA and trust (eg Chen et al. 2016, Demir, McNeese and Cooke 2016, Schaefer et al. 2017). Most of the experimental platforms being implemented utilise visual solutions to achieve that communication. Some are graphical, for example, Battiste et al. 2018 discuss the implementation of an "electronic flight bag" where participants interacted with a touch-screen tablet, and some have used a text messaging application (window) to provide conversational communication (eg Demir, McNeese and Cooke 2017).

2.4 Part 3 – Speech Interfaces with Automation

Interestingly, whilst researchers have noted that HAT is highly dependent upon a good intra-team communication capability (eg Demir et al. 2015) and at the same time have observed that the anthropomorphism of the automation (or autonomy) will assist the human to accept it as a synthetic agent and team-member (eg Wynne and Lyons 2018), the research into implementing a human-like audio-voice conversational communication capability within a HAT appears to be quite limited and difficult to find. Recently O'Neill et al. (2020) conducted a meta-study of experiential research on HAT, and found that of the 76 studies evaluated, 10 implemented communication using visual representation, 39 using text-based chat, and just 2 using audio-voice conversation. O'Neill et al. (2020) identify that more research examining communication in HATs is needed, to aid understanding of how various media, types of team processes addressed by the communication, and communication frequency and quality all relate to team outcomes.

The paucity of research into implementing a conversational interface for a HAT is interesting as speech is probably the most intuitive, and thus effortless and convenient, mode of

communication for humans (Veena et al. 2018). There does appear to be plenty of research and discussion on voice recognition and synthetic speech as a form of computer interaction with humans. However, much of it would fail to qualify as HAT communication research either because there is no apparent autonomy in the system, or because the autonomy is not implemented as a team-member. Generally, much of the audio-voice or speech research can be placed onto one of four categories:

- Research into HAT Speech using either Text or the Wizard of Oz technique.

- Research on conversational interfaces with "chatbots" used for commercial and domestic transactions (eg Bunz 2019, Przegalinska et al. 2019, Lago, Dias and Ferrira 2020).

- Research on the use of voice activated control systems (eg Simpson and Levine 2002, Poirier, Routhier and Campeau-Lecours 2019)

- Research on the technology used for voice recognition (eg Tamura, Iwano and Furui 2004, Lavrynenko, Konakhovych and Bakhtiiarov 2016, Veena et al. 2018)

The first category of research, in particular the use of text messaging for speech (eg Demir et al. 2015) or using the Wizard of Oz technique (eg Demir et al. 2019b) with its findings on effects such as the value of anthropomorphism (or disadvantages with lack of it) on performance and trust was partially covered earlier (Section 2.3.4 above). However, what was not discussed is the effect that a talking assistant (synthetic agent) can have on more general human cognition, in particular attention and workload. Research is increasingly showing that there can be a cognitive penalty associated with talking to automation equivalent to talking on a mobile phone (Large et al. 2016), with Strayer et al. (2016) observing that drivers can take up to 27 seconds to regain full attention on the road after providing auditory instructions to an infotainment system. However, other research has used this increase in cognitive workload to deliberately increase

attention and alertness with Mahajan et al. (2021) demonstrating that a voice assistant can be used to counter the effects of passive fatigue.

Whilst on the surface the third category of use of a voice activated control system would appear to cover research into human automation interaction, the systems being discussed are at a very low Level Of Automation (LOA), often the baseline of LOA1 or Manual Control. In many cases the audio-voice control interface appears to be a relatively simple audio substitute for a mechanical control device and not a complex communication system offering new interaction capabilities. For example, Nishimori, Saitoh and Konishi (2007), and Avutu, Bhatia and Reddy (2017) independently conducted research into using voice commands to control a wheelchair as an alternative to operating a hand control stick; no new interaction or controlling capability was introduced by the voice communication.

Occasionally the research does cover systems where the interaction is more than a simple baseline LOA of Manual control (LOA1) and the measurement is on effect of communication rather than success of communication. For example, Maciej and Vollrath (2009) discuss the effect on driver distraction of using an in-car navigation aid, where the human operator provides an address vocally and the system is silent and calculates route options to be selected vocally or manual by the operator. Simpson and Levine's (2002) wheelchair had an Assistive Wheelchair Navigation Systems that allowed the operator to provide gross directional commands, effectively abstracted outcome demands rather than specific manipulation demands. However, even in these systems, communication tends to be one-way from the human operator to the automation, with no exchange of task or processing information to prevent the operator losing system SA and thus prevent them becoming out-of-the-loop.

However, that is not to say that the research into audio-voice interfaces for chatbots and audio voice control systems is of no value to this research project. Quite the contrary. The research by Waytz, Heafner and Epley (2014) on autonomous car interfaces, and Przegalinska et al. (2019)

on Chatbots both demonstrate that providing the automation (as a car or chatbot) with a recognizably human voice improved the anthropomorphism of the automation and significantly improved the subjective trust of the human operator. This suggests that implementing an audio-voice communication capability could provide the benefit of making the automation more anthropomorphic and trustworthy. This would in turn encourage the human to perceive the automation as a distinct agent, an autonomous synthetic agent that the human could team with.

Simpson and Levine (2002) observed that providing the wheelchair with an automation "navigation assistant" that interprets generic instructions and provides semi-autonomous fine movement control allowed the number of controls to be reduced to a small dictionary that was consistent and intuitive, suggesting that advantage can be gained by limiting the input dictionary to a standard sub-set of natural language that is interpreted into complex activities. Later, Atrash et al. (2009) observed that allowing operators of a voice-controlled wheelchair to form their own sentences and commands lead to very poor speech recognition rates; the voice recognition software was simply unable to recognise the words and meaning of the sentences. The HAT research of Demir, McNeese and Cooke (2016) identified a similar issue, with the inability of their limited language synthetic teammate to understand all the unstandardised human conversational speech messages negatively impacting on the team performance. Thus, previous research has indicated that with currently limited capability speech systems it is better, for performance, to limit and standardise the conversation between the human and synthetic teammate.

Interestingly, the aviation teaming model of Crew Resource Management (CRM) draws similar conclusions about natural free formed speech used to communicate between human team-members; observing that natural language, with its colloquialisms, slang phrases, and non-standard phraseology can cause serious communication problems (CAA 2014a). The UK Civil Aviation Authorities CRM guidance suggests that "standard operational terms and phrases are

essential" and that "clarity of communication can be improved by methods such as standardising and restricting vocabulary, using short messages" (CAA 2014a: 139). This guidance is widely followed in aviation. Thus, in aviation human teams at least, not only is teaming possible with a deliberately limited and standardised conversational capability, it actual offers some benefit to the team, improving the clarity of communication and positively affecting performance.

Not all research results on implementing speech communication systems are positive. Maciej and Vollrath (2009) observed that voice control of in-vehicle information systems and navigation support aids improved overall driver performance compared to manual control of those same systems, but significantly reduced the driver's reaction time to all tasks, driving and other. Poirier, Routhier and Campeau-Lecours (2019), observed a similar time penalty when using voice control over manual control of a wheelchair; however, they found that participants subjectively preferred the voice control system, finding it to be more intuitive, requiring less concentration and effort to use than a manual joystick.

Comments made by researchers on the efficacy of previously available technology with which to conduct applied use of auditory speech may provide an indication as to why there is more research into how to build and implement speech systems than on how to apply them and their human factors affects. When discussing previous research and providing design advice for Chatbot systems McTear, Callejas and Griol (2016) observed that early Chatbot and Telecom voice-controlled systems were not easy to use, with speech often stilted and vocabulary restricted, which effected their capability and acceptance. Tamura, Iwano and Furui (2004) observed that whilst voice recognition in cars was effective in quite conditions with clear speech such as the simulation laboratory, in noisy conditions such as encountered when driving on the road, the accuracy of the recognition decrease significantly. This complaint was echoed by Rodemer and Wessels (2011) who determined that "today's handsfree systems for in-vehicle use have reached their performance limits". Strayer et al. (2016: 3) found similar issues with

speech activated in car entertainment systems, observing "many of the systems that are currently available tend to be complex and error prone, with inconsistent behaviour".

Simply put, the speech synthesis and speech recognition technology has not been sufficiently capable to support applied human factors research of truly synthetic systems; research has been difficult to carry out or has had to rely upon humans roll playing the synthetic agent and talking on its behalf (known as the Wizard of Oz method). Thus, the apparent lack of research into the use of audio-voice communication in operational systems may simply be a consequence of the current shortfall in the useable technology with which to conduct research. Incidentally, this reasoning could also explain why there is still considerable research being conducted into the core technology use to provide a voice recognition capability rather than the application of voice interaction; researchers are still trying to get the capability to work.

Irrespective of the reason for the lack of applied HAT research, as O'Neill et al. (2020) identifies, the majority of HAT experimental research on intra-team communication is limited to conversation utilising a visual interface (text), even when researching speech-based interaction. For examples, in a series of studies Demir, McNeese, and Cooke and others (eg Demir et al. 2015, Demir McNeese and Cooke 2016, Demir et al. 2017, McNeese et al. 2017, Demir McNeese and Cooke 2018a, Demir McNeese and Cooke 2018b, Demir et al. 2019a) evaluated the capability and efficacy of synthetic communication verbal behaviours on a range of aspects of teaming yet did so using a text-based messaging system rather than an audio speech synthesiser and voice recognition technology. Ironically, one of their key findings was that "even if human team members are properly communicating with a synthetic teammate, its errors and lack of human-like behavior can still result in a negative effect on team performance" (Demir, McNeese and Cooke 2016: 7). Of particular value, one of the key findings of the series of studies was that proactive communication where information is pushed from one team-member to another is associated with improved team performance; Demir, McNeese and Cooke (2017: 11) conclude

"In order to make HAT more effective in terms of teamwork, we need to develop mechanisms to enhance pushing information within HAT".

There could also be a negative consequence of the tendency to implement the communication interface as visual. Modern cognitive psychology (eg Baddeley 2010, Sternberg and Sternberg 2012, Eysenck and Keane 2015) provide theoretical models and practical evidence of how human cognition uses different areas (resources) of the brain for audio and visual processing, thus allowing for both audio and visual processing to be conducted simultaneously. The Wickens (2008) multiple resource model theorises that the ability to conduct multiple tasks simultaneously can be improved if the tasks each utilise these different cognitive resources (ie one uses audio and one visual), a phenomena Wickens identifies as cognitive timesharing.

Thus, using visual for both communication (eg text) and the primary task (eg aircraft control) could be imposing an additional load on the single (visual) resource of the human participants of HAT communication research, potentially even adversely affecting the performance outcome. However, conversely, theoretically implementing the HAT communication as audio and the primary task as visual could offer an opportunity to take advantage of dual cognitive processing systems and cognitive timesharing proposed by Wickens (2008). As explained by Wickens (2002: 162) "dual task performance is poorer when two visual tasks must be time shared than in a configuration in which the equivalent information for one of the tasks is presented auditorily".

2.4.1 Conclusion on Speech Interfaces with Automation

The reported shortfall of research into implementing a HAT audio-voice intra-team communication capability (O'Neill et al. 2020) provides the clearest indication of the research gap that this research project aims to address. Whilst there is not a large quality of research into implementing an audio-voice communication capability, there is sufficient research into similar fields such as direct voice control of mechanical systems and interaction with chatbots

and in vehicle information systems to draw some elementary conclusions on successfully implementing an audio-voice interface:

- To increase trust and acceptance the synthetic agent speech should be anthropomorphic, sounding human by offering cognitive detail and natural sounding sentences.

- To increase the likelihood of voice recognition success participant speech should be moderated to a standardised sub-set of natural speech.

- To reduce the temporal penalty on performance of using speech, messages should be short and succinct.

- To improve team performance the voice communication should be used to proactively push information, ideally information to build and support SA.

A further discussion on the practical implications of these findings and more detailed discussion of some of the factors identified is provided in the Methodology (Chapter 3) in order to provide a more defined scope and design for the audio-voice system to be implemented in this research project.

2.5   Literature Review Conclusions

The aim of the literature review was to carry out an analysis of past and present theoretical and experimental research on how to address the concern that when interacting with highly automated or autonomous systems humans can easily lose SA and become "out-of-the-loop" with potentially dangerous consequences on human decision-making and system safety as happened in AF 447 in 2009.

The Literature Review was broken down into three primary sections: a review of the most accepted and used models of individual and team SA; a review of research on implementing automation that attempts to identify solutions to addressing human SA issues when using autonomous systems; and a review of relevant research into implementing a voice communication capability in highly automated systems.

The opening study into SA started with a discussion of the existing theories and knowledge on SA in human individuals and then in human teams. The review highlighted that whilst at first the predominant theories of SA appeared quite different, as the theories have been developed over the past 25 years common themes within the theories have emerged:

- Individual SA covers both the knowledge product and processes of build that knowledge;

- Individual SA is goal orientated and is developed to address those goals;

- Individual SA is highly dynamic and temporal in that it takes time to develop and is only valid in the instant;

- Individual SA is highly dependent upon existing schema and dynamic mental models;

- SA in a Team is not singular (shared) but is achieved through deliberate proactive exchanges of raw knowledge and abstracted information from Individual SA;

- SA is not necessarily a human-only construct, but can be generated, stored and transferred from synthetic agents.

The second part of the literature review was an examination of research, both theoretical and experimental, into designing human-automation interfaces that address the issue of humans losing SA. Initial proposals attempted to define and set specific limits on the scope and interaction of automation as Level of Automation (LOA). More recently researcher have proposed that instead system design (separation of tasks between human and automation) and interface design should be based upon the concept of teaming, with the human and automation collaborating and cooperating to achieve common goals. Researchers with a background in aviation have suggested that the human approach to teaming known as Crew Resource Management (CRM) provides an ideal model for implementing HAT.

As with SA, a key underpinning requirement to implement the concept of HAT is the ability of the human and autonomy to be able to communicate to achieve the collaboration and coordination required of an effective team, and to follow the CRM model for implementing HAT. In fact, it seems that specifically audio-voice communication supports the achievement of the key goals of HAT:

- it supports the deliberate exchange of goal and task data between the teammates;

- it provides the conduit for the deliberate exchange of dynamic goal orientated SA information;

- It also provides a conduit for exchange of SA information about the team and team-member cognition that improves transparency of the team;

The final part of the literature review examined some of the existing research into implementing audio-voice interaction technologies, observing that:

- audio-voice communication can assist anthropomorphise the automation, giving it entity;

- human-human team best-practices of standardising audio-voice communications should be implemented to improve the likelihood of communication success;

- To reduce the temporal penalty on performance of using speech, messages should be short and succinct.

The overarching conclusion of the literature review is that there is a gap in the past and current research (see O'Neill et al. 2020) that indicates a need to research the implementation of an audio-voice communication capability as a solution, or at least part solution to address the issue of humans losing SA and becoming out-of-the-loop when working on highly automated systems.

The review of the literature also indicates the presence of at least five Independent Variables (IV) for consideration in this research programme:

- Presence of Voice: the prime IV of the presence or absence of the audio-voice messaging capability.

- Team Structure: the workflow and hierarchical structure of the team as that determines what and how communication is used.

- Operator Speech: whether the operator does or does not speak to the synthetic agent.

- Reasoning Transparency: the link between improved transparency on synthetic agent activities and reasoning for those activities and both trust and improved SA (especially SA about the other team-members) could be evaluated.

- Automation Degradation: as Human Autonomy Teaming is (in part) intended to assist prevent the loss of SA that leads to human operators becoming out-of-the-loop, especially in situations where the automation degrades, it would be appropriate, if time and scope of research permit, to test whether the implementation of the audio-voice capability will improve the ability of the human operator to not become out-of-the-loop during periods of automation degradation.

Chapter 3 – Methodology

3.1 Introduction

This research project will be conducted following a mixed methods approach. Wherever possible, quantitative analysis will be carried out to objectively test the effect that adding an audio-voice communication capability to the autonomous synthetic agent of a Human Autonomy Team (HAT) has on the SA and behaviour of the human teammate. To supplement this information, qualitative methods will also be used to sample the participants reported experiences and perceptions of the autonomous synthetic teammate. This chapter will discuss the methodologies selected, the variables to be tested and measured and provide insight into the general design of the studies and apparatus used in those studies.

The chapter will start by discussing two key factors that were identified as fundamentally affecting the design of the research studies that thus needed to be discussed before a research design and methodology could be fully prepared; the scope and limitation of audio-voice communication to be implemented and the selection of a primary Situation Awareness (SA) measurement technique. The first part of this section will detail the considerations made and literature consulted in order to determine what limitations to set on the scope of the voice communications and will then provide details of the structures and standards selected that would inform the design of the audio-voice message constructor. The second part of the section will discuss the most frequently used methods of measuring SA that were identified and will explain the selection of the method used.

The chapter will then move onto discuss the actual design of the research project, detailing the basic programme outline and the sequence of studies selected, although the more detailed information on each study will be provided in the relevant study chapter of this thesis. The two sets of apparatus used for the studies (two desktop simulators of aviation air traffic control systems) will be described in detail with explanations of the tasks set, how the participant was

expected to interface with them and how the autonomous synthetic agent of the Human Autonomy Team (HAT) would interact with the participant.

The chapter will then conclude with a discussion of the Independent Variables (IVs) identified, the conditions that would be used to control and implement deliberate variations of those variables, and the Dependent Variables (DV) that would be measured during the studies. Finally, the chapter will discuss the basic administration of the methodology, covering participant selection, ethics, participant training and post-trial data analysis.

3.2 Factors Affecting Audio-Voice Communication Implementation

Prior to designing the research mechanism and apparatus upon which to conduct the research, it was necessary to consider providing some definition or at least limitation to the scope of the research, in particular to the scope of the audio-voice communication capability to be implemented and evaluated. This scoping and derivation of limitations or standards was required to ensure that the design of the study provided participants with a sufficiently similar experience to allow comparison of measures between conditions and thus allow the conduct of the research as either a Within-Groups or Between-Groups experimental evaluation. The following section discusses the key factors identified that would affect the implementation of audio-voice communication.

3.2.1 Natural Language vs Plain Language

As the primary aim of the research was to evaluate the overall effect of providing a synthetic agent with a human-like voice communication capability it would seem logical that the voice communication messages would be designed to appear as human-like conversation using natural language. This deduction is supported by previous HAT researchers who have concluded that for synthetic speech communication to be effective it should use natural language, either because this is a direct requirement of the teaming model (Strybel et al. 2017) or because of an

observed negative correlation between team performance and human like communication behaviour (Demir, McNeese and Cooke 2017).

The expectation is that providing a natural language voice will assist the human participant suspend dis-belief and engage with the synthetic agent as they would a human teammate. 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 HAT recommended giving the automation "a better voice interface that uses natural language".

However, human voice communication is extremely complex and varied, and the aim, style and structure of the messages are communicated and can have their own effect on performance and SA, as demonstrated by Demir, McNeese and Cooke (2016). The experience of the aviation industry is that natural language is not a reliable and safe format to ensure a successful outcome from a 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 (CAA 2014a).

Furthermore, natural language can be extremely difficult to implement with any degree of realism using a real time synthetic agent without resorting to a "Wizard-of-Oz" methodology of using a human actor to imitate the synthetic agent (eg Demir et al. 2019b). For this research, which aims to investigate interaction with real synthetic agents, using a human actor is undesirable for two reasons. Firstly, it is undesirable as it means that the research is not directly applied as neither the voice nor the message is computer generated; it requires the decoupling of the real-time data processing of the actual synthetic agent from the speech production. Secondly, as the speech is generated by a human actor responding to cues and likely producing messages based upon a script, there is a direct risk that a human error (eg a simple stutter or

"er") from the actor could either provide a clue to participant they are interacting with a human or could generate a non-standard (and inconsistent) miscommunication that could skew results.

These concerns over implementing a natural language voice communication capability led to the conclusion that it would have been inappropriate and indeed, impracticable, to attempt to evaluate the full spectrum of natural speech possible. Arguable it would also be nugatory to attempt to test the full spectrum of natural speech in a dynamic task setting as realistically only a task-orientated sub-set of speech would ever likely be used even in a human only team (during the intense task work period modelled in this research). Thus, an alternative and practicably implementable compromise had to be found.

Interestingly, the aviation industry addresses the dichotomy of the requirement to use natural speech for teaming verses the requirement for error free communication through the Crew Resource Management (CRM) teaming model. CRM utilises over fifty years of both scientific research findings and employment best practice to maximise the efficacy of human-human teams in dynamic and safety critical situations. The United Kingdom Civil Aviation Authority (CAA) in their Human Factors Handbook CAP 737 (that provides instruction on the implementation and training of CRM) recommend the use of a more limited and standardised content of voice communication identified as "Plain Language" (CAA 2014a). The CAA also identify that the delivery of the communication should be standardised, emphasizing the importance of a steady speech rate, even in situations of high stress, and the use of clear, short and simple messages. The adoption of the general tenets of CRM are also recommended by the National Aviation and Space Agency (NASA) Ames Research centre as an "of-the-shelf" best practice solution to implementing the HAT teaming relationship, as explained by Shively et al. (2017) and Battiste et al. (2018).

For this research, it was determined to follow the CAA and NASA Ames recommendations and impose a reasonable and consistent standard format, structure and complexity to the

communication messages used by and heard by the synthetic agent for all studies. As well as eliminating negative effects from potentially poor, confusing messages open to misinterpretation and errors in comprehension identified by the CAA, implementing a standard form and structure of communication also provided the benefit of ensuring a consistent participant experience. Furthermore, from a practical perspective it eliminated the need for a Wizard of Oz actor or a complex AI to generate "natural speech" patterns and messages, making it possible to implement the communication using speech synthesis and recognition software that was directly triggered by the data processing elements of the simulator in its role as a synthetic agent. Thus, the research was able to genuinely evaluate audio-voice communication between a human and automation (the automation presented as an autonomous synthetic agent).

3.2.2 Types or Categories of Communication

To meet the primary aims of the research to investigate the effect of using audio-voice to improve SA, and to meet the generic requirements of transactive SA, the use of audio-voice communication was centred on the delivery of SA information. However, rather than make apparently arbitrary determinations on what types of message should be used and what SA information should be transferred, it was determined to base the studies on examples of operational aviation technology and to thus lever off aviation best practice on the scope and content of communication used to build SA.

Following the practices of the aviation industry, only a limited range or set of messages were implemented, the scope of the range of messages derived from the UK CAA instructions on voice communications between aircrew (pilots) and flight navigation services (air traffic control) in two key manuals: the CAA Radiotelephony Manual Civil Aviation Publication (CAP) 413 (CAA 2016) and the CAA UK Flight Information Services CAP 774 (CAA 2014b).

- CAP 413. In CAP 413 Chapter 2 the CAA identifies a range of categories of communication. For mandatory Flight Control three elements or types of communication are used: Clearance (directions that must be followed), Instructions (directions that should be followed) and Information (used to assist the safe conduct of the flight). The definitions of Clearance and Instructions also identifies a fourth element; the requirement to Acknowledge a communication. The CAP 413 also identifies "advice" as a sub-category of Information "An Aerodrome Flight Information Service Officer provides advice and information" (CAP 413:4-30). Thus CAP 413 identifies five categories of speech: Acknowledge, Clearance, Instructions, Information-information, and Information-advice.

- CAP 774. Throughout CAP 774 the CAA identifies that a Fight information Service Officer are expected to pass "advice and warnings on high-risk conflictions". The CAP 774 identifies that information, advice and warnings are given that are useful for the safe and efficient conduct of flights and intended to improve the pilot's SA; and that the information covers a range of topics including flight rules, serviceability of facilities at aerodromes, meteorological conditions, other traffic information and warning, terrain warnings, and other information likely to affect safety. Thus CAP 774 identifies three categories of speech: Advice, Information and Warnings.

From the two CAPs six potential message types or categories of communication for SA were identified: Clearances; Instructions; Acknowledgements; Information General; Information Advice and Information Warning. However, as the Clearance and Instruction types of direction are extremely similar and in this research are likely to only be implemented for the human participant, for simplicity in this research they were consolidated to form a single category of Direction. The final list of categories is provided in Table 3.1 below.

Communication Category	Description	Rank
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.	2
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.	4
Information	Provide general information to build SA that directs attention but does not necessarily require a response. Similar in scope to the Information communication of Air Navigation Services (CAP 413)	5
Direction	Provide instructions for action that the recipient must implement. This is a merge (for simplicity) of the Clearance and Instruction categories and will only be used by the participant.	1
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")	3

In order to follow the tenets of aviation CRM and standards of aviation communication, as directly applicable to the context of the desired contribution of this research, communication between the human and synthetic agent in this research was limited to messages that fall into one or more of this short catalogue of message types.

The separation of speech messages into categories also solved a practical question raised; which message should be spoken when either the system generates a new message whilst an existing message is being spoken or when the system generates multiple messages to be spoken at the same time? The solution was to ascribe a relative level of importance, a rank, to each category and then use that rank to determine which message should be played first, and which messages can interrupt and be played over.

Following the "safety first" ethos of CRM, it was determined that a Warning should be given precedence over an Information message, and a Direction should be given precedence over a Warning (as that Direction may initiate an action designed to address a Warning). Ironically this logic also indicates that an Acknowledgement should be more important than a Warning

message. The resultant hierarchy of the messages that was implemented in this research programme is shown in the ranking column of Table 3.1.

3.2.3 Message (Sentence) Structure

Continuing to follow aviation industry best practice, the CAA CRM advice of using succinct natural language in the synthetic agent messages was followed and messages were kept short, to the point, and presented in a timely manner. To reduce participant learning burden, all participant communications (Directions) were kept to one- or two-word sentences (eg "Add 1", "Prioritise 4").

To assist with standardisation of synthetic agent speech and to assist with software coding of the message generator, a model for assembling synthetic sentence structure was sought. Whilst aviation does not appear to have a standard format for sentence structure (apart from the radiotelephony practice of always starting a communication with the identification of the caller), the medical profession frequently uses the Situation, Background, Assessment, Recommendation four element model or SBAR model (NHS 2010), apparently borrowing that model from aviation along with the 5-step advocacy approach (Brindley and Reynolds 2011).

A modified version of the SBAR model was adopted for use in constructing synthetic agent sentences; an "Attention Getter" as recommended by the 5-step advocacy model identified by Brindley and Reynolds (2011) was placed at the beginning of each message, and the "Background" element was removed in an attempt to shorten the total message length. Whilst not all elements of the model are needed for all communications, where possible the synthetic agent speech generator would generate messages using the following modified SBAR structure:

- Attention Getter. Where possible all synthetic messages would start with an attention getter to ensure that the first words or content of the messages sentence was not missed due to operator attention being directed elsewhere. Examples of an attention getter are "Acknowledge" or "Warning" or "Proximity Warning".

- Situation. A report of the current activity or problem. Examples of a Situation element are "Target 1 Prioritised for clearing" or "U21 is heading towards Coventry".

- Assessment. An assessment of the current situation that could be phrased as explanation for a recommendation about to be provided. This assessment could also be a justification of actions taken if those actions do not require operator decisions or intervention. This element is where the primary reasoning transparency SA information would be placed. Examples of an assessment explanation are "as it is fully compliant, I have reduced its risk category to low".

- Recommendation. A recommendation of what should be done or what decision should be taken. Examples of a recommendation are "Recommend change to Distance Strategy to clear Target" and "I recommend I Contact it".

Finally, the aviation practice of limiting communications to simple and purposeful exchanges of information was also implemented, and in the absence of a Wizard of Oz actor or speech AI, the ability to conduct exploratory counter-factual conversations was not provided.

3.2.4 Message Content

One of the primary arguments for implementing the audio-voice communication was that it would provide the main conduit for transactive SA; it would allow the members of the HAT to proactively exchange SA data, whether that data was pushed or pulled. However, as identified in the models of SA and much of the research into SA, when one individual is provided with additional information that another is not there will inevitably be a difference in the SA of the two individuals. This is because the individual with more information has experienced an increase in interaction with the environment which both the Endsley 1995a and Smith and Hancock 1995 models predict will result in the building of SA. As the primary aim of the study was to determine if the presence and use of the audio-voice channel would affect human behaviour, not if the content of the message would affect the behaviour, it was determined to

ensure that the audio-voice messages did not provide additional SA information that was not readily available on the graphical display. If thus implemented, the silent baseline and audio-voice conditions would be SA information comparable; the participants in both conditions would have available the same core SA information, therefore, any difference in SA observed between conditions would be a consequence of the presence or absence of the communication not a consequence of the message content providing additional (non-visually available) SA information.

Additionally, for confirmation, it was determined to experimentally test whether it was the content of the message delivered rather than the mode of communication used that would affect SA. This was achieved by implementing an additional condition in the Pilot Study of "Text" where the content of the message that would normally be delivered as a spoken message was instead displayed on the screen in a text box. The text box was positioned on the screen so as to be directly visible to the participants while they carried out the main visual task (it was placed immediately below the radar screen within easy line of sight).

3.2.5 Level of Automation (LOA) and Team Structure

In this research programme, the HAT would consist of two members, the human participant and a single autonomous synthetic agent. The intent was that the two team members would work together to carry out a simulated air traffic management task. However, as demonstrated by the review of articles on autonomy's impact on SA and other HAT research (eg Endsley 2018, Calhoun et al. 2018, Battiste et al. 2018) there are many different formats or structures of teaming possible between the human and synthetic team-members that determine how tasks, workload, and responsibility were divided between them. For example, in pilotage, the team could be structured so that the human is in control of most activities including manoeuvring a vehicle and the automation simply provides separation or threat warnings. Alternatively, the team could be structured with the human and automation dividing labour between them, for

example in air traffic management the automation could monitor aircraft looking for those behaving suspiciously and passing those it considers high risk to the human to further evaluate and if necessary, process with security services.

These different team and team role structures come with different needs for communication (Schaefer et al. 2017) in the HAT which in turn would have an effect on the style and content of communication between the human and autonomous synthetic members of the HAT.

This view can be most easily seen by examining the descriptions of Levels Of Automation (LOA) such as those provided by Endsley and Kaber (1999). Whilst the LOA concept appears to have generally been overtaken by HAT, the descriptions still provide insight into broad generic forms of teaming and 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 agent at a low level "Batch Processing" LOA would pass complex strategic decision support information; if it were to it would by default have behaved as a medium level "Decision Support" LOA.

Therefore, as the fundamental structure of the team would likely affect the communication, which in turn could affect the SA and performance of the participant, it was necessary to consider establishing a consistent teaming structure for each study, especially the more advanced second and third study; that teaming structure set to optimise the effect of the independent variables being studied. To do this, it was determined to first evaluate the impact of a range of teaming structures, in the form of LOA, on the effect of communication.

It is accepted that the LOA concept has faced criticism for being prone to "imply that there are discrete levels of intelligence for autonomous systems, and that classes of vehicle systems can be designed to operate at a specific level for the entire mission" (Murphy 2012:4). This was not the intent or expectation of this research; it is not suggested that LOA identifies the only teaming structures possible in HAT; nor that a HAT will exist in its totality at a singular LOA. Rather, for

this research the LOA framework provided a vehicle to isolate and model individual generic forms of interactions found within a complex teaming relationship such as Task Lead, Task Sharing, Task (Decision) Support and Task Supervision that in real world systems could be dynamically allocated to team-members on a task-by-task basis according to a job and team description. The expectation is that the results of evaluating the impact of voice communication in these forms of interaction (LOA) could then be collectively considered, even integrated when considering the larger more complex teaming structure found in real system.

For this research, four LOA (see Table 3.2) were chosen for implementation from the Endsley and Kaber (1999) ten level model, the LOA being those that appeared to match descriptions of human autonomy teams commonly implemented by previous HAT researchers: the description of a HAT with a semi-autonomous subordinate implementing predetermined sequences of actions or "plays" by Calhoun et al. (2018) is similar to the Endsley and Kaber (1999) description of LOA3 – Batch Processing; the description of an automation able to explore variance in factors and provide suggestions for selection by the human of the system by Battiste et al. (2018) is similar to Endsley and Kaber's LOA5 – Decision Support; and automation with full delegation of a functional area with human oversight described by Demir, McNeese and Cooke (2016) is similar to Endsley and Kaber's LOA9 – Supervisory Control. These three LOA were therefore chosen for implementation, along with a baseline of LOA1 – Manual Control.

Team Structure	Structure Definition	Communication Categories
Manual Control (LOA1)	Effectively a baseline condition where 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 (LOA3)	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
Decision Support (LOA5)	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 Battiste et al. 2018). The synthetic provides the same messages as before; however, in addition the synthetic provides suggestion for the human activity.	Warning, Acknowledgement, Advice, Information
Supervision (LOA9)	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 team member communication shifts towards providing reports on what and why it is doing what it is.	Information (and if participant intervened, Warning, Acknowledgement, Advice)

3.2.6 Summary of Factors Affecting Audio-Voice Communication Implementation

In summary, for this research the autonomous synthetic agent would be implemented to meet the intra-team communication recommendations given by aviation human factors and aviation CRM. The autonomous synthetic agent would not attempt to communicate using the full spectrum of human sentences and conversation made up using natural language, but instead the scope of its communication would be limited to a plain language sub-set of English and its range of conversation limited to generating messages about the tasks being undertaken. A catalogue of possible message types was constructed and a standardised message structure (Attention Getter, Situation, Assessment, Recommendation) was used. The use of categories of the message catalogue was to be determined by the LOA (or team structure) implemented.

3.3 Situation Awareness Measurement Methodology

In addition to preparing the limitations to the implementation of the synthetic agent audio-voice communication capability, it was also necessary to consider which of the established and proven methodologies for measuring SA was to be used. As identified by Salmon et al. (2009) there are a plethora of different approaches to measuring SA and plenty of debate over which is the most appropriate for use, and it is expected that different approaches to measuring SA will likely also measure different aspects of operator SA. It was also observed that the selection of the method for measuring SA would likely effect the design of the studies and experimental apparatus (eg some techniques require the simulation to be paused and some require it to be predicable).

Three of the more commonly used (or popular) non-physical methods of assessing SA were considered for use:

3.3.1 Post Activity Subjective Assessments

One of the most popular measurement tools for measuring SA is the Situation Awareness Rating Technique – SART (Taylor 1990) which requires participants to take a survey of 10 generic questions that give a measure of their subjective opinion of their own SA. The questions are fixed and scenario (situation) independent and provide a view of the cognitive function and capacity of the individual. The questions are generally posed at the end of the scenario or trial being evaluated. The advantage of SART is that it is quick and simply to use and analyse, it is non-intrusive as it does not interrupt the task or simulation activity, and it provides an evaluation of SA that is comparable between conditions and studies (Stanton et al. 2013).

However, SART has drawn criticisms and questions over its validity, primarily because of the subjective nature of the assessment and the timing of use. Concerns raised are that the results of the survey can be affected by memory degradation and poor recall and the method only focuses on aspects of SA associated with attention, complexity and variability, all of which are more closely associated with measuring Workload than SA (Salmon et al. 2009). Furthermore,

and perhaps of greatest concern is that as a subjective measure, the operators' self-assessment of SA does not always match the reality of the situation (Jones 2000), especially in the case of poor SA, "how can one be aware that they are not aware?" (Stanton et al. 2013:264). Finally, Salmon et al. (2009:499), when experimentally evaluating the effectiveness of SART, reported "The findings from this study (and from previous research) raise doubts over the validity of SART as a measure of individual participant SA during simulated scenarios". For this latter reason, as well as the earlier reported potential for poor reliability and inability to detect poor SA, neither the SART methodology, nor in fact any other subjective assessment technique such as Situation Awareness Subjective Workload Dominance or SA-SWORD (Vidulich and Hughes 1991), will be used in this research.

3.3.2 Real Time Questioning Probes

An alternative form of measuring SA is the use of real time probes, such as the Situation Present Assessment Method or SPAM (Durso and Dattel, 2004) or, the more recently proposed Quantitative Analysis of Situation Awareness or QASA (Edgar et al. 2017). However, the use of real-time probes presents an issue with this research project as the probes themselves are effectively conversation messages and could thus directly interfere with the synthetic agent speech (and compromising the results of the study), especially if those messages are delivered close to a synthetic agent warning or recommendation for action. Furthermore, they could be confusing for the participant who may not be able to distinguish between a real time probe and a real time synthetic agent message. For this simple reason real time probes will not be used in this research program.

3.3.3 Freeze/Interrupt Probes

Arguably the most popular SA measurement technique is the use of a freeze interrupt probe, with the most popular of those being Endsley's (1988) Situation Awareness Global Assessment Technique (SAGAT). Using an interrupt technique (eg SAGAT) the participant undertakes a well-

planned linear exercise in a simulator that is periodically stopped and all visible data hidden (the screens blanked) and the participant asked probing questions to test their knowledge of the situation (their awareness). Questions are theoretically global in scope, not biased towards measurement of a single model of SA, although care should be taken to ensure that they cover the expected SA requirements of the participant (Endsley 1988). Post exercise the responses are marked, and a numerical value of SA is determined (usually a % of accuracy).

Whilst it is popular, some researchers have expressed concerns over the SAGAT method, observing that the intrusion is a distraction on task performance and that there are difficulties using the approach in a live environment (Salmon et al. 2009). Furthermore, the method is highly dependent on the quality and appropriateness of the probe questions, "the foundation of a successful SAGAT data-collection effort rests on the efficacy of the queries" (Jones and Kaber 2005: 42-2). As the questions are generated and framed by an expert on their opinion of what appropriate or "good" SA is, those questions could easily provide cues as to what is considered by the research to be "good" SA, making it difficult to achieve Endsley's (1988) goal of avoiding inadvertent bias of participant's attention. Finally, some researchers have even challenged the validity of the technique asking, "is it SA or memory being assessed?" (Salmon et al. 2009:491), although ironically, with SA being identified as a form of mental model (Endsley 1995a) and thus a form of memory, the answer to the latter question would appear to be "yes", as both are apparently being assessed.

As this research programme will be conducted exclusively on a simulator, concerns over measurement in a live environment are considered largely immaterial. Furthermore, as for this research programme it is intended to use an adaption of an Endsley and Kaber (1999) methodology for two studies that come complete with a description of the SAGAT questions to use, design of the probe questions is of reduced concern, limited only to the third and final study (see experimental design below). Finally, and most importantly, with the questions only

delivered during an interrupt period it is expected that there will be no opportunity for interference between the audio-voice communications and the SA question communications; they will be mutually exclusive in time.

Therefore, of all the three different methods of measuring SA directly, the SAGAT method is the most viable and most prepared method to be used for this research and is thus selected as the primary methodology for measuring SA.

3.3.4 Additional Measures – Performance and Situation Assessments

As discussed in the Literature Review (Chapter 2.2.1.5.4) SA is identified by Endsley (2015) to cover both the product (state of knowledge) and process of building SA (situation assessments). The SAGAT methodology only tests the state of knowledge. Therefore, it is possible to expand the scope of SA measurement to include the frequency that the participant takes situation assessments and the scope of information gathered in each assessment. These two measures could also be used to evaluate the efficacy of the participant's SA building. Whilst the measurements on frequency and direction of situation assessments does not necessarily demonstrate that information observed has been inculcated into SA, they do provide evidence of effort required to generate the SA and generate a broader picture of a participant's capability to build SA. Therefore, where possible a measurement of the frequency of situation assessments (quantitative) and queries as to the direction of attention (qualitative) will also be taken to provide an evaluation of the SA of an individual.

Furthermore, there may be an opportunity to measure SA indirectly through the measurement of decision-making or output performance as used by Lo et al. (2016) to evaluate SA when the explicit SAGAT methodology underperformed. The argument for using decision-making and performance as a measure of SA (apart from the argument of its pervious use by Lo et al.) is discussed in the conclusion on Individual SA in the Literature Review (Chapter 2.2.1.6) and has its genesis in the original Endsley (1995a) discussion and model of SA where SA leads to decision-

making and then to performance. The argument brought forward from the Literature Review is that whilst it is easy to accept that an individual decision and improvement in performance is not always the product of an improvement in SA, it would be harder to accept that the wholesale improvement of a large cohort of individuals' decision-making and performance is not the consequence of an improvement in SA. However, as no Literature had been found that provided a proof (Lo et al. 2016 withstanding) that this method of vicariously assessing SA has been used to successfully to provide a quantitative measure of SA that could be processed statistically, it was determined that this measure would be better used as supporting evidence that can inform the research about the type and extent of the SA being built.

3.3.5 Summary of Situation Awareness Measurement Methodology

As discussed above, of the three different methods of measuring SA, the SAGAT freeze probe methodology is the most viable and considered to be the most reliable at sampling dynamic SA. As explained in 3.3.1 above, post activity subjective assessment techniques such as the SART methodology will not be used due to concerns over its reported poor reliability and inability to detect poor SA (as argued by Salmon et al. 2009).

Furthermore, real time probes (such as SPAM) will not be used due to concerns that the primary probe, delivered alongside the audio-voice communication, could interfere with that communication (and thus affect the outcome of the study) and could be confusing to the participant (identifying which is a communication and which is an SA probe).

With the post-activity and real-time probe options discarded, the freeze/interrupt probe method appears to offer the most viable and valid method of objectively measuring participant SA, with the SAGAT methodology the most successful and well documented freeze probe methodology available. Therefore, for this research project, SAGAT will be used as the primary method for measuring the SA of the participants.

In addition, where possible the frequency of situation assessments will be measured and also used to quantitatively evaluate the effort required to build SA. Furthermore, qualitative information on the scope of situation assessments being conducted will be gathered during a post-study debrief in which participants will be asked to provide feedback on where their focus of attention was, thus providing context of what elements were being used to construct SA (ie what was the source of their SA rather than what was the "measure" or "value" of their SA).

Finally, evaluations of performance, in particular decision-making, and of situation assessments will also be taken to provide secondary evidence of the presence and variance in SA.

3.4   Experimental Design

Once the factors of audio-voice structure and content and SA measurement technique had been discussed and evaluated, and solutions for implementation had been identified, it was possible to consider in more detail the actual design of the research programme and the methodology to be used for the experimental studies of that programme.

3.4.1 Research Design and Approach

The aim and ambition of the research was to evaluate the effect of using the audio-voice channel of communication on the behaviour of a single human interacting with a single autonomous synthetic agent, with the intention, if possible, of basing the research in a relevant aviation setting (an aviation setting for which the results of this research can potentially provide a contribution towards flight safety). The observations that future Unmanned Aircraft Systems (UAS) Traffic Management (UTM) (FAA 2020), will be designed to be highly automated provided inspiration for an appropriate simulation scenario, that of a human and synthetic agent working cooperatively on a simulated air traffic management task.

Reviewing both non-research and research literature, it was observed that Endsley and Kaber (1999) had designed an abstracted air traffic management simulator with which they evaluated

the impact of LOA on participant SA and performance. In their articles (Kaber and Endsley 1997, Endsley and Kaber 1999) had provided a highly detailed description of the experimental apparatus and the implementation of 10 LOA, effectively providing a blueprint for a simulator that would meet the basic identified research design needs of this research programme.

However, the Endsley and Kaber (1999) simulator did have some shortfalls in that it was highly abstracted and therefore not directly applicable to modern UTM use. Therefore, it was determined to use the Endsley and Kaber (1999) abstracted simulator for the early studies of the research, but to then base the last study on a more realistic interpretation of contemporary UTM concept of operations (eg Corus 2019, FAA 2020) and modern UAV detection and display systems (eg Altitude Angel 2021, Operational Solutions 2021) for the final study.

3.4.1.1 Research Study Plan

The original research study plan was to conduct at least four studies; two would be preparatory and quite wide in scope (Pilot Studies) and two would be more applied. Unfortunately, the advent of the 2020 Covid-19 crisis and subsequent implementation of the UK Lockdown resulted in one of the applied studies being cancelled and the aims of the two applied studies concatenated into a single study.

The resultant design to answer the research questions was for a series of three experimental studies to be conducted that progressively worked through the detailed research questions as hypotheses using the two simulator designs identified; the first two studies using the abstracted Endsley and Kaber (1999) design, and the last study using the contemporary UAV detection simulation.

3.4.1.2 Pilot Study – Chapter 4

The first study was also the pilot study and was broad in scope, evaluating the primary hypotheses of the presence of the audio-voice communication from a synthetic agent having an effect on participant SA, performance and teaming, and evaluating how the teaming structure

could possibly vary those effects. In the study four LOA were implemented, with the participants experiencing interaction with just one of the four LOA. The communication interface implemented was limited to the synthetic agent speaking; the participant speech capability was not to be implemented until the second study, the Shisa Kanko (Chapter 5) study.

3.4.1.3 Shisa Kanko Study – Chapter 5

The second study expanded upon the results and findings of the Pilot study by studying the added effect of the participants talking aloud to the synthetic agent. The study was conducted as an extension and adjunct to the Pilot study, therefore, the apparatus from the Pilot Study was used, but unlike the Pilot Study, the teaming structure was set at a single LOA (LOA3) for all participants.

3.4.1.4 Applied UTM Study – Chapter 6

The third and final study was a more applied implementation of the synthetic agent audio-voice communication capability; it provided a vehicle to test the effect of more detailed reasoning audio-voice messages from the synthetic agent communication on participant SA, teaming and decision-making and it also provided a practical evaluation of the use of synthetic communication in a more realistic applied traffic control task.

3.4.2 Experimental Apparatus

3.4.2.1 Abstracted Air Traffic Control Simulator

As discussed above, for the first two studies of the research programme (Pilot and Shisa Kanko Studies) an elementary and highly abstracted air traffic control simulator was built to mimic a simulator used in study by Endsley and Kaber (1999) into Levels of Automation (LOA). The study article provided a highly detailed description of a computer-based simulator called "Multi-Task" with 10 LOA that could be replicated and modified and used to evaluate the interactions between humans and synthetic team-member engaged in different teaming structures. The description of Multi-Task was used as a blueprint to build an abstracted air traffic control

simulator for this research programme, with the implementation designed to be as near a replication as possible (within the constraints of available technology) of the original Endsley and Kaber (1999) simulator. However, for this research only four of the 10 available LOA were implemented: LOA1 (Manual Control), LOA3 (Batch Processing), LOA5 (Decision Support) and LOA9 (Supervision). The following description of the simulator, its method and timing of "processing", its scoring mechanisms etc. are all derived from the Endsley and Kaber (1999) description of Multi-Task.

The simulator was a simplified and schematic simulation of an air traffic radar station (Figure 3.1) in which a human operator and a synthetic agent form a HAT to achieve a dynamic safety critical task. The radar "sweep rate" was set at 1Hz. The team was presented with a continuous stream of square targets (aircraft) that slowly transited from the outside of a radial display towards a central deadline (red central circle Figure 3.1). The team was required to process and clear the targets off the screen before they either reached the central deadline or collided with each other.
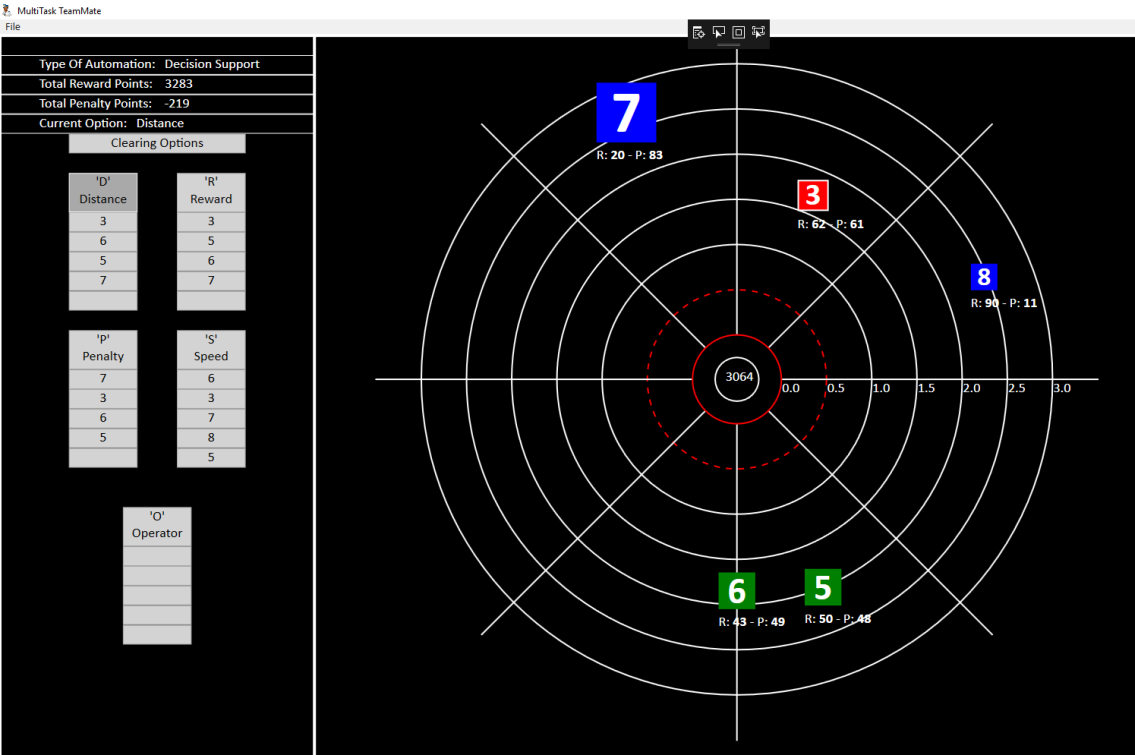
Figure 3.1: Multitask Teammate Display in Decision Support Mode (developed from Endsley & Kaber 1999)

As per the Endsley and Kaber (1999) design proscription, 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. To clear the targets in LOA1 Manual Control the human participants had to mouse-click on a target and in LOA3 Playbook and above the participant had to direct the synthetic agent to clear it. Targets being cleared were progressively reduced in size until they reached a minimum size and were then removed from the radar screen. The maximum rate at which targets were reduced in size 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 ever present on the screen at any time. As soon as a target was cleared a new target was inserted at the edge of the radial display.

Participants were given reward points for successfully clearing targets and penalty points if a target reached the deadline or collided with another target. The reward and penalty scores for each target was determined by a combination of colour and starting size as per Table 3.3 below:

Table 3.3: Reward and Penalty Points of Targets (reproduced from Endsley & Kaber 1999)

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

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

- Prevent targets reaching the deadline.

- Prevent targets colliding with each other.

- Maximise the score achieved.

To complete the task the human autonomy team had to complete three sub-tasks: a) identify a sequence for clearing targets, b) select the current priority target, and, c) process the target. Which member of the team the tasks was allocated to depended upon the LOA:

- Manual Control (LOA1) – all activities were conducted by the human. The synthetic agent provided Warning messages.

- Playbook (LOA3) – the human completed a) and b) and the synthetic agent then completed c). The synthetic agent provided Warnings, warning solution Advice and Acknowledgement Messages.

- Decision Support (LOA5) – the synthetic agent generated four options for a), the human then selected one, and the synthetic agent implemented that option as b) and c). The synthetic agent provided Warnings, warning solution Advice, strategy selection Advice, and Acknowledgement Messages

- Supervision (LOA9) – the synthetic agent carried out all 3 tasks, with the human monitoring and able to over-ride to any lower LOA. The synthetic agent provided Information Messages unless degraded by the participant to a lower LOA.

Care was taken to ensure that the graphical displays would provide the SA information that would be a part of or trigger the audio-voice speech: distance markers to show time to impact; target colour and size to show goal and penalty risk; and a warning marker to indicate when targets were too close to the deadline.

The computer-based simulator was a C# program constructed at Coventry University. The performance speed of the simulator was controlled and standardised through use of programmed timers, and the screen objects were scaled to replicate the screen sizes of the original 1999 computer screen. The graphical display for all trials was a 23-inch graphical display

at 1920 x 1080 resolution, with participants placed at a viewing distance of 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°-19° and

horizontal viewing angle of 24°-30° (see Figure 3.2)

Figure 3 2: Multitask Teammate Participant Positioning (developed from Endsley & Kaber 1999)

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 of the desktop, which was placed next to the visual display, with the volume set at 62%.

To provide the voice message as a text communication, the computer generated the audio voice message using the normal voice message decision tree, but the message was sent to a text window placed immediately below the abstracted radar display rather than being spoken aloud. The text message was displayed on the screen for the same period as it took the voice synthesiser to speak the message, thus ensuring data was presented to the participant for equal durations as both text and audio-voice.

3.4.2.2 Abstracted Unmanned Traffic Management Simulator

A slightly abstracted and gamified Unmanned Traffic Management (UTM) simulator was used for the third and final study of the research programme. The UTM simulator was designed to implement much of the Eurocontrol (Corus 2019) UTM Concept of Operations and provided a simulation of a UTM control station in which a human operator and a synthetic traffic control agent formed a Human Autonomy Team to undertake a traffic control task. The team was set the Counter Unmanned Aircraft System (C-UAS) task of managing and controlling Unmanned Aircraft Vehicles (UAVs) that appear on the screen with three primary goals:

- Preventing unauthorised UAVs from flying over specified urban and other critical infrastructure (which are identified as Control Zones);

- Report unauthorised UAVs that cannot be prevented from overflying Control Zones and place a GeoFence warning zone around those UAVs;

- Obtain as high a score as possible.

The simulated control station provided an overhead satellite map view of the West Midlands covering the urban areas of Coventry, Leamington Spa, Lutterworth, Rugby and Warwick, an area approximately 12 x 23 km (Figure 3.3). Two different types of control zone were identified:

Controlled Airspace (coloured Blue on the map) over urban areas; and Restricted Airspace (coloured Red on the map) over high value infrastructure such as airfields, prisons, and turbine windfarms. UAVs were displayed on the map as filled circles with an arrow head pointing in the direction of flight and were coloured to show their "risk" value: Green, low risk; Orange, medium risk; and Red, high risk.
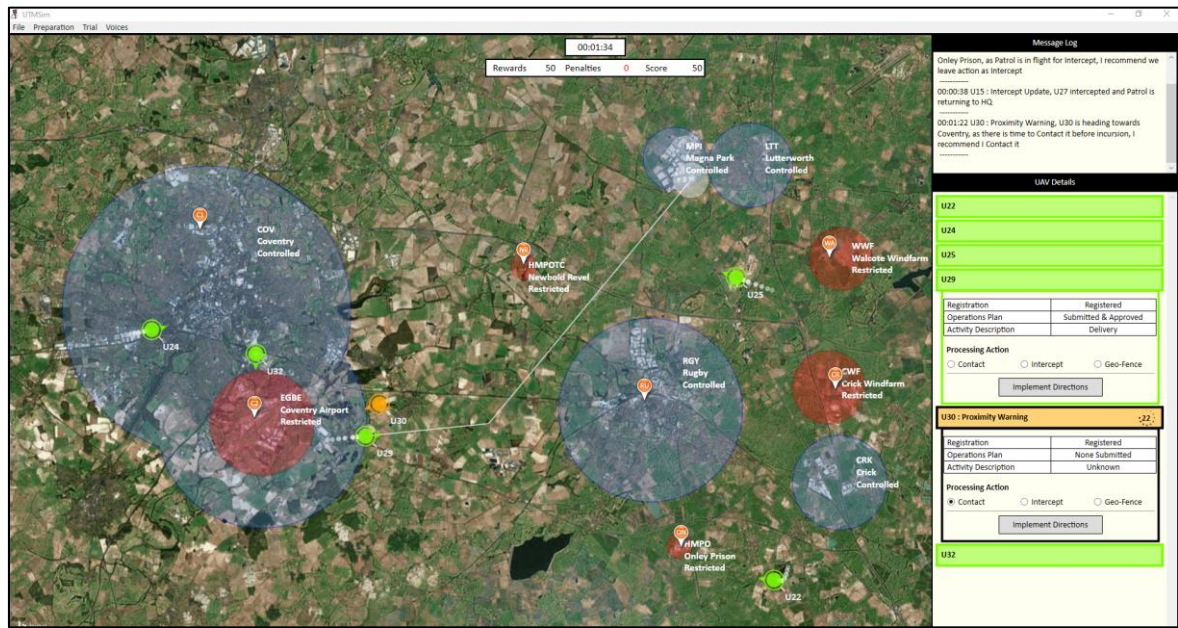
Figure 3.3: Unmanned Traffic Management Simulator. The Data Pages for U29 and U30 are open for viewing.

A UAV could be authorised to overfly a Control Zone if it was Registered and had an Operations Plan to overfly that specific zone, but unregistered UAVs and registered UAVs without operations plans could not be permitted to overfly a Control Zone. Flight strips were presented to the right of the screen as Data Bars (see Figure 3.4). Information on the registration state and operations plan for a UAV was displayed in a Data Page (see Figure 3.4) underneath the appropriate Data Bar, and Operations Plans were drawn graphically on the map as white lines and geo-zones. To view Operations Plans the participant clicked on the UAV Icon (see Figure 3.4). To view the Data Page the participant had to click on the Data Bar. When the synthetic agent was able to speak when the participant viewed the Data Page the synthetic agent also said the registration state and whether the UAV had an operations plan (as well as showing the Data Page).
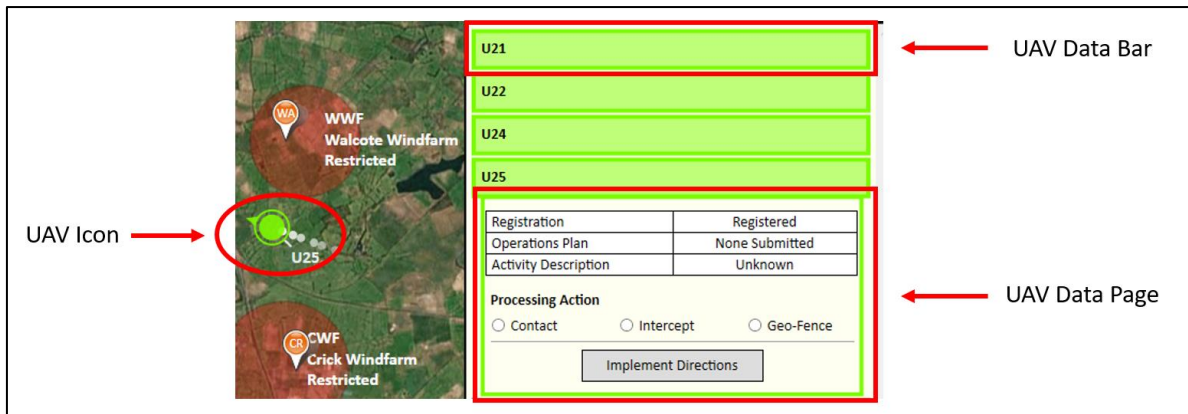
Figure 3.4: Unmanned Traffic Management Simulator, UAV Icon, Data Bar and Data Page.

As with the Multi-Task simulator, the UTM simulator was designed to provide the SA information that would be a part of or trigger the audio-voice speech: the Data Page showing UAV registration and operations plan status; a visual change in target and Data Bar colour and the automatic opening up the Data Page to provide a warning; the Data Page "processing action" radio buttons pre-set at an option to show the synthetic recommendation; and post warning the Data Page showing a rotating countdown marker with the time to incursion.

The human autonomy teaming structure was set to the relatively high LOA of Decision Support with the participant informed that the synthetic agent was to conduct the majority of the tasks including radar detection analysis, map preparation, and communicating with the UAVs. The synthetic agent would also be assessing the behaviour of the UAVs, identifying UAVs it judged at risk of acting illegally and presenting the participant with options to deal with those "at risk" UAVs. The synthetic agent would alert the participant to an "at risk" UAV approximately 30-35 seconds before the UAV illegally entered a control zone. Participants could wait for the synthetic agent to make a risk identification and then follow the synthetic recommendations, or they could pro-actively make their own identifications of UAVs at risk and make their own decision selection, or in fact any combination of the two.

The team had three options to deal with UAVs behaving illegally. They could either a) attempt to Contact the UAV pilot with the intent of getting it to retrospectively apply for clearance and

95

thus become authorised to enter, b) send a security services Patrol UAV to Intercept the rogue UAV or c) warn security services about the UAV and place a moving avoidance "GeoFence" around the drone. The Contact and Intercept actions were the highest risk options as they both took between 10-20 seconds to complete leaving the participant with limited time for a second action should the first fail.

The participant was given the team task of deciding what action to take, with the synthetic agent providing recommendations on actions as both graphical information and as voice messages provided as text and audio-voice. Once the participant had made the action selection the synthetic agent would then carry out the action process,

Team (participant) performance was scored dynamically, and the display shown on the screen, with participants being given a reward for successfully dealing with an unauthorised drone and a penalty for letting an unauthorised drone enter a control zone.

Three scripted scenarios were prepared for the UTM simulator trials designed to provide the participant with equal numbers of "at-risk" UAVs, those UAVs flying similar profiles and having similar characteristics although the sequence and exact geographic flight profile were varied between scenarios to obscure this fact to the participant. This ensured that the participant experience for each condition was consistent and thus comparable for data analysis.

To allow the implementation of a reasonably short trial period of 10 minutes but also ensure a relatively high number of human-synthetic agent interactions the UAVs were programmed to fly at relatively high speeds of between $30 - 120$ knots, approximating to the speeds of current light commercial drones and anticipated future light-aircraft drones. UAV density was set at between five to seven UAVs on the screen at a time, with only one of the UAVs at any time behaving illegally and causing an incident. UAV illegal incidents were programmed to take between 20 to 50 seconds and were provided at the rate of approximately one per minute.

3.4.2.3 Technology

As the experimental study was to be conducted during the recent Covid-19 lockdown periods and delivered on the participant's home PC, the UTM simulator had to be constructed to be more platform independent than the apparatus of the previous two studies and was designed to operate to a minimum common operating system specification, chosen in this instance to be Windows 10. The UTM simulator was purpose built for the study using Windows Foundation Presentation C# and the C# Microsoft Bing Maps library. As with the previous study the performance speed of the UAVs and refresh rate of the "radars" was controlled through programmed timers, and the screen map and objects were dynamically and programmatically scaled to ensure that they appeared on the screen as the same size irrespective of actual screen resolution, ensuring all participants had the same screen experience.

Audio voice was provided through a C# 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 neutral relative speed of 10. Text duplicates of the audio-voice message were posted in a scrollable general messages window placed at the top right-hand corner of the screen, providing participants with the ability to read current and past audio-voice messages. Voice recognition code was attempted; however, during the configuration trials considerable technical issues were encountered using the Windows 10 voice recognition software; therefore, the voice recognition was not used, and the participants entered all commands using the mouse.

3.4.3 Independent Variables

From the discussion of factors and the conclusions of the Literature Review (see Chapter 2.5) five Independent Variables (IV) were identified for evaluation.

3.4.3.1 Presence of Voice Communication

The primary independent variable was the presence or absence of the audio-voice communication capability as discussed earlier, giving two conditions of "None" and "Voice". However, as there was concern that any change in observed SA or performance could be brought about by the message content rather than the message delivery medium an additional third condition was also implemented in the Pilot Study to test whether it was the presence of the message or the medium of delivery that caused any observable variance in participant SA, performance, teaming or workload. This additional message content condition was only tested and implemented in the Pilot condition. Thus, the full range of conditions available for the Presence of Voice was None, Text and Voice (see below), although only the two conditions of None and Voice were used in the Shisa Kanko and UTM studies:

- **None**: The None condition provided the general baseline condition for all studies. In the None condition the participant communicated with the synthetic agent using keyboard and mouse, and the synthetic agent would use graphical representation alone to pass information to the participant.

- **Text**: In the Text condition the synthetic agent communicated with the participant using both the baseline graphical representation as the "None" condition and short audio-voice messages delivered as text messages in a text window. The audio-voice messages were constructed using the same message generator software of the Voice condition below to ensure that the message content was as identical and thus comparable as possible.

- **Voice**: In the Voice condition the synthetic agent communicated to the participant using both the baseline graphical representation as the "None" condition and short audio-voice messages. The participant would always communicate with the synthetic agent

using at least the keyboard and mouse, and if voice recognition was and could be implemented, through the use of audio-voice Directions.

3.4.3.2 Teaming Structure

The second independent variable was the structure of teaming which was be evaluated in the Pilot study. However, rather than set the variance as a Within-Groups (requiring a separate study), this independent variable was tested using a Between-Groups evaluation allowing it and the primary independent variable of Presence of Voice to be tested in a single Pilot study. The variable was evaluated by splitting the initial study participants into 4 groups, each of the groups working on the simulator in a different teaming structure. The four teaming structures were implemented through use of four selected LOAs:

- **Manual** (LOA1): The human conducted all task functions.

- **Playbook** (aka Batch Processing LOA3): The human and synthetic agent jointly shared the task, with the human conducting the first primary element of the task that includes decision-making. The synthetic agent did not contribute towards decision-making but would simply undertake delegated work and tasks.

- **Decision Support** (LOA5-6): The synthetic agent undertook most tasks and prepared decision options to be presented to the human for approval and selection. The human could then select those options or make their own alternative decisions.

- **Human Supervision** (LOA9): The synthetic agent undertook all tasks and decision-making and only informed the human what decisions and actions it was taking. The human could intervene by deliberately degrading the teaming to a lower LOA.

3.4.3.3 Reasoning Transparency

From the Literature Review conducted in Chapter 2 there was an expectation that in addition to the audio-voice providing SA information, the synthetic agent carrying out the simple act of

providing explanations for decisions in their communications could increase transparency and human trust in the agent. Chen et al. (2018) recommended that transparency and trust in synthetic agents can be improved by providing evidence of cognition or a rational for decisions, and Waytz, Heafner and Epley (2014) found that agents designed to be anthropomorphic and display apparent mindful decision-making engender increased trust.

As any variation in trust could have a direction effect on the teaming, and evidence of cognition and rationalisation for decisions could have an effect on SA and thus performance, the effect of the audio-voice communication showing evidence of cognition or rational thought on human decision-making was to be evaluated by establishing two different levels of message detail as separate conditions. One of the conditions would be an implementation of audio-voice with explanations and one without. These conditions could then be compared to the baseline condition of no audio-voice communication, in this instance identified as Silent:

- **Silent**: The Silent condition was the same as the None condition, just renamed to assist identify the study that was being implemented. In the Silent condition the synthetic agent only provided information graphically. This provided a baseline condition against which to compare the two audio-voice conditions.

- **Agentic**: In the Agentic condition, the synthetic agent provided an audio-voice message with a short sentence structure that did not include an explanation or reasoning element. This condition effectively tests the effect of the implementation of audio-voice without Reasoning Transparency.

- **Anthropomorphic**: In the Anthropomorphic condition the synthetic agent provided a longer detailed message that included a short reasoning explanation for deductions and decision recommendations. This tested the presence of Reasoning Transparency.

3.4.3.4 Operator Speech

The NASA Ames recommendation for implementing HAT is that all communication should be bi-directional (eg Battiste et al. 2018). However, research into Finger Pointing and Calling – FPC (Shinohara et al. 2013) indicates that when an operator is conducting FPC on their own, not interacting with a synthetic or human teammate, the act of making observations vocally aloud can positively affect attention and memory, in turn improving SA and performance. This gives the risk that when the human communicates with the synthetic agent it is the overt verbalisation of information not the communication itself that gives rise to any observed change in SA. Therefore, to isolate the effect of the communication from any potential FPC effect, the implementation of operator speech was treated as an Independent Variable, with two conditions, in both conditions the synthetic agent spoke:

- **Quiet**: The Quiet condition was the same as the None condition, just renamed to assist identify the study that was being implemented. In the Quiet condition the operator would not speak, providing instructions only using the keyboard and mouse.

- **Speaking**: In the Speaking condition the participant communicated with the synthetic using audio-voice messages.

This IV was tested in the second study, called the Shisa Kanko (Chapter 5) study after the Japanese expression for FPC.

The third study, the applied UTM (Chapter 6) study, was originally planned to be implemented with the Operator Speech set to Speaking. However, as technical issues were encountered implementing voice recognition on participants' home PCs the final study was implemented with the Operator Speech set to Quiet.

3.4.3.4 Automation Degradation

The original scope of the research project was to evaluate automation degradation in a separate applied study. However, as a consequence of the Covid-19 Lockdown and having to cancel an applied study, it was not possible to implement a separate study purely to test automation degradation. Instead, the evaluation of this IV was added to the scope of the final UTM study with only a small adaption to the methodology of that study needed.

When implementing automation failure as an IV, previous researchers have tended to model a systemic or catastrophic automation failure (eg Endsley and Kaber 1999), however, more recent research tends to focus on the human factors of operators dealing with partial failures where the autonomy continues to carry out most actions successfully but makes a partial error that is difficult for the monitoring human to identify (eg Endsley 2018, Bruder and Hasse 2020). For this research, this more recent anticipation of partial Automation Degradation rather than catastrophic automation failure was evaluated as two conditions:

- **Reliable:** In the baseline condition of Reliable the synthetic agent performed without malfunction.

- **Uncertainty:** In the Uncertainty condition the synthetic agent reported a functional failure that would cause it to be uncertain in its advice provision. From that point onwards 50% of all the synthetic agent's would be faulty.

These two conditions were implemented in the final UTM study of the research programme.

3.4.4 Dependent Variables and Measurement Techniques

The following discussion identifies the dependent variables and the measures and methodologies to be used to evaluate those variables.

3.4.4.1 Performance

The operator performance was primarily assessed by taking measures that provided observations of how successful the participants were at completing the task, at *Task Completion*. Task Completion measures included average tasks completed successfully, average rate of processing, and average rate of task failure. Much of this information was also passed dynamically to the participants during each trial run through the scoring mechanism designed to motivate participants by providing Rewards, Penalties and a Total Score. Therefore, rather than take additional nugatory measurements, the results of the scoring mechanism were used as the primary data for measuring Task Completion, along with the rate of work.

However, these Task Completion measures only identified when performance changed; they do not necessarily provide information on the cause of performance changes. To gain further insight into possible causes of behavioural changes that led to performance change two other sets of secondary performance data were gathered: measures of the decision *Selection Behaviour* of the participants (used in the Pilot and Shisa Kanko studies); and when that could not be used to measure performance (in the UTM study), measures of the decision *Processing Efficacy* of decisions.

3.4.4.2 Situation Awareness

Whilst the literature review indicated that other researchers have reported concerns about the validity or reliability of the SAGAT methodology, as it appeared to offer this research more advantages and less disadvantages compared to the other methods, and as it was, according to Salmon et al. (2009) the most accurate method of measuring SA, the SAGAT method was used as the primary SA data gathering tool for all three studies. For the Pilot and Shisa Kanko study the SAGAT questions were taken from the Endsley and Kaber (1999) description of the study methodology. For the UTM study the questions were derived from a recommendation on data needed to complete the goal tasks provided by a Counter UAV system expert (see Appendix A).

In addition to this primary tool, in all studies participants were asked a question in the post-activity debrief on aspects and information of the graphical interface they paid attention to, that data used to provide an evaluation of the scope of situation assessments made. Furthermore, in the final UTM study where the participants could be observed to request SA data the frequency of with which participants made these requests for SA data could be used as an indication of the frequency of situation assessments.

Finally, variance in participant performance and particularly participant decision-making was used to provide secondary evidence of a change in SA but not necessarily a measure of that change. Simply put if a participant was observed to have different decision-making behaviours in two or three different conditions that was taken as an indication that they also likely had different SAs in those two or three conditions; however, the observed behaviour change could not be quantified to give a measure of change in SA.

3.4.4.3 Perception of Teaming

Ideally a quantitative measure of human autonomy teaming and teamwork would be taken rather than a qualitative measure, as it is difficult to determine the relative value of the any reported participant response (what is subjectively "good" teaming to one participant could be "excellent" to another) even when using a taxonomy measure such as a Likert Scale. Unfortunately, the measure most commonly used by other HAT researchers to quantitatively measure teaming is intra-team interaction or communication behaviour, eg communication rate, communication category and content (Demir, McNeese and Cooke 2016, Sorensen and Stanton 2016). As in this research programme the communication rate, content and category are deliberately and directly being manipulated along with frequency, duration and timing, using communication as a measure would therefore (unfortunately) be invalid as it would not be objective (it could/would be set by the researcher!).

As a result, despite the disadvantages of using qualitative measures, for this research project teaming (and trust) measurements were gathered using a post-activity survey to gather participant evaluations of teaming and trust in the synthetic agent; the Perception of Teaming held by the participant. As at the start of the thesis the researcher was unable to find examples of existing and proven surveys used by other HAT researchers, an internally generated (researcher generated) survey was constructed using characteristics of teaming identified in the Literature Review as essential for synthetic team-members (eg Klein et al. 2004). This internal survey was used for the Pilot study; however, following the publication of "Measuring Effectiveness of Human Autonomy Teaming" (Lashley et al. 2019) the internal survey was replaced with an excerpt of the Collaborative Adaptive Proficiency Test Evaluation Assessment Methodology (CAPTEAM) and also the HAT Trustworthiness Assessment Protocol (both described by Lashley et al. 2019). These two surveys were used in the third and final study.

3.4.4.4 Workload

The most commonly used tool for measuring workload is the NASA-TLX (Richards 2020) and that methodology was used in this research programme. It was originally planned to use the tool for all three studies; however, in light of the poor results obtained in the Pilot Study and the length of time that taking it added to the duration of the trials (that were already long in duration because of having a large number of SAGAT interrupts and at least two surveys) it was determined not to continue its use in the Shisa Kanko (Chapter 5) and Applied UTM (Chapter 6) studies.

3.4.5 Study Implementation

3.4.5.1 Ethics

The methodology for the literature review and the three individual studies was reviewed for ethical integrity and approved by the Coventry University Ethics board. The ethics project references are P69106 for the Literature Review, P86210 for the Pilot Study, P93432 for the

Shisa Kanko (Chapter 5) study and P109048 for the Applied UTM (Chapter 6) Study. The Literature Review study was determined to be low risk and the remaining three were evaluated to be Medium Risk as they required the research to gather data on human participants. To meet the ethical requirements of this risk, all data collected was anonymised as soon as practically possible after it was processed and original information on participants identify destroyed. As part of the ethics submission process the researcher was required to carry out a project risk assessment for each study proposed.

All participants were provided with a Participant Information Sheet that provided details of the research to be conducted, gave the dates by which data was to be anonymised and destroyed, reminded them of their rights under the General Data Protection Regulation 2016 (GDPR) and the Data Protection Act 2018, and gave instructions on how to withdraw from the study if they wished and how to access information held about them. All participants were asked to complete and sign an Informed Consent Form in order to take part in the study, thus ensuring that it was recorded that all participants were volunteers and had agreed to participate in the study.

3.4.5.2 Target Audience Selection

Aviation operational professionals, particularly aircrew and air traffic controllers, are trained to communicate in very routine, standardised and predictable patterns using the basic tenets of CRM, the CAP air navigations regulations. In addition, their employers (airlines) will often have very proscriptive communication routines for standard operations. As a result, it is considered likely that these aviation operations professions would be critical to the point of sceptical of a autonomous synthetic teammate that did not follow their precise aviation standards, which would ironically vary between individuals depending upon their employment. This would present a risk of prejudicial negative bias for the studies; therefore, it was determined to target recruitment for participants away from aviation operators towards the general public. Choosing the general public as the target audience for participants also provided the benefit of

significantly increasing the overall accessibility to individuals for recruitment and simultaneously increasing the age and gender range of participants. This recruitment criteria did not preclude the recruitment of aviation personnel, but also did not limit the recruitment to only aviation industry personnel.

Details of the participant groups for all three studies has been included in the individual study chapters.

3.4.5.3 Training and Preparation

For all three studies the participants were provided with a 10 to 15 minute task briefing explaining the goals of the task, the capabilities of the synthetic teammate and the system interface (the simulator "controls"). For the first two studies this was provided by the researcher in person, but for the final UTM study, this briefing was provided as written instructions with the research contactable for questions by email or MS Teams.

After the briefing the participants were provided with five minutes of training that covered both how to use the simulator and also how to achieve the task, followed by ten minutes of practice to ensure that the participants had sufficient exposure to become task competent and were no longer learning how to conduct the task during the trials. For the first two studies this training and practice was provided and supervised by the researcher. However, due to UK winter 2020/21 lockdown it was not possible for the researcher to be present with the participants, so for the UTM study a "training tutor" module and a practice module were created and embedded within the UTM simulator.

Participants in all three studies were required to undertake and complete the training module and a practice session before they were able to undertake the trial conditions, ensuring that all participants were provided a standard preparation experience and undertook the trails in a consistent state of knowledge and capability.

3.4.5.4 Data Analysis

The simulator programs built were designed to constantly sample and save all interaction and communication data the instant the interaction occurred. Graphical data such as target position on the screen was saved at the rate of 1Hz. Data collected from the studies was then processed using a purpose build software program produced in C# that automated the data processing.

The automatic data processing programs were constructed to read all the collected raw data files, process the whole cohorts' information in the files to calculate the measures required at the sample rate required, and then collate that information into text comma separated value (.csv) files that were saved ready for statistical analysis using IBM SPSS 26®. This significantly reduced the time taken to process the large amounts of data collected and made it possible to easily correct and repeat analysis of the whole cohort if an error was found to have been made in the initial data processing algorithm or extra information was needed to provide context to results found.

The data analysis software and simulators of the studies are open source and available for future research.

Chapter 4 – Pilot Study

4.1   Introduction

The first study was designed to be the pilot study of the thesis, and as such was broad in scope in its exploration of the impact of audio-voice communication from the autonomous synthetic agent of a Human Autonomy Team (HAT) on human operator performance, Situation Awareness (SA), perception of teaming and workload. The primary focus of the study was to evaluate the impact of the two independent variables of Presence of Voice Communication and Team Structure.

As explained in the Literature Review (Chapter 2.5) previous HAT research has indicated that implementing a conversational communication interface between the human and synthetic agent of the team can improve trust, performance and situation awareness of the human team-member (eg Chen et al. 2016, Demir McNeese and Cooke 2017, Battiste et al. 2018, and Guznov et al. 2020). However, much of the previous HAT research has tended to use graphical or textual channels to achieve that communication, potentially missing out on an opportunity to lever of the anthropomorphic appearance of audio-voice communication messages and cognitive performance gains from greater cognitive timesharing postulated by multiple resource theory (eg Wickens 2008) and multicomponent models (eg Baddeley 2010).

The expectation and general hypotheses were that providing the synthetic agent of a human autonomy team with a speech capability with which to converse with the human team member would improve the likelihood that the gains of HAT would manifest. Furthermore, the new conversation channel would be the vehicle for the exchange of data that would be useful for stopping the human operator becoming "out-of-the-loop".

4.2   Study Aim and Hypotheses

4.2.1   Aim

The aim of this study was to experimentally evaluate the general hypothesis of the introduction, identified formally as:

> *Would providing an autonomous synthetic agent of a Human Autonomy Team with a conversational interface using the audio-voice channel of communication improve the situation awareness, task performance and workload of a human operator and would facilitate the creation of a teaming relationship between the human and synthetic agent?"*

4.2.2   Study Hypotheses

The overarching hypothesis for this study was that the implementation of the audio-voice would result in an improvement in the human team-member's SA, which in turn would lead to an increase in task performance and a reduction in perceived workload; therefore, the hypotheses were prepared to test those three terms (SA, performance and workload). However, as it was possible to take a near continuous measure of performance, but only possible to take infrequent discrete samples of SA, priority was given to testing performance:

> *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 it was also expected that the addition of an anthropomorphic voice would have a positive impact on the attitude of the human in the HAT and would assist that human accept the automation as an autonomous 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.

4.3 Methodology

4.3.1 Experimental Apparatus

As explained in the Methodology (Chapter 3.4.2.1), the experimental apparatus used as a vehicle for the study was a desktop simulator of an aviation task adapted from the Endsley and Kaber (1999) "Multitask" simulator and modified with the addition of a text messaging window and a voice synthesis engine that would provide the synthetic agent with the ability to provide communication messages. The simulator was located in a laboratory room at Coventry University allocated specifically to the research project for the duration of the study that allowed the participants to undertake the experimental trials in quiet without interruption or distraction.

4.3.2 Research Conditions

The design of the research was to evaluate the effect of the independent variables of Prescence of Voice Communication and Teaming Structure as a 3 x 4 condition study. The Prescence of Voice Communication was tested using a Within-Groups evaluation across three conditions:

- **None**: a baseline condition in which the participant instructed the synthetic teammate on the target processing order to follow using the keyboard with no communication from the Synthetic.

- **Text**: an intermediary condition where the synthetic agent would provide messages, such as warnings, information on synthetic activities and workload, requests for work, and recommendations for clearance strategies messages as text messages in a text window.

- **Voice**: an audio communication condition in which the synthetic agent would generate the same speech messages as the text condition, but those messages would be spoken aloud.

The Text condition was introduced to evaluate whether any change in SA and performance was a consequence of the data of the message rather than the new delivery mode of audio-voice communication. Only information readily available in the graphical interface was included as SA data in the message; that is message content was limited to data available in the visual display.

The Teaming Structure independent variable was set to be tested 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:

- **Manual** (LOA1). The human conducts all tasks.

- **Playbook** (LOA3). The human carries out part of the task (target selection) and the synthetic agent conducts the second part of the task (remaining part).

- **Decision Support** (LOA5). The synthetic agent conducts the majority of the task with the human providing strategic direction.

- **Human Supervision** (LOA9). The synthetic agent operates autonomously with the human supervising and able to intervene and override the synthetic agent, reducing it to a lower LOA.

A detailed explanation of the LOA and associated communication categories is provided in the Methodology Chapter 3 Table 3.2.

For this study the Independent Variables of Reasoning Transparency was implemented as "Anthropomorphic" (although reasoning statements were kept succinct), Automation Degradation was set as "Reliable" and Operator Speech was set as "Quiet".

4.3.3 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. No participants had visual or auditory impairments or disabilities.

The participants were randomly assigned into one of the four LOA groups (with no bias for gender or age) with six participants assigned to each group. LOA 1, 3 and 9 had 2 female and 4 male participants, and LOA5 had 3 female and 3 male participants. 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, during which time participants were presented with an average of 140 targets to clear. 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 4.1)

							Trial Post Activity	
Normal Conditions	SAGAT Period 1	Normal Conditions	SAGAT Period 2	Normal Conditions	SAGAT Period 3	Normal Conditions	NASA TLX Survey	Teaming Perception Survey
2-3 min	< 5 min Blank	2-3 min	< 5 min Blank	2-3 min	< 5 min Blank	0-4 min	No time limit	No time limit

0 min   10 min
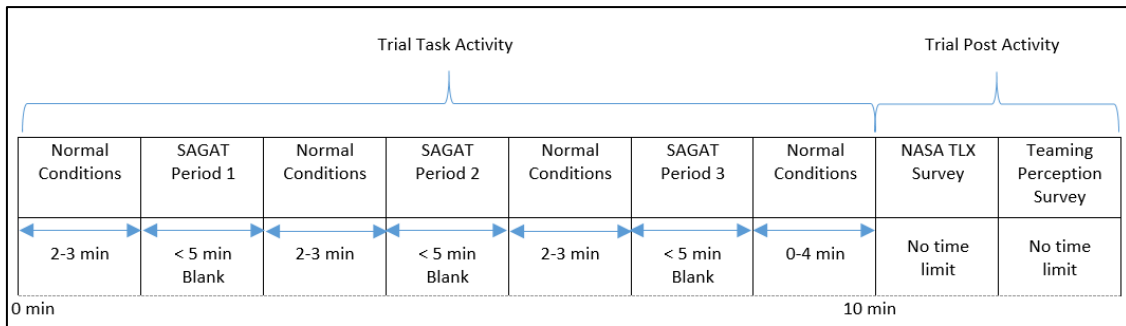
Figure 4.1: Trial Schedule Including SAGAT Freezes (developed from Endsley & Kaber 1999)

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 semi-structured debrief that gathered limited qualitative information on their perception of teaming and their SA. The participants were asked one simple question on SA "What information on the screen did you focus on and why?" and two simple questions on perception of teaming and trust: "Which of the three configurations did you prefer and why?"; "In which of the three configurations did you trust the automation the most & why?". The answers to the questions were documented. The qualitative data from this post-study debrief was then used to provide additional context to hypotheses 2 (SA improves in Voice condition) and hypotheses 4 (teaming improves in Voice condition).

4.4 Results

4.4.1 Data Sampling

In the original LOA study Endsley and Kaber had taken a sample of each performance measure every minute of their 20 minute simulation giving them 20 samples per trial for each participant. This rate of one sample per minute was implemented in this study, providing 10 samples of each measure in each 10-minute trial and thus 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.

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 *Task Outcome* performance were taken as the primary measure for performance 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 Task Completion measures, six *Selection Behaviour* measures were also taken to provide further detail on the actions that likely led to a performance. Two of the Selection 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 relatively 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 a question for each of the Endsley (1995a) SA Levels for between one to five targets: one question for SA Level 1 (Perception) asking for the size and colour of each of the up to five targets; one question for SA Level 2 (Comprehension) asking for the reward value, penalty value, distance and speed for a maximum of 4 targets; and, one question for SA Level 3 (Projection) asking for an estimate of the time to impact of one target.

Perception of Teaming was measured using a post-trial survey that asked participants to provide a subjective rating of five characteristics of teaming identified from HAT research as essential characteristics for synthetic teammates. Perception of Workload was measured using the NASA-TLX survey without pairwise comparison.

All hypothesis 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 to non-normal was not consistent across measures with some measures having a mix of normal and 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 (2016) 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. Between-Groups post hoc pairwise comparisons from the Mixed ANOVA were subject to the Bonferroni correction.

In addition, for Hypothesis 1 (audio-voice improves performance), further Within-Groups Repeated Measure one-way ANOVA were conducted for the Selection 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.

4.4.2 Confirmation of Adherence with Previous Research

Evaluating the Task Completion 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 Task Completion 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 4.2). The results provide confidence that we had correctly interpreted and appropriately modelled the Endsley and Kaber (1999) experimental apparatus.
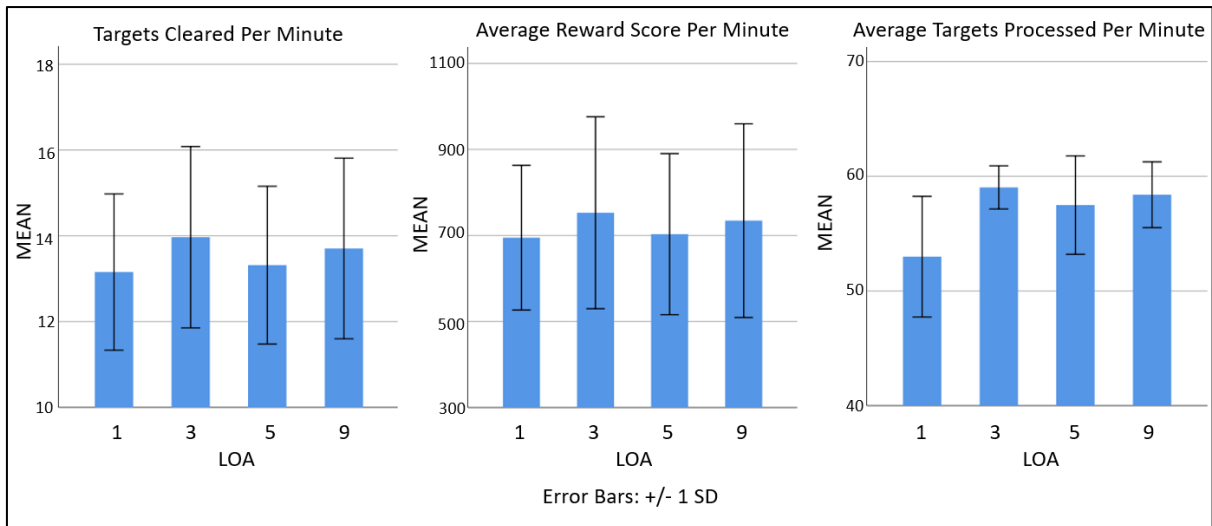
Figure 4.2: Clearance Performance Measures by LOA (taken across all Voice Conditions)

4.4.3 Hypothesis 1 - Improved Performance with Synthetic Audio Communication.

Significant difference was found for the Within-Subjects communication conditions for two of the three Task Completion 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 4.3) indicate a general rise in Task Completion performance from None to Text to Voice for all measures, supporting the Hypotheses.
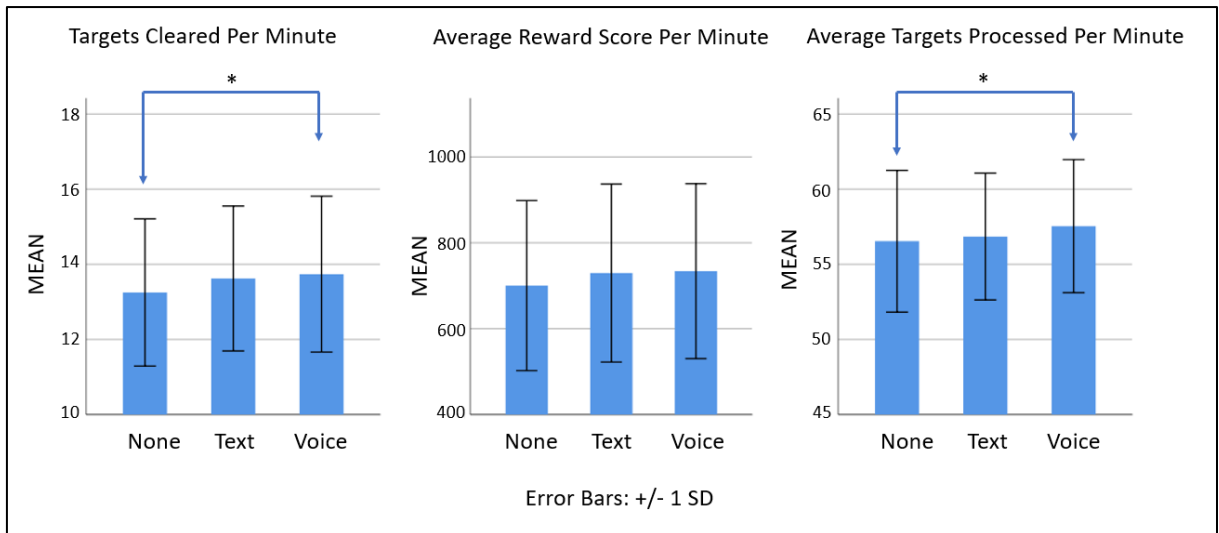
Figure 4.3: Performance Measures by Voice Condition (taken across all LOA). Statistically significant effects are indicated by asterisks

The Between-Subjects analysis also provides significant variance for LOA in Targets Cleared ($F(1,236)=6.086$, $p=.001$), Targets Processed ($F(1,236)=65.955$, $p<.0005$) and Reward Score ($F(1,236)=3.163$, $p=.025$). The Post Hoc tests using the Bonferroni correction indicate that for Targets Cleared the variance was between LOA1 and LOA3 ($p=.001$), and LOA3 and LOA5 ($p=.014$), for Targets Processed was between LOA1 and LOA3 ($p<.0005$), LOA1 and LOA5 ($p<.0005$), LOA1 and LOA9 ($p<.0005$), and LOA3 and LOA5 ($p=.008$), and for Reward Score was between just LOA1 and LOA3 ($p=.046$).

The detailed profile plots showing both Within-Groups (communication) and Between-Groups (LOA) data present a complex pattern of variances that show that the variance in Task Completion measures of performance was inconsistent between LOA, making it hard to draw conclusions. In general, the Voice provided the greatest improvement in LOA1 and LOA5 (apart from target processing where text provided the greatest improvement), and Text provided the greatest improvement in LOA3, and in LOA9 both communication conditions actually marginally degrading Task Completion performance for both targets cleared and reward score (Figure 4.4).
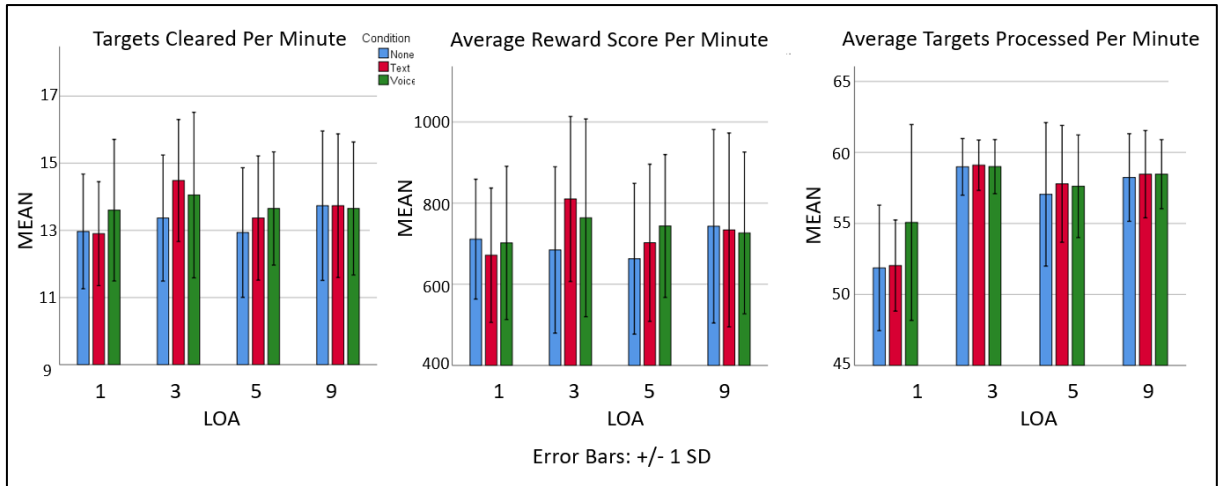
Figure 4.4: Task Completion Performance Measure by Communication and LOA conditions

To explore this inconsistency further, the Task Completion measures were separated and examined at the individual LOA. Furthermore, the Selection Behaviour measures for each LOA were also used to assist identify likely causes for the observed inconsistent variation between LOA.

4.4.3.1 LOA1 - Manual Control

The ANOVA for LOA1 (Manual Control) showed significant variance in the Task Completion 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 4.5 plot 2) shows that in targets were cleared physically closer to the deadline in the None and Text conditions than in the Voice condition.

The Selection 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 (Figure 4.5 plots 3-5)

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 4.5: Hypothesis 1 Measures with Significant Variance for LOA1. Statistically significant effects are indicated by asterisks

4.4.3.2 LOA3 – Playbook

For LOA3 (Playbook) the ANOVA for the Task Completion 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 the Selection Behaviour measures 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 4.6) for the significant measures show that in both Text and Voice, the participants increased the number of targets cleared and their total reward score and tended to clear targets further away from the centre and further away from each other.
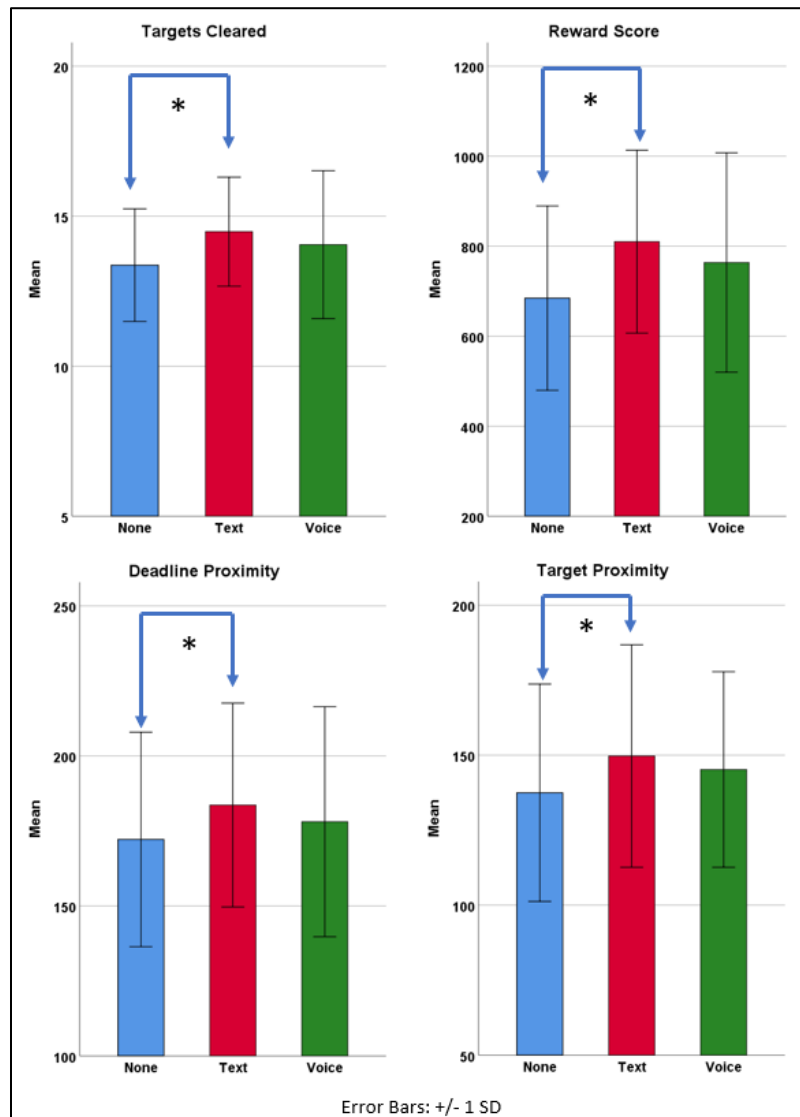
Figure 4.6: Hypothesis 1 Measures with Significant Variance for LOA3. Statistically significant effects are indicated by asterisks

4.4.3.3 LOA5 – Decision Support

In contrast to the two lower LOA, the ANOVA for LOA5 (Decision Support) did not show any significant variance in any of the Task Completion performance and Selection Behaviour risk management measures.

However, in contrast to the Task Performance measures, there was significant variance for the remaining four Selection Behaviour measures. 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 Selection Behaviour measures (Figure 4.7) 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.
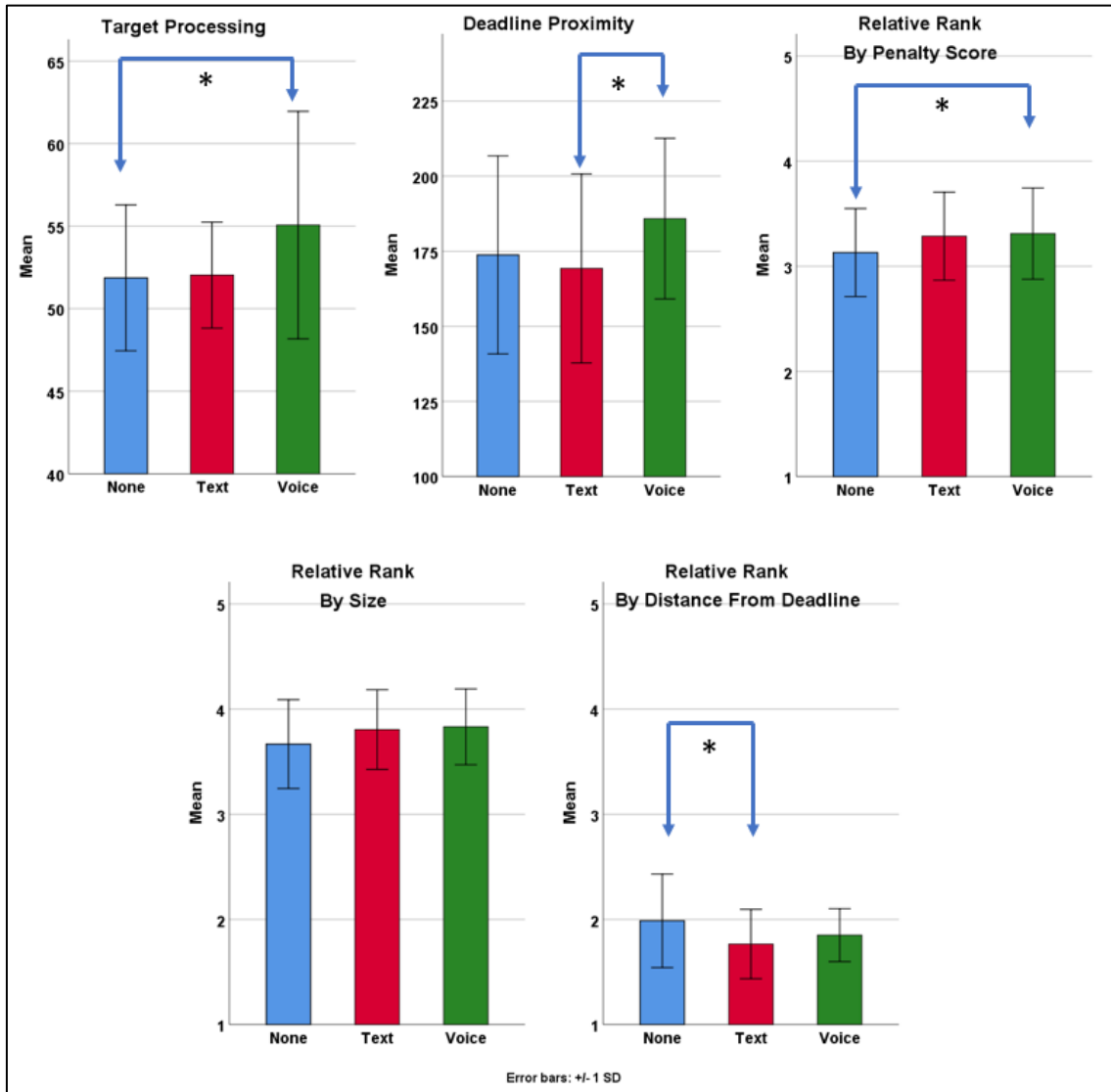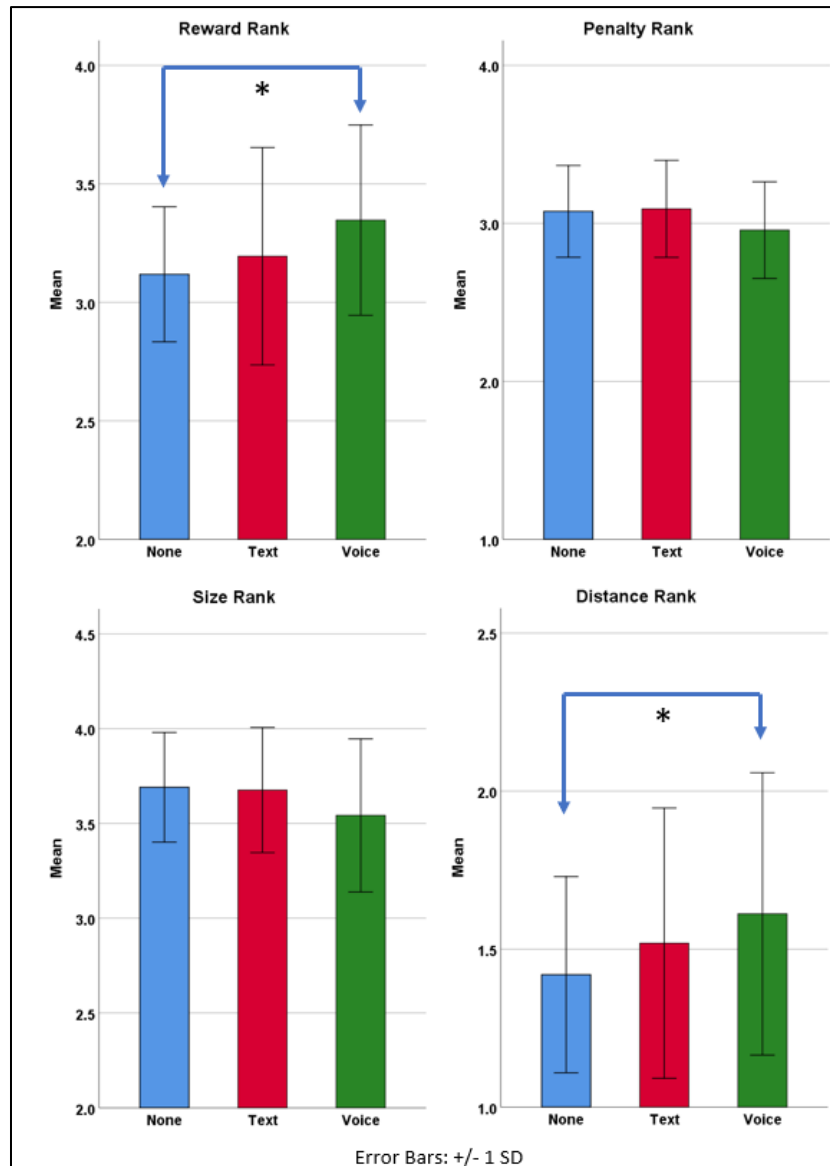
Figure 4.7: Hypothesis 1 Measures with Significant Variance for LOA 5. Statistically significant effects are indicated by asterisks

The primary difference in the scope of communication for LOA5 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.

4.4.3.4 LOA9 - Supervision Control

In LOA9 (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 Task Completion performance measures, nor for any of the Selection Behaviour measures. Thus, it would appear that in LOA9 the participants largely left the synthetic teammate to carry out all tasks without interference, irrespective of whether the synthetic was talking or not.

4.4.4 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 8) 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.
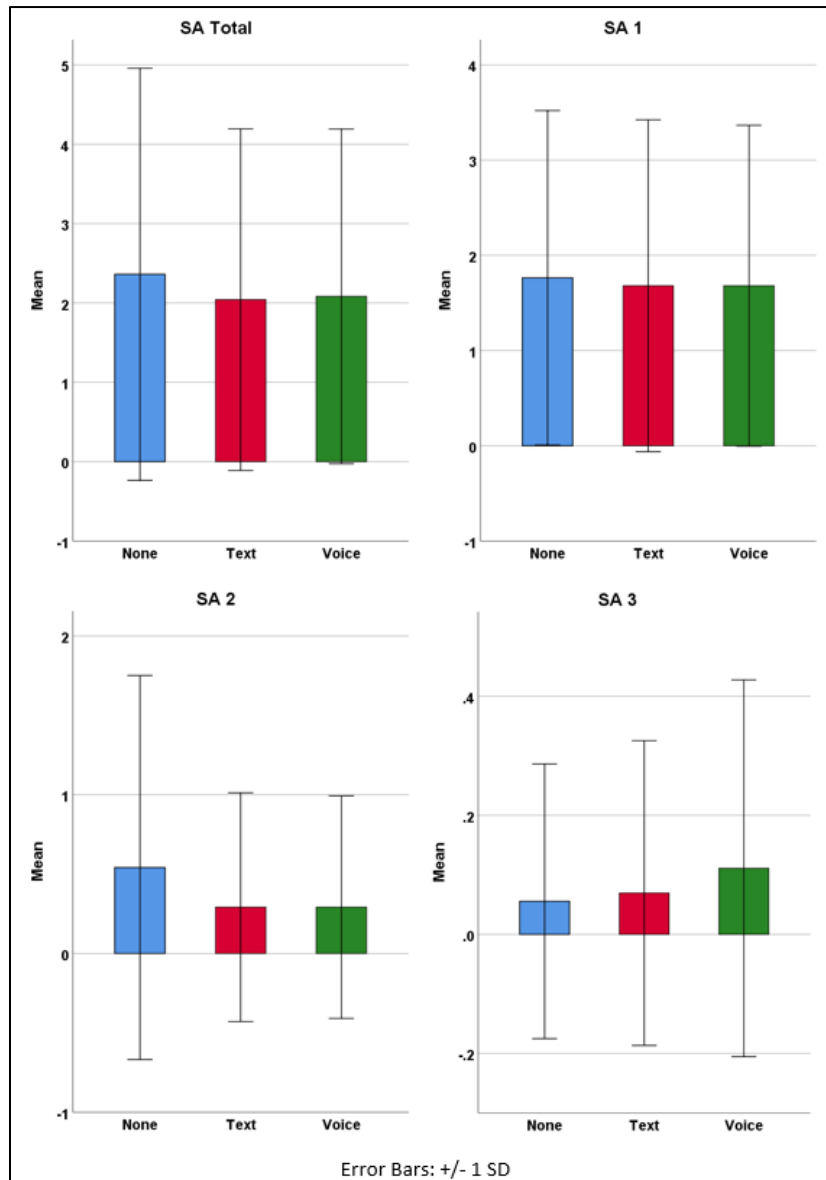
Figure 4.8: Profile Plots for Hypothesis 2 Measures

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 consider dis-regarding the SAGAT results and actually consider the hypotheses not reliably tested rather than not proven.

In contrast to the invariance of the SAGAT quantitative data, the answers to the teaming questions of the qualitative post-study debrief indicate that participants did find that the addition of the audio-voice (the Voice condition) effected and improved SA, with 18 of the participants making a positive comment about how the addition of the voice provided information that can be identified as contributing towards participant SA, either by allowing visual attention to remain focused on a particular element (eg participant comment that "allowed me to do what I was doing without checking other parts of the screen") or by providing knowledge of the Synthetic Cognition (eg participant comment that "it allowed you to hear its thought processes") or by providing attention directing information (eg participant comment that "it warned me about collisions"). Only two participants made a negative connection between Voice and SA (eg participant comment "it made me jump and distracted me"), but eight participants made a negative connection between Text and SA, often explaining that they easily forgot the messages (eg one participant commented "only remembered two messages on screen" when the data shows they received 76 messages) or that it forced their attention away from the radar screen (eg participant comment "The text assist was useless because it required diverting attention to a part of the screen that didn't help the game play") even though that only required a vertical change in view of approximately 5°.

Thus, the qualitative post-study debrief does provide some subjective evidence for a positive improvement in SA in the Voice condition compared to the other two conditions. Furthermore, accepting and following the conclusion drawn in the Literature Review Conclusion on Individual SA (Chapter 2.2.1.6), that positive variance in individual SA could be inferred from observations of a variance in decision-making behaviour that lead to a positive change in performance, the positive results drawn from Hypothesis 1 provide secondary evidence to support an improvement in SA in the Voice condition over the other two conditions. Therefore, combining these two secondary evaluations with the primary ANOVA evaluations, whilst it is determined

that the hypothesis may not be proven statistically, there is sufficient secondary evidence for some effect to warrant that further research continues to attempt to measure SA variance.

As an aside, before attempting in future studies to measure SA variance it is important to determine why the SAGAT scores were so poor. Fortuitously, the post-trial questionnaire included the question "What information on the screen did you focus on and why?" and the answers to that question provide a possible clue as to why participant SAGAT scores where poor. The majority of participants (16 participants or 66%) stated in the debrief that the focus of their attention was on evaluating the relative positioning and/or relative movement of the targets with only eight remembering focusing on the Colour and Size and even less on the actual Number, Reward or Penalty (5 participants). Thus, it appears that the information many of the participants paid most attention to could be classified as location and motion information which has been identified in cognitive psychology to be likely processed using the dorsal cognitive "where pathway" (Sternberg and Sternberg 2012), which, according to Eysenck and Keane (2015: 48) would be non-declarative: "processing the ventral stream typically but by no means always leads to conscious awareness, whereas processing in the dorsal stream does not". Thus, it was considered that the SAGAT in this study may have provided poor absolute results simply because it was being used to attempt to sample memory that was primarily none-declarative. This observation was taken into consideration when preparing additional SAGAT questions for the next study.

4.4.5 Hypothesis 3 – Reduced Workload With Synthetic Audio 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 was not supported. The profile plots for the data sets are provided in Figure 4.9.

4.4.6 Hypothesis 4 – Improved Perception Of Teaming With Audio Communication.

The mixed ANOVA of the Perception of Teaming survey results indicated 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 was 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 was supported.

The profile plot for communication (Fig 4.9) indicated 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 provided the most transparency information.
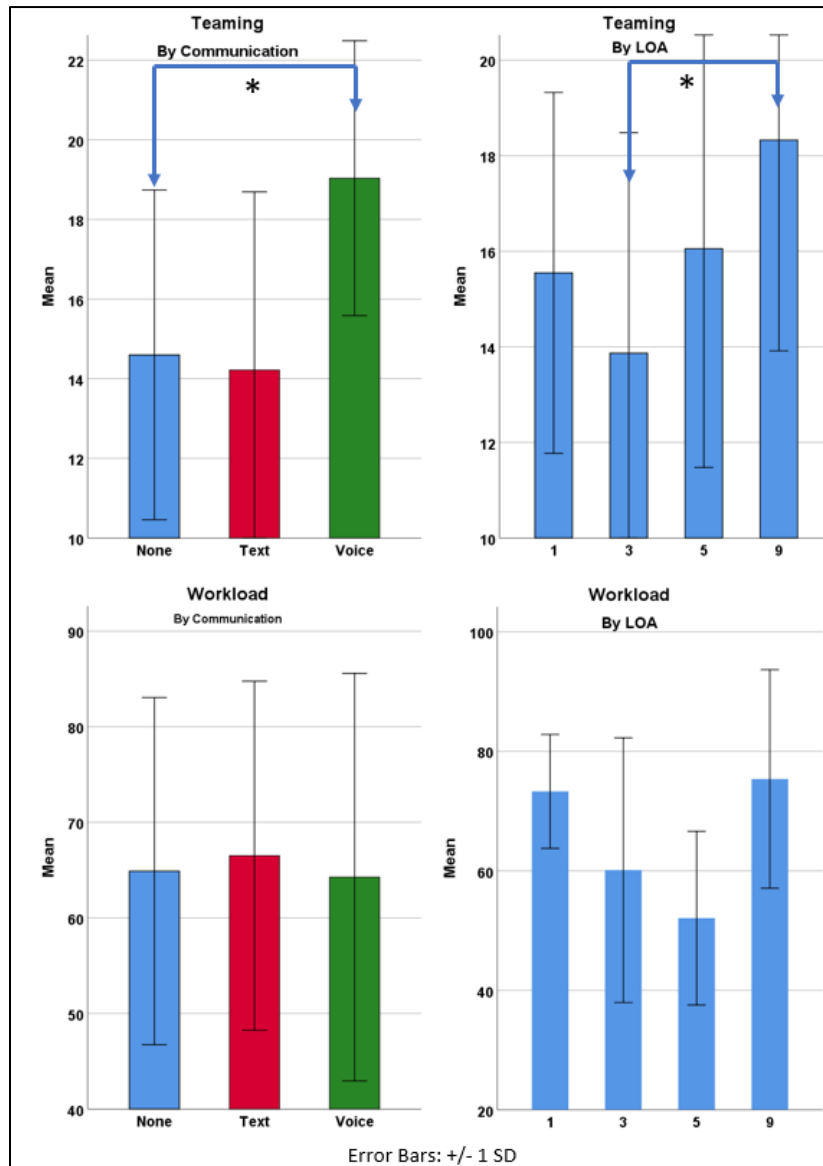
Figure 4.9 : Profile Plots of Teaming and Workload. Statistically significant effects are shown by asterisks

Additional support for the hypothesis came from the post-trial interview, where a large majority of the participants (21) expressed a subjective preference for the Voice condition over the other two conditions, and a majority (15) explained that they trusted the synthetic agent more when it had a voice. This tallied with the repeat-measures ANOVA of reported trust (F(2,46)=5.114, p=.010) which the pairwise comparisons had shown there to be an increase in trust between the Voice (M=3.942, SD=.182) and None (M=3.194, SD=.206, p=.018) conditions. Interestingly 10 participants gave responses that indicate they also saw the talking agent as anthropomorphic (eg participant comments like "because its more human" or "nice to have someone in the

130

background instead of silence") and more companionable (eg participant comments such as "just nice hearing a voice" or "get some comfort there is backup – relaxes you"). Thus, the hypothesis was supported.

4.5   Discussion

The aim of the study was to provide an experimental scenario in which a human and synthetic teammate 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. In the study performance was measured by evaluating the achievement of the three goals, in particular the achievement of the final goal which was "achieve the highest score". To achieve a high score participants needed to attempt to clear as many targets as possible and place a priority on clearing targets that offered higher reward scores (which generally meant clearing smaller targets as they would have a higher score and would clear relatively quickly).

Overall, the results directly support Hypotheses 1 and 4 in that human performance and perception of teaming were positively influenced by providing the synthetic teammate with a conversational communication channel that was 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 (1995a, 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 Controllers) have had similar issues with poor SAGAT scores and in fact even registered a negative correlation between measured SA and task performance. The observation of this effect warranted further investigation and the "Shisa Kanko" study in Chapter 5 was adjusted to include examination of this unusual phenomenon.

The pairwise comparisons of the results demonstrated 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. Whilst the profile plots (Figure 4.3) show some increase in performance due to the message content (as text), on its own the content of the message is not sufficient to significantly improve performance; the message has to be delivered through the audio-voice medium for there to be a statistically significant improvement.

Furthermore, just as Endsley and Kaber (1999) observed that the participant performance varied with the LOA (and between LOA), so our results revealed that the effect on performance of adding the voice also varied with the LOA. The results showed that there were similarities in behaviour at LOA1 and LOA3; however, they also showed that the behaviours in LOA5 and LOA9 were quite dissimilar from each other and from the behaviours of 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 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 LOA) and thus decrease the risk of making an error. This "risk reduction" behaviour was most pronounced in LOA1 where participants actively prioritised clearing the largest targets with the highest penalty (and therefore greatest risk); and potentially more balanced in LOA3 where there was no significant bias in target selection. Thus, the presence of the communication appears to have acted as a motivator for work rather than a behaviour changer.

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 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 teammate (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 LOA1 and LOA3 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 LOA5 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. However, before deciding that this behaviour represents over-trust and over-compliance, the results also show that whilst the number of recommendations accepted was high, the participants appear to retain a balanced level of trust in them, overturning 66% of those recommendations within 8 seconds in favour of their own strategy decision before any subsequent prompt or recommendation from the synthetic agent.

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.

4.6   Conclusions

The original intent of the study was to be able to determine whether providing an audio-voice speech capability to a synthetic teammate in a Human Autonomy Team (HAT) could improve the Situational Awareness (SA) of the human, 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 Task Completion 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 the somewhat novice participant group even though performance was improved. This finding was contrary to 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 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, the determination is that the 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 an individual's viewpoint, the study demonstrated that using the synthetic audio-voice capability to deliver advice and recommendations could have a significant impact on the fundamental behaviour of the human operator. In the 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 were strongly influenced by the synthetic voice, and their behaviour leads to the deducted anticipation that they would likely accept any 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 the conclusion from the study results is 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 would be of benefit to human operators of complex systems almost irrespective of LOA.

4.7   Impact on Methodology and Research Approach

The positive results obtained for performance and perception of teaming indicated that both the hypothesis and methodology for evaluation of the impact of audio-voice communications were suitable and effective at providing valuable and useable data that could be analysed successfully to discern variance in the dependent variables. Therefore, the planned study programme could continue without sever modification or review of the hypotheses.

Furthermore, the results at each of the LOAs indicated which of the LOA to select for the two remaining studies. Use of LOA1, which was effectively a baseline LOA, would not be appropriate as there is no sharing of tasks, no teaming. The apparent abdication of work and lack of variance in work rate or task behaviour for LOA9 rules that LOA as inappropriate. This left the selection for future studies as either LOA3 or LOA5. As the rate of work and Task Completion performance at LOA3 is more easily affected (than LOA5) by the whether or not the participant provides work to the synthetic agent, LOA3 was selected for the second study the "Shisa Kanko" study. The finding that the implementation of a synthetic agent providing decision support at LOA5 was highly influential on participant decision-making indicated that this LOA should be used (as planned) for the third study into the effect of audio-voice communications on human team-member decision-making behaviour. Thus, the Shisa Kanko study would be implemented at LOA3, and the UTM study at LOA5.

However, the poor results produced from the implementation of the SAGAT methodology and the lack of variance in the workload indicate that both need to be reviewed to determine whether their use should be continued.

Whilst the SAGAT results were disappointing, reflection and review of the literature on the implementation of SAGAT published after the Endsley and Kaber study (1999) indicate that the SAGAT used in this Pilot study was incomplete in that only recall questions were used and no recognition questions were included. Reviewing guidance on using the methodology further, Endsley (2004) indicates that both question types must be used to ensure that both explicit and implicit situation awareness is sampled. Thus, it was determination to continue to use SAGAT, but to modify its implementation away from the simple scope of only using recall questions as described in the original Endsley and Kaber (1999) methodology to one that included recognition questions.

Finally, reading the debrief interviews of the participants the lack of variance in the results of the NASA-TLX workload methodology were, on reflection, not unexpected as no participant made any comment about feeling any workload pressure at any LOA. As it was observed that the total duration of the exercise for the participant was considerably longer than originally expected (it took between 1½ - 2 hours to complete the exercise) and anticipating that the workload results of future studies could also likely not vary making the workload reporting of limited value, it was determined to cancel the implementation of workload measures in the subsequent studies to either reduce the participant duration or provide additional time for SA or Perception of Teaming measurement.

Chapter 5 – Shisa Kanko Study

5.1   Introduction

In the initial Pilot Study (Chapter 4) the Operator Speech was implemented as Quiet, with the participants only listening to the synthetic agent. The primary reason for this was that during the process of the Literature Review, research had been found of the Japanese practice of "Shisa Kanko" or Finger Pointing and Calling – FPC (Shinohara et al. 2013) that indicates that when an operator is conducting FPC on their own, not interacting with a synthetic or human teammate, the act of making observations vocally aloud can positively affect attention and memory, in turn improving SA and performance. Thus, there was concern that this "speaking to yourself" effect could compromise the integrity of any findings on implementing the synthetic agent speaking in the Pilot study. As a result, the requirement for the participant to speak was removed and speech was limited to the participant receiving communications from the synthetic agent as text (to test the effect of the message content alone) or as audio-voice messages (testing both message and mode of delivery). Instead of giving directions as speech, the participant provided directions to the synthetic agent through use of the mouse (LOA1) or keyboard (LOA3, LOA5, and LOA9).

The results obtained in the Pilot Study (Chapter 4.4.3) were positive and indicate that receiving audio-voice communication did have a significantly positive effect on performance and perception of teaming. The results had shown that the synthetic agent providing audio-voice warnings, information and recommendations to solve problems would encourage the participants to improve their safety orientated performance; they would work harder and faster and focus on activities that would reduce the risk of goal failure. The results had shown that they could also be influenced in their decision-making by recommendations provided as audio-voice messages.

However, the results of the Pilot study had also shown that the content of the messages (delivered as text) on its own did not significantly improve performance or perception of teaming; the message had to be delivered through the audio-voice medium for there to be statistically significant improvement in participant performance or teaming.

The Pilot study results had also shown that, contrary to expectations, neither the delivery of the content of a message (text), nor the delivery of that message by audio-voice had any significant effect on SA. In fact, the graphical plots of SAGAT results (Figure 4.8) had indicated a marginal but non-significant reduction in SA when the synthetic agent provided information as a message (text) and when it delivered that message as audio-voice.

With knowledge now present that that SA had not been affected by either message content or the synthetic agent talking, it was now known that it was possible to observe (and thus test) whether getting the participant to talk would affect SA; with the other conditions kept constant but Operator Speech varied, any change in SA would likely be a result of the participant talking. Furthermore, the performance results of the Pilot study had provided an indication of the variance in performance to be expected due to the synthetic agent talking. Therefore, it was known that any variance in performance other than the predicted variance would likely be the result of the participant talking. Thus, it was now possible to evaluate the IV of Operator Speech and test the effect of audio-voice communication from the participant to the synthetic agent on participant SA and performance.

5.1.1 Background: Implicit and Explicit SA

As explained in the introduction, the Pilot Study had successfully demonstrated variance in the participants' Task Completion performance and subjective perception of teaming; however, no statistical variance in SA had been found between any of the conditions (None, Text, or Voice) although it was argued that the observed change in performance was indicative of a likely change in SA. In contradiction of expectation neither the addition of potentially extra

information (the message content), nor its delivery as an audio-voice message had any effect on SA.

In fact, perhaps surprisingly, the results from the Situation Awareness Global Assessment Technique (SAGAT) questions were actually consistently sparse, with many left unanswered, and in a post-activity debrief almost all participants complained that when undertaking the SAGAT they found that they struggled to naturally recall any SA information they were asked; they simply could not remember information about the targets (where they were, what colour they were, what size they were etc). This observation was backed by statistical analysis of the SAGAT results which showed that the average participant answer rate was approximately 2 ± 2 bits of information (M=2.16, SD=2.290).

Lo et al. (2016) reported a similar outcome when using the SAGAT to measure SA in simulator-based research of a dynamic train traffic control task. They obtained low absolute values for SAGAT answers but did find some variance in SA between conditions and were able to demonstration correlations between the SAGAT scores and performance (although the correlation switched from positive in their pilot study to negative in their main study) showing that despite the low absolute scores SAGAT was successful at providing a relative measure of SA. In their discussion Lo et al. (2016) argue that the low absolute scores are not a reflection on the SAGAT methodology, but rather are explained by the presence of implicit SA.

The proposal that some aspect of SA could be implicit and non-declarative is not new; in an early discussion on progress with the use of the SA model and SAGAT, Endsley (2004) acknowledges the presence of implicit SA. Furthermore, modern cognitive psychology appears to offer support for the presence of implicit SA. The dual process theory of cognition proposes that all reasoning is conducted as a duality of heuristic and analytic processes (Evans 2006), identified respectively as unconscious (non-declarative) and conscious (declarative) processing (Evans 2019). Examples of such systems are the dual visual processing systems identified in the human brain; the ventral

"what pathway" and the dorsal "where pathway" (Sternberg and Sternberg 2012). The ventral "what pathway" is primarily declarative and the "where pathway" is not; Eysenck and Keane (2015: 48) explain "processing the ventral stream typically but by no means always leads to conscious awareness, whereas processing in the dorsal stream does not". Furthermore, cognitive psychology also indicates that not only are some physical parts of the brain naturally non-declarative, it is also possible through repetition to make the cognitive processes used in the brain non-declarative. For example, the more practiced that an individual becomes in executing a procedure or a task, the more the cognitive processes used to guide that activity are automated and habituated into a procedural skill, using non-declarative procedural memory (Sternberg and Sternberg 2012).

Reflecting on the relatively easy and repetitive tasks of the simulator used for the Pilot Study (watching targets move, continuously assessing their relative positions to each other and a fixed location, and selecting a target or series of targets for processing) and examining the qualitative data (indicating that 66% of participants took SA cues from the relative location and motion of the targets) it was retrospectively hypothesized that participants of the Pilot Study (Chapter 4) were likely making primary use of the cognitive dorsal "where pathway" (non-declarative) and were also able to habituate those tasks (making them non-declarative), with the overall result that a substantial proportion of the information used for cognitive processing could have become non-declarative which would in turn lead to SA becoming implicit. The suggestion that much of the SA of the participants in the Pilot study was thus implicit would go some way to explaining why the recall questions used in the Pilot Study and by Lo et al. (2016), that would primarily sample explicit knowledge, had such low absolute values.

Thus, with the results and Literature indicating that the reason for the low SAGAT scores was that a potentially substantial portion of SA was implicit, the thought of the researcher was that to successfully measure SA, either an alternative method had to be found to that would sample

implicit memory, or a vehicle had to be found to assist the participant improve SA by converting implicit SA into explicit SA that could then be used to answer SAGAT recall questions.

Ironically, the very research on "shisa kanko" or Finger Pointing and Calling (FPC) that prompted the researcher not to implement participant speech in the Pilot Study, offered a potential solution for this latter thought on attempting to convert implicit SA into explicit SA. Shinohara et al. (2013) explain that in the safety critical awareness FPC method human operators use hand movement to force attention to a "target" and deliberate vocal declaration of observed information to facilitate the memorisation and recall of safety critical information pertinent to that target. The aim of FPC is that by deliberately requiring operators to memorise and recall information they will become more conscious of their environment, their state of progress through a procedure, and the actions that they have left to do. The intent is to improve the safety behaviour of the operator, slowing down their processing, making them focus harder on environmental cues, and decreasing the risk they will miss a stage (Shinohara et al. 2013).

Considering this effect of improving conscious memory and attention through overt verbalisation, it was theorised that the deliberate vocalisation of observations might assist transfer knowledge that is routinely unconscious (either because it is being generated in non-declarative portions of the brain (eg Eysenck & Keane 2015, Evans 2019) or because is automated and habituated as a procedural skill (Sternberg & Sternberg 2012)) into conscious working memory. Therefore, it was determined that there was a likelihood that making the operator pass information to the synthetic agent would result in the operator having better conscious and explicit access to that information.

5.1.2 SAGAT and Measuring Implicit SA

As discussed above, with the deduction that the likely reason for the low SAGAT scores was that the participant's SA had been largely implicit, an alternative to using recall questions had to be found to that would sample implicit memory and thus implicit SA.

Ironically, Endsley, despite using only recall questions in the methodology that was replicated for this study (Endsley and Kaber 1999), in a later publication indicated that SAGAT should be able to gather data on implicit memory through the "use of recognition and not just free recall in the probe response measures" (Endsley 2004: 335). Thus, for this current study it was determined to add recognition questions to attempt to directly sample and measure the participant's implicit SA.

5.2   Study Aim and Hypotheses

5.2.1 Aim

The primary aim of the study was to evaluate the effect of implementing human operator (participant) speech to a synthetic teammate in a HAT, on the SA of the human operator. In addition, the study also aimed to test the use of recognition questions in SAGAT interrupts to evaluate participant SA.

5.2.2 Hypotheses

Taking the primary aim of the study and the discussion above three primary hypotheses were proposed to be tested using data collected during the trials. Two of the hypotheses were directly linked to the aim, the test of any change to SA. However, as the results of the Pilot study had shown no variance in SA despite a variance in performance, and Endsley (1995a) had argued that there should be a correlation between SA and performance, it was determined to also evaluate performance to provide a secondary measure of change in SA should the SA measurement results again prove to be inconclusive.

The hypotheses were:

> *Hypothesis 1:   Participants Speaking audio-voice messages to the synthetic agent will have improved SAGAT scores over participants in the Quiet condition.*

Hypothesis 2:   Participants Speaking audio-voice messages to the synthetic agent will obtain higher scores for SAGAT questions covering the SA data communicated than SA data observed.

Hypothesis 3:   Participants Speaking audio voice messages to the synthetic agent will have improved task performance.

5.3   Methodology

5.3.1 Changes to Experimental Apparatus

The same experimental apparatus used in the Pilot Study was used in this study but was set fixed at LOA3 (Playbook). The synthetic agent was provided with a new voice recognition capability. This allowed the synthetic agent to convert participant speech into text data that could be used to trigger the same directions provided by a keyboard or mouse click. Participants were provided with a USB "wired" headset with headphones and microphone, which was adjusted for volume (hearing and speaking) during a "set-up" period prior to the practice and trials.

In addition, the communication capability of the synthetic agent at LOA3 was expanded to include the Information category (not normally included at LOA3 – see discussion in the Methodology Chapter 3.2.5), thus giving the synthetic agent all four categories of communication: the ability to Acknowledge instructions; to Warn of risk situations; to give Advice on which target to prioritise and to give Information on which target was being processed.

Finally, in order to allow the synthetic agent to be able to complete all of its messages without interrupting itself, the overall simulation in both conditions was slowed down by 20% by increasing the radar clock rate from 1s to 1.2s (simulated time for a radar sweep). This was necessary as the synthetic messages (eg "Processing One Red Large") took slightly longer than 2 seconds to say aloud and with the clock rate set at 1s the synthetic agent would interrupt itself

when it cleared a small target (which would clear in just two seconds). Providing the synthetic with 2.4 seconds gave it enough time to speak all processing Information messages and allowed for a short pause before the next message.

The change of clock speed had no effect on performance rate comparisons within the study, as those comparisons are conducted relatively between conditions (rate of clearance in one condition verses rate of clearance in another condition) not absolutely. However, as the clock rate of the simulator for this Shisa Kanko study was now different to the clock rate of the Pilot Study simulator it was not possible to directly compare the data sets of LOA3 from the Pilot Study directly to the data sets generated from this study using statistical analysis tools such as ANOVA in SPSS. However, comparison of overall variance effect was possible, as both studies had the same (effective) baseline. Thus, observations on the performance variance between the "Voice" and "None" conditions of the Pilot Study (Chapter 4) were compared to the performance variance between the "Speaking" and "Quiet" conditions of this study (in the same way results of each LOA could be compared to each other in the Pilot Study).

5.3.2 Research Conditions

The study was to continue to evaluate the impact of Presence of Voice, but also evaluate the effect of the IV of Operator Speech. The IV of Teaming was set at LOA3 "Playbook", Reasoning Transparency was set at "Anthropomorphic", and Automation Degradation was set as "Reliable".

The study was conducted as a Repeat-Measures with two participant-to-synthetic communication conditions: Quiet and Speaking (which included the Presence of Voice conditions of None and Voice respectively). For practical timing limitations, as the results of the Pilot study had been that the presence of the message content on its own (the Text condition) did not cause a significant improvement in SA, performance or teaming, the use of the Text condition was discontinued.

The conditions finally used were:

- **Quiet**: a baseline condition in which the participant instructed the synthetic teammate on the target processing order to follow using the keyboard with no communication from the Synthetic.

- **Speaking**: as well as the synthetic agent speaking, the participant also spoke, providing instructions on which target to process. The synthetic agent would acknowledging instructions, provide safety warnings, advise on which target to process to negate a warning, and would inform the participant when it started processing a new target. The voice messages included goal specific data. The participants had to include colour information to their commands, eg to task the synthetic to process target one the participant would say "add one red". The synthetic teammate communications included the SA information of colour and size eg "acknowledge, adding one red large".

The Colour and Size data were selected for addition to the human and synthetic speech as the combination of colour and size determined the reward and penalty value of the target (see Chapter 3 Methodology Table 3.3) which in turn directly affected the risk of the target to the achievement of the three goals the participants were given. The anticipation was that an improved awareness of the colour and size would lead to a better understanding of the current risks through improved SA.

Through this process of deliberately articulating colour and size SA information the principles of FPC were implemented, with the expectation that information that in the pilot study had been just seen and thus would have been implicit information (implicit SA) would in this study be vocalised or heard and would thus be explicit knowledge (and explicit SA).

5.3.3 Participants

In total 26 participants aged between 21 and 52 (M=32.27, SD=9.569, 11 female and 15 male) took part in the study. None of the participants had previous air traffic control experience. No

participants had visual or auditory impairments or disabilities. None of the participants reported any difficulty comprehending the task requirements. Participants were volunteers taken from Coventry University staff (administrative and academic) or students from a wide range of disciplines (eg Social Sciences, Psychology, Information Technology, 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. All participants undertook both conditions, the sequence of conditions randomly assigned from a 26 x 2 Latin-Square Matrix.

Each participant was provided with a task briefing and explanation of conditions and role of the synthetic teammate, and then given five minutes of training in the task and the SAGAT interrupts. The participants were then given five minutes of practice in each of the two conditions before conducting the two trials. Most participants initially found saying the target colours aloud in the Speaking condition unusual and quite difficult, but all quickly learned how to do so and were fluent (skilled) by the end of the Speaking condition training session having made an average of 65 practice vocalised statements.

After the last trial, post-study, the participants were asked to complete a written debrief of their experience. The debrief posed three semi-structured questions designed to elicit the subjective opinion of the participant on their SA. The first two questions gathered information on how the participants had built SA, asking them "What information did you focus on and why?" and "Which of the task goals were your priority and how did you ensure that you achieved them?". The third question attempted to gather opinion on the relative workload required to generate SA in conditions, asking "In which of the two trials did you find it easier to answer SA questions and why?". The qualitative data from the questions was then used to provide additional context to Hypotheses 1 (improved overall SA in Speaking) and 2 (in Speaking, better SA for spoken than observed).

5.4 Results

5.4.1 Data Collection

Situation Awareness data was collected using the SAGAT methodology. Following advice from Endsley (2004), recognition questions were used to gather data on implicit SA and recall questions on explicit SA. Two recognition questions were posed: one querying recognition of the radar screen (eg Figure 5.1) and the other recognition of the processing queue. Each recognition question provided the participant with four screen capture images of a section of the screen, one of which had been taken immediately prior to the interrupt.
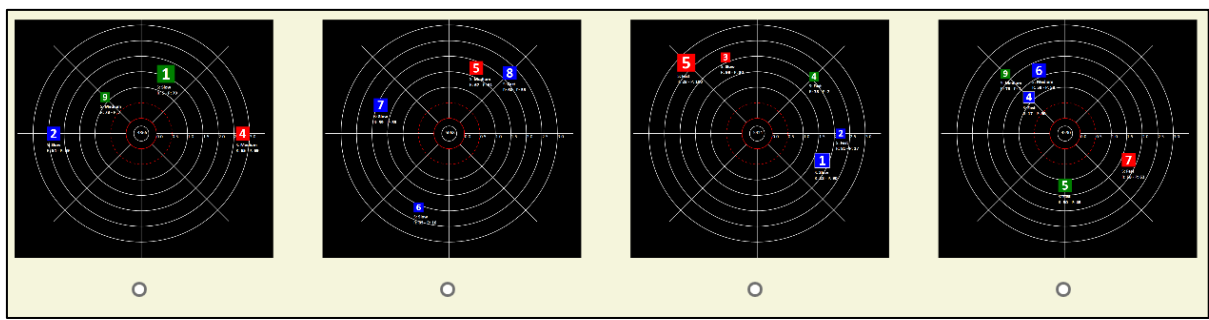
Figure 5.1 Example of SAGAT Recognition Question for Radar Screen

Three SAGAT recall questions were asked, the questions taken from the original study (Endsley & Kaber 1999 : 475) that the methodology had been adapted from:

> the queries [..] asked for: (1) target colour and size identification for each target (level 1 SA); (2) four questions concerning the reward, penalty, distance and speed of targets (level 2 SA); and (3) a single question concerning when a target would reach the deadline at the centre of the display (level 3 SA).

As the recognition questions showed an image of the radar screen and targets, they also showed the primary information of the recall questions (the size, colour, location and speed of the targets); therefore, the recognition questions could not be used in the same interrupt as the recall questions. As a consequence, a total of six SAGAT interrupts were conducted; three interrupts for recognition questions and three interrupts for recall questions. The two types of

148

interrupt were interlaced alternatively, the first interrupt recognition, the second recall, the third recognition etc. and occurred approximately one to one and half minutes. The resultant trial duration was 11 minutes. The same format of recall and recognition question was used for each interrupt.

The task performance measures of the Pilot study were also used for this study. Performance was measured continuously and then summarised for each minute of the trial. Three measures of *Task Completion* performance were taken: the average number of targets cleared, average reward score per minute, and the average amount of target processing per minute. In addition, six *Selection Behaviour* measures were taken: two risk management measures (average target distance from deadline or Proximity to Deadline when cleared, and average target distance from other targets or Proximity to Targets when cleared) and four target selection behaviour measures (the relative reward value, penalty value, size and distance from deadline ranked against the other targets on the screen at that time).

The data sets gathered from all performance measures and SAGAT were analysed through SPSS to test for significant variance between conditions. The Shapiro-Wilk test for normality was applied to all data sets. All SA and almost all performance data sets failed the test; therefore, all variance evaluations were conducted using the non-parametric Friedman for multiple related samples and Wilcoxon signed-rank test for pairs of related samples. As recommended by Fritz, Morris and Richler (2012), Cohen's r was calculated from Z and N values gathered from the Wilcoxon signed-rank tests to provide numerical and relative effect size (Cohen 1992: 157, Table 1).

5.4.2 Hypothesis 1 - Improved SAGAT Scores When Speaking.

The hypothesis was tested using both recognition and recall questions with an expectation, as per Endsley (2004) and our earlier "SAGAT and Measuring Implicit and Explicit SA" argument,

that recognition questions would serve to sample implicit SA and recall questions would sample primarily explicit SA.

5.4.2.1 Recognition Questions (Implicit SA)

The scores for recognition questions were consistently high, with 100% of questions attempted and a mean score of 84% (M=.840, SD=.337) in the Quiet condition and 86% (M=.859, SD=.254) in the Speaking condition. No statistical variance was found for recognition answers between the two conditions overall (Z=-.614, p=.539), or for each of the two questions (Q1 Recognition of Radar Screen Z=-1.604, p=.109, Q2 Recognition of Processing List Z=-.577, p=.564).

5.4.2.2 Recall Questions (Explicit SA)

The scores for the recall questions were considerably lower than those for recognition questions, with just 57% of all recall questions attempted and an overall average score of 13% correct in Quiet and 17% in Speaking. However, before determining that these results show that participant recall was poor, it must be observed that the six recall questions asked participants for an average of 25 individual facts, and that the number of facts participants could recall (approximately 3±3 facts in Quiet, M=3.27 SD=3.17 and approximately 4±3 in Speaking, M=4.22 SD=2.82) fits the theorised expectations of working memory of around four blocks of information as discussed by Baddeley (2010).

Significant variance in recall SAGAT scores between the two conditions was found for overall SA (Z=-3.464, p=.001) and Level 1 SA (Z =-3.347, p=.001). In both cases the Cohen's r was between .3 to .5 and therefore identified (using Cohen 1992:157, Table 1) as a Medium effect (overall SA, N=78, r=.3922, SA Level 1, N=78, r=.3790) with the average score for both higher in the Speaking condition than in the Quiet condition (SA Overall: Speaking M=.169 SD=.114, Quiet M=.134 SD=.127; SA L1: Speaking M=.331 SD=.196, Quiet M=.249 SD=.225) as shown in Figure 5.2. However, there was no variance identified for Level 2 explicit SA (Z=-.972, p=.324) and Level 3 explicit SA (Z=-.816, p=.414) with both having very low average scores of between 4%-7%.
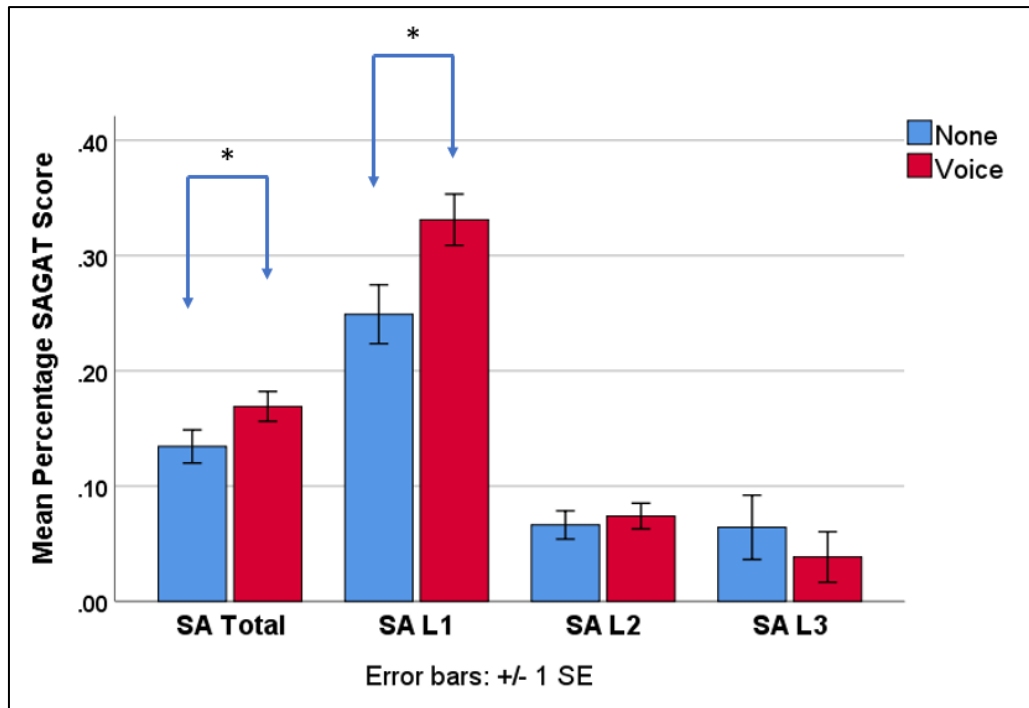
Figure 5.2: Situation Awareness Recall Scores. Statistically significant effects are indicated by asterisks*

The qualitative data also supports the hypothesis with the majority of the participants (20 participants or 77%) stated that they found it easier to answer SA data in the Speaking condition, with 11 of the participants (42%) explaining that the speech helped them remember data for the SAGAT questions.

Thus, with implicit SA remaining constant (and consequently not degrading) and explicit SA improving, the general hypothesis that the operator Speaking improve overall SA is supporting, with the improvement lying exclusively in explicit aspects of SA.

5.4.3 Hypothesis 2:   Higher SAGAT Scores For SA Data Communicated.

5.4.3.1 Specific SA Questions

In total seven recall SA questions were posed, asking the participant to recall information about Target Colour, Size, Distance from Centre, Relative Speed, Reward, Penalty and Time to Deadline. The data on only two of those questions was articulated in the Speaking condition, with the participant stating the Number and Colour of the target to be processed. When the synthetic agent spoke, it would state the Number, Colour and Size of the target it was discussing.

151

Thus, the SA data deliberately communicated was Colour and Size and the remainder (Distance, Speed, Reward, Penalty and Time) of data for the SA questions was observed.

A Wilcoxon test for each of the seven SA questions between the two conditions indicates that there was significant variance between the conditions for the answers to the Colour question (Z=-2.810, p=.005) with the Cohen's r indicating a Medium size effect (N=78, r=.3182) with the average Colour score improving from 32% (M=.319 SD=.311) in the Quiet condition to 42% (M=.420 SD=.232) in the Speaking condition, an overall improvement of 31%. Similarly, significant variance was found for the Size question (Z=-2.286, p=.022) with the Cohen's r indicating a Small to Medium effect (N=78, r=.258) with the average Size score improving from 18% (M=.179 SD=.207) in the Quiet condition to 24% (M=.242 SD=.245) in the Speaking condition, an overall improvement of 41%. There was no significant difference between conditions for answers to any other SAGAT question (see Figure 5.3).

The greatest relative improvement (as opposed to absolute value change) in recall occurred for SA data heard (Size, only spoken by the Synthetic) over SA data said and heard (Colour said by both participant and synthetic agent).
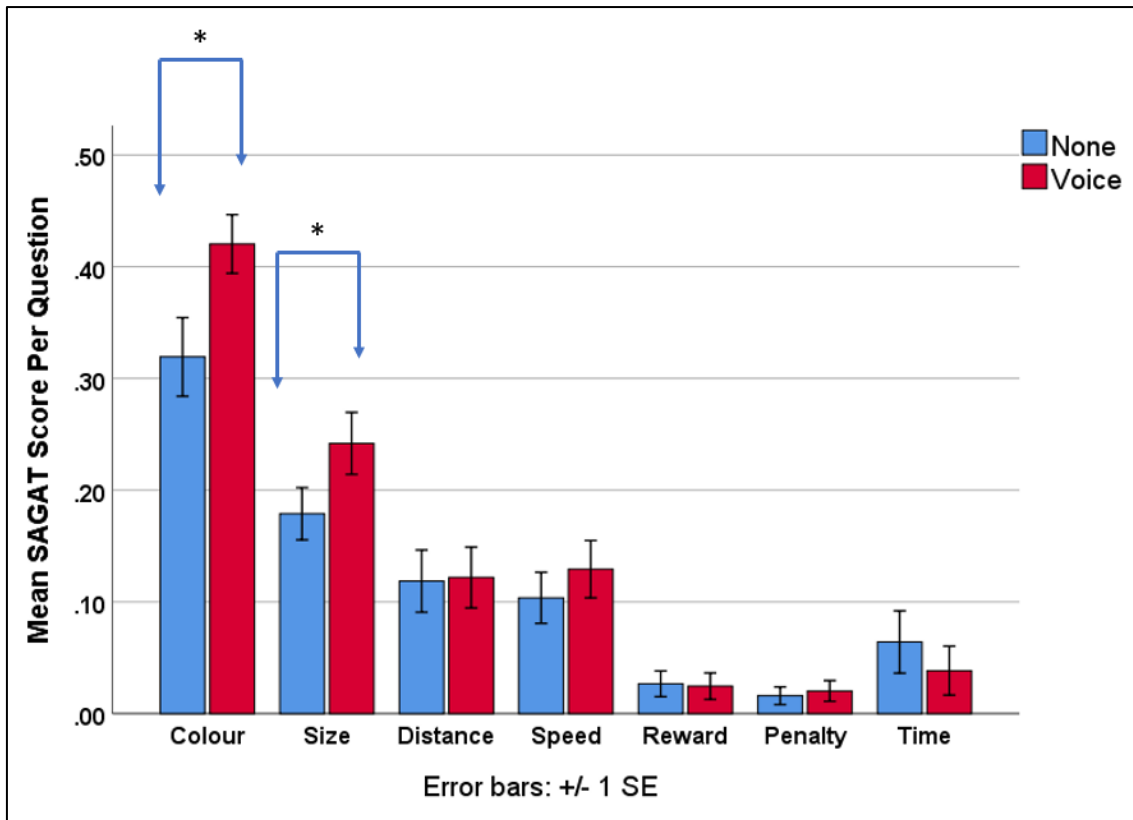
Figure 5.3: Average SA Recall Scores for each of the six SA Recall Questions. Statistically significant effects are indicated by asterisks*

Thus, the requirement to articulate the Colour and Size did seem to improve the participants' attention and their ability to build SA; however, the expansion of the SA was very much limited to just what was being said.

The answers to the post-study qualitative question "What information did you focus on and why?" give context to the importance of this significant variance. Most participants answered that they focused their attention on the relative position and movement of the targets (18 participants or 70%) and/or the colour and shape of the targets (17 participants or 65%) and/or the number of the target (10 participants or 38%). Few participants identified that they paid attention to the reward or penalty (4 participants or 15%). Thus, participants were not only aware that they would be asked colour and size questions, the requirement to articulate colour and size appears to affected their focus of attention in both conditions (although no participant indicated that they varied their SA between conditions). However, the SAGAT results show that

153

despite focusing on Colour and Size in both conditions, only when the Colour and Size were said aloud did the ability to recall that information change; the information apparently became explicit.

5.4.3.2 Recency of Targets Articulated

A further detailed analysis was conducted to determine if the explicit SA recall was affected by the recency of the participant's interaction to see if there was any variation of recall of SA information as the participant's attention moved between targets. SA results for targets appearing in SAGAT interrupts were grouped in rank order by how recently the participant had instructed the synthetic to process the target (either using the keyboard in Quiet or via speech in Speaking). The target most recently assigned for processing was ranked 1st, the next most recently assigned 2nd etc, though to those that were present on the screen at the time of interrupt but had not been assigned for processing which were ranked as "Not".

When comparing recency of interaction within conditions, there was significant variance for the Speaking condition (χ^2(5, 24)=15.240, p=.009), with the average score of the targets most recently interacted with much higher (M=.376 SD=.286) than the 2nd (M=182 SD=.248), 3rd (M=.149 SD=.193), 4th (M=.145 SD=.179) and 5th (M=.096 SD=.224). There was no significant variance between the average SA scores for recency of target interaction for the Quiet condition (χ^2(5, 15)=8.457, p=.133), with the average scores marginally higher for the most recent (M=.223 SD=.292) and second most recent (M=.164 SD=.226) but reducing to a low at 3rd (M=.091 SD=.147), then levelling off for the 4th (M=.117 SD=.166) and 5th (M=.122 SD=.160) most recent. The average recall of targets "Not" interacted with was M=.144 SD=.227.

When comparing between conditions (Figure 5.5), a significant variance was found between the average SA scores for the 1st (Z=-3.062, p=.002) of medium effect size (N=86, r=0.330) with Speaking (M=.376 SD=.286) and Quiet (M=.223 SD=.292), and for the 3rd (Z=-2.038, p=.042) with small to medium effect size (N=69, r=0.245) with Speaking (M=.149 SD=.193) and Quiet (M=.091

SD=.147). No other significant variance was found for other interacted targets, nor for the "Not"

target; however, the overall trend was for consistently higher scores (better recall) for targets

interacted with in the Speaking condition over the Quiet condition (see Figure 5.4) showing that

the addition of the Speaking provided not only a significant improvement in recall of targets

most recently interacted with, but generally improved recall of data about all targets spoken

about. The only exception to this was for the targets not selected for processing (not spoken of

in Speaking or keyed in Quiet), where recall of information about the "Not" targets was actually

lower in the Speaking condition (11%) than the Quiet condition (14%).
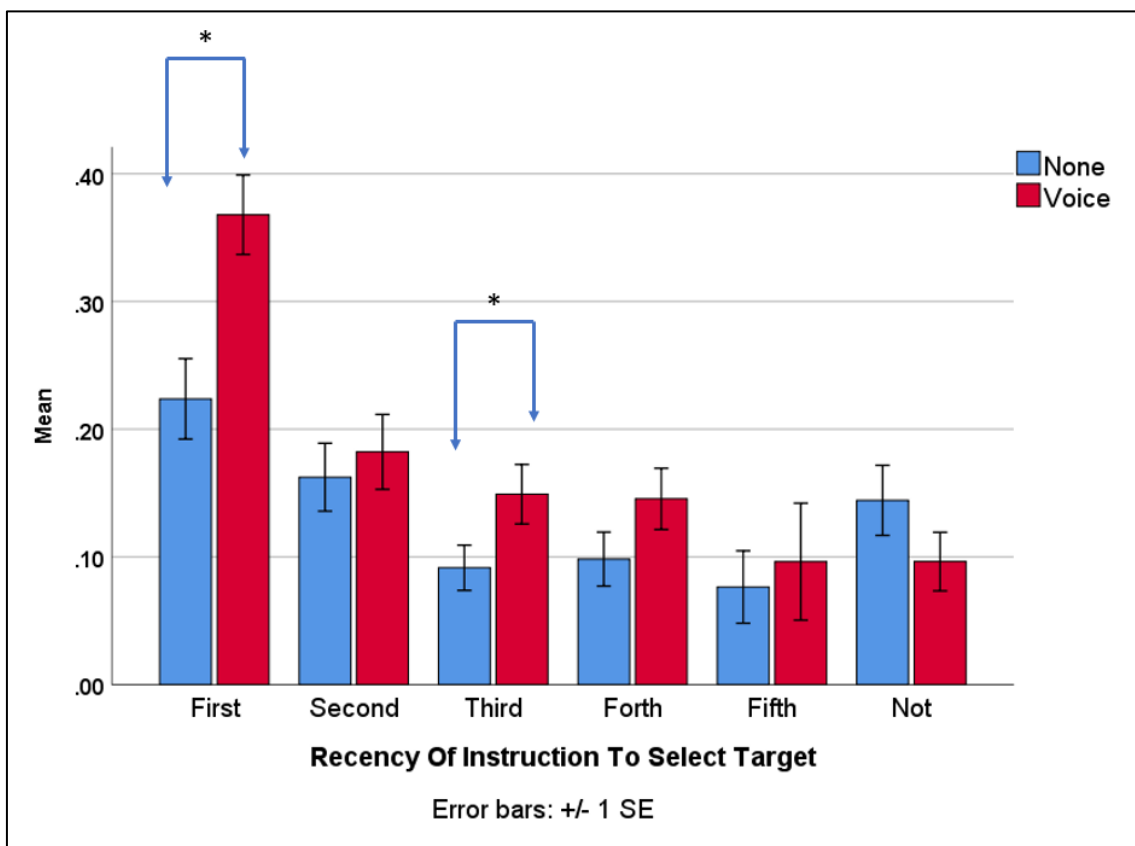
Figure 5.4: Average SA Recall Scores for Targets Based Upon Recency of Selection. Statistically significant effects are indicated by asterisks*

Thus, overall the hypothesis was supported in that there was a significant increase in recall and

thus SA of the data spoken, but there was no significant improvemet in data not spoken, and

there may in fact have been some evidence that recall of some the data not spoken actually decreased (eg Reward and Time and Targets not yet sent for processing).

5.4.4 Hypothesis 3: Speaking Improves Task Performance

The impact of Speaking had widespread effect amongst the three Task Completion and six Selection Behaviour measures, with a total of seven of the nine measures showing significant variation between conditions indicating not only a significant change in performance but also a significant change in behaviour.

Significant variance was found between conditions for all three performance measures (Figure 5.5): Targets Cleared (Z=-4.666, $p < .0005$) with a medium effect size (N=286, r=.2759); Reward Score (Z=-5.264, $p<.0005$) with a medium effect size (N=286, r=.3113) and Processing Rate (Z=-4.430, $p<.0005$) with a small to medium effect size (N=286, r=.2620). The profile plot for the Task Completion measures (Figure 5.6) show that these correspond to an overall drop in performance in the Speaking condition; when participants verbally communicated their processing rate decreased, resulting in less targets cleared per period and thus less reward points gained.
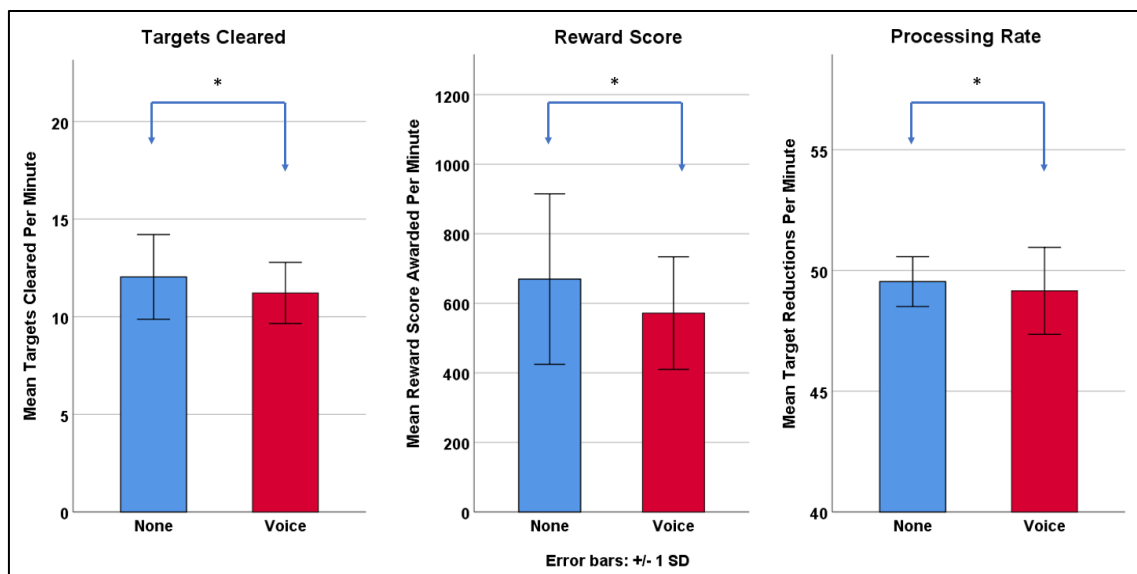
Figure 5.5. Task Completion Performance measures with statistically significant variance. Statistically significant effects are indicated by asterisks*

Four of the six Selection Behaviour measures also showed significant variance. The risk management measure of average Proximity to Deadline showed a significant variance between conditions (Z=-4.516, p<.0005) with a small to medium effect size (N=286, r=.2670) with participants letting the targets get closer to the deadline before they were cleared, however, there was no significant variance between the Proximity to Targets measure.

Of the target selection behaviours Reward Rank, Size Rank and Distance Rank show significant variance; with average Reward Rank (Z=-3.139, p=.002) a small effect size (N=286, r=.1856) and lower in the Speaking condition, average Size Rank (Z=-2.244, p=.025) a small effect size (N=286, r=.1327) higher in the Speaking condition, and average Distance Rank (Z=-3.679, p<.0005) small effect size (N=286, r=.2175) lower in the Speaking condition. The selection behaviour measure of Penalty Rank did not show a difference.
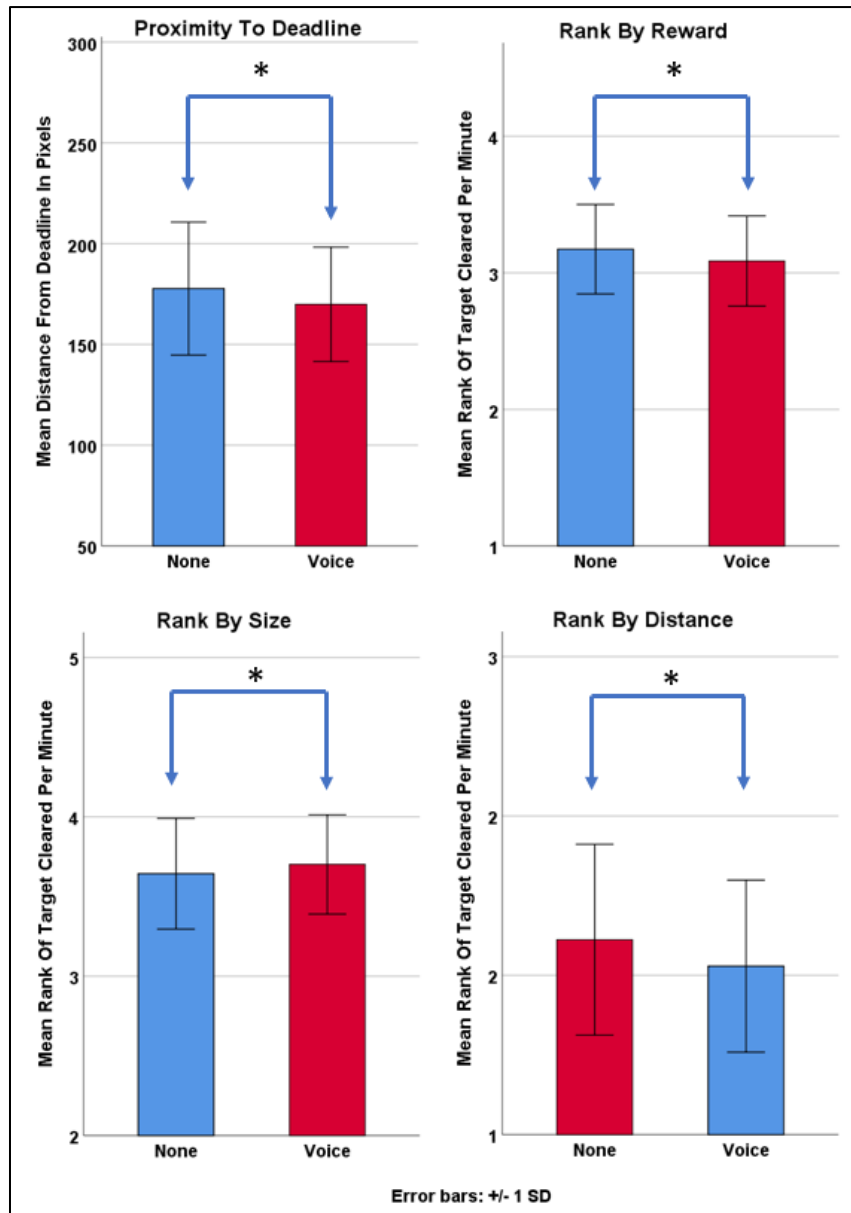
Figure 5.6. Behaviour measures with statistically significant variance. Statistically significant effects are indicated by asterisks*

Overall participants appeared to change their behaviour considerably in the Speaking condition to a more safety orientated risk reduction approach, slowing down the rate of processing and thus reducing targets cleared and reward scored, by prioritising the processing of larger targets (with lower rewards) that are closer to the deadline.

If only taking a task completion perspective of the performance, as determined earlier in the Methodology (3.4.4.1) for this simulator, the results show that Hypothesis 3 is not supported; However, an alternative view is that the change in behaviour towards a more safety orientated behaviour, placing greater emphasis on the quality of target selection rather than the quantity

of target clearance, should actually be taken as a positive (even if it does not support the hypothesis), especially in the aviation context. This is view is discussed below.

5.5   Discussion

The aim of the study was primarily to determine whether the implementation of human operator speech to a synthetic team member would improve the SA of the human operator of the system. The secondary aim of the study was to test whether recognition questions in SAGAT interrupts would demonstrate the presence of implicit SA.

Starting with the second aim, the results from Hypotheses 1 above show that the participants had far greater success at answering recognition questions than recall questions, which in turn does seem to support the hypothesis that much of the SA of the participants is implicit SA. The results of the qualitative interviews also provide further support for the hypotheses, with the majority of participants indicating they were evaluating relative movement and positioning rather than factual information for recall. However, the lack of variance in the scores for recognition questions between conditions indicates that the implementation of speech between the human and synthetic has no impact upon recognition and by extension, likely has no impact on implicit memory. These conclusions are made with some caution, as only two recognition questions were posed per interrupt, and the extremely high recognition scores could be attributed to the fact that humans are significantly better at answering recognition questions than recall (eg Sternberg & Sternberg 2012, Eysenck & Keane 2015).

Despite the message of caution, the deduction of this study based upon the quantitative and qualitative results is that participants likely had a significant proportion of their SA naturally form as implicit SA. The presence of this implicit SA in this study in turn offers an explanation for the low absolute SAGAT recall scores of the Pilot study (Chapter 4.4.4) and other previous research (e.g. Lo et al. 2016); the participants of all studies, using implicit cognition and generating implicit SA, were likely simply unable to declare (recall) the dynamic observations, deductions and

predictions that they were making. That is not to say that participants were unaware of targets being processed, just that they were unable to articulate the data that they were using for processing. Nor is the observation made to criticise SAGAT, rather it is to opine that SAGAT in its recall version is not particularly effective at measuring implicit SA and it is necessary to ensure recognition questions are present if there is an expectation that SA will be partially implicit.

Returning to the primary aim, to determine whether participant speech would improve participant SA, the results for Hypothesis 1 (that SA would improve when the participant spoke) demonstrated that the SA of the human operator did improved significantly in the Speaking condition. As the results of the Pilot study had already shown that viewing messages from the synthetic agent and receiving audio-voice messages from the synthetic agent did not improve SA, then the deduction is that it was the addition of the human operator speaking to the synthetic agent of HAT had likely caused that improvement in SA. As an note of interest, the detailed analysis of results for Hypothesis 2 (Figure 5.3) on specific SA questions indicates that the improvement in ability to recall information was greater for information heard (Size) than for information both spoken and heard (Colour).

Researchers have for some time associated the deliberate pushing of information with improvements in team SA (Demir, McNeese and Cooke 2017), and identifying that providing transactions of SA can be used to build (improve) the SA of other team members (eg Salmon et al. 2010). The results appear to build upon these findings indicating that deliberately articulating SA information in a push communication, in addition to benefitting the SA of the other team member being spoken to, can actually assist the speaker improve their own internal SA. Whilst the additional SA data chosen to be placed in messages in this study (Colour and Size) may seemed potentially trivial, the principle of the results stand; requiring the participant to verbalise key SA information helped them remember and recall that information and helped them use that information when building SA.

Whilst the results of the study were that SA improved, the results of the performance evaluation Hypothesis 3 show that the task completion performance apparently degraded significantly, with the participants in the Speaking condition slowing down their rate of work and clearing fewer targets. This effect appears to be similar to the negative correlation between SA and performance reported by Lo et al. (2016), where, in their study, Train Traffic Controllers with poor processing performance (worse train departure and arrival punctuality) got higher average SAGAT scores.

However, before conclusively determining that our participants' drop in work-rate was a drop in performance it must be observed that there was also significant variance in Selection Behaviour measures, specifically the variance in Proximity to Centre, Reward Rank, Size Rank, and Distance Rank. The variances in these target selection behaviours show that in the Speaking condition participants changed their behaviour to focusing on clearing the targets most at risk of causing a failure of the prime goal "prevent targets reaching the deadline" and prioritising larger targets (that would take longer to clear) nearer to the deadline, which was identified earlier in this study (see 5.4.4 above) and in the Pilot Study (see the discussion in Chapter 4.5) as a safety conscious, risk-adverse behaviour.

Interestingly, Finger Pointing and Calling (FPC) is designed to be a safety improvement practice and is often implemented to deliberately slow down an operator's work-rate and reduce the incidence of errors in task execution. Shinohara et al. (2013:130) explain:

> "One of the important potencies of FPC is to intentionally make a worker control each step of an action sequence. As a result, the behavior of FPC slows down and its accuracy is thus promoted."

In the study the Task Completion performance results certainly demonstrated a reduction in speed, and the Selection Behaviour certainly demonstrated an attempt to reduce the risk of errors. Thus, when considering the complete set of performance results (outcome and selection

behaviour) in the context of the aims and expectations of FPC (slow and safe) an alternative interpretation of the performance results is that the participant speaking actually lead to an improvement in the (safety) performance of the participant; something that is of paramount important for operators of safety critical systems (eg power plants) or personnel employed in safety critical industries (eg transport and medicine). This deduction is supported by the qualitative data with the majority of the participants (18 participants or 69%) explaining that they prioritised achievement of the two "safety goals" they were given (Prevent the targets hitting the centre, and, Prevent the targets hitting each other) over the third "performance" goal (Get the highest score possible).

Finally, it is worth observing that there was a marked difference in performance results between this Shisa Kanko study and the same LOA in the Pilot study (LOA3 – Playbook), which is especially interesting when considering that both studies were conducted using the same task simulator, the same methodology and taking the same measures. In the Pilot Study at LOA3 the audio-voice communication from the synthetic agent appeared to encouraged participants to focus on safety by working faster and clearing all targets further away from the centre Deadline but did not make them prioritise targets on size or colour or value (they just cleared everything faster). In this current study, also at LOA3, the participants' behaviour changed to working slower and prioritising the clearance of larger targets located closer to the central Deadline (ie prioritising on size and risk). Thus, adding the requirement for the human operator to speak to the synthetic agent appears to have had its own effect on participant behaviour; it appears to have made them more selective about which targets they prioritised for clearance.

Finally, worthy of note is that the data from Hypothesis 2 (Figure 5.4) shows that whilst there was a significant improvement in recall of information on the target most recently spoken about in the Speaking condition, there was also a visible trend for reduction in SA for the targets not yet spoken about. This latter decrease could be a symptom of a potential masking effect of the

voice communication, with the improvement in SA of objects articulated covering and reducing the SA on objects not articulated. This potential for masking certainly needs to be investigated through further research.

5.6   Conclusion

The overall conclusion of the study was that the introduction of voice communication between the human and the synthetic agents can improve the human's declarative, and thus accessible Situation Awareness (SA), by encouraging the speaker to transfer implicit information into explicit information, enabling conscious reflection and risk-analysis. However, this improvement in SA is limited to information communicated and could potentially result in information not communicated being masked.

The results from the study also indicated that the improvement in explicit memory was achieved by both listening and talking, with listening appearing to provide a greater improvement in recall SA than talking. That listening is better at creating SA than talking would perhaps not be surprising if the message heard provides new information not accessible elsewhere, but our study shows that even hearing information that is readily available as visual data (and thus is already "known") improves SA. The deduction is that the communication by audio-voice (saying or hearing) of specific data assisted in copying (not transferring) data from implicit SA into explicit SA; improving explicit SA without degrading implicit SA resulting in an overall improvement in SA.

Finally, our results also indicate that not only did the participant talking assist with the transfer of SA knowledge from implicit SA to explicit SA, the performance results show that the addition of participant speech also led to a change in decision-making behaviour. However, and perhaps most pertinently, that change in behaviour was itself new and different to the decision-making behaviour observed in the equivalent LOA in the Pilot study. In the Pilot Study (Chapter 4.4.3.2) at LOA3, the addition of the communication had resulted in participants working faster and

simply attempting to clear all targets more quickly. In this current research at LOA3, when expanding the implementation of audio-voice communication to include participant to synthetic agent commands as well as the original synthetic agent to human communications, the participant behaviour was completely different again. In this current study the participants' work rate actually slowed down and participants became more selective about which target they prioritised (they prioritised larger targets).

The new change in behaviour observed in this current study indicates that SA was not simply expanded in scope without effect on outcome, but that the newly expanded SA was significantly structurally different leading to the participants making new decisions. Thus, the implementation of participant speech appears to have created a cascade change; a change in SA, causing participants to make new (SA L1) perception observations, new (SA L2) comprehensions and new (SA L3) projections, that in turn led to a change in decision-making behaviour. These findings resonate with the aims and practices of Finger Pointing and Calling (FPC) that encourage operators to be more considered, accurate and thus 'safer' in their activities. Whilst the use of participant speech is not a complete replication of FPC, the effect of that speech appears to be extremely similar to that intended of FPC.

These findings on the presence of implicit SA and the use of participant speech on transferring implicit SA into explicit SA to change decision-making behaviour will be of particular relevance to industries running highly automated safety critical systems. It is suggested that in systems such as future aviation Air Traffic Control and Commercial Flight Decks that will fundamentally change away from Human-Human Teams towards a Human Autonomy Teaming (HAT) model of human operators, often solo, working with an autonomous synthetic teammate, it will be highly advantageous to apply the strategies of FPC to the team through routine, active exchange of SA data as audio-voice communication, even if that data is not "new" or is simply to acknowledge understanding that the other team member is undertaking assigned work. To do this will

support not only the creation of team SA; it will also improve the innate SA of the human and will encourage the human into more deliberate, considered, accurate and safety biased behaviour.

Chapter 6 – UTM Study

6.1   Introduction

The originally planned methodology had separated the research into five separate experimental evaluations of the Independent Variables (IVs) of Presence of Voice, Teaming Structure, Operator Speech, Reasoning Transparency and Automation Degradation. The first two studies, the Pilot study and Shisa Kanko study, were to evaluate the effect of three of those IVs, Presence of Voice, Teaming Structure and Operator Speech. This left this Unmanned Aircraft Systems (UAS) Traffic Management (UTM) study to evaluate whether getting the autonomous synthetic agent of a Human Autonomy Team (HAT) to provide spoken messages that included Reasoning Transparency information would improve participant's SA and teaming, or, conversely, would be too long, disruptive and actually negatively impact on participant's SA. It also left the UTM Study to evaluate whether that effect on the participant's SA, performance and teaming would be further effected by the Automation Degrading for a period.

However, the results of the Pilot study had shown that at higher decision-support LOA the audio-voice delivered synthetic agent recommendations had a significant influence over the participant decision-making behaviour. When given a recommendation as an audio-voice message, participants were highly likely to follow it, even if it led to participants taking risks that they would not normally contemplate. There was concern that this change in behaviour might be the result of an unbalanced over-trust in the speaking synthetic agent. Therefore, it was determined to slightly expand the scope of this UTM study to also to evaluate whether any observed increase in trust (resulting from the spoken recommendations) might be over-trust. The observation of behaviour in both the Reliable and Uncertain periods of Automation Degradation would be used to make an assessment on the presence of over-trust as discussed below.

6.1.1 Over-Trust and the addition of Automation Degradation

In the Pilot Study (Chapter 4) when presented with audio-voice Decision Support advice (at LOA5) on how to "increase their score" the participants were highly likely to heed that advice, even when the recommendation was to carry out an activity that appeared risky and increased the likelihood of failing a goal task and getting a penalty. This behaviour was in stark contrast to the participants who, because they were engaged with the synthetic agent at a lower LOA, did not receive decision-support recommendations. Those participants at the lower LOA, when presented with warning advice (effectively a passive recommendation to pay attention) had behaved in a risk adverse manner, taking actions that decreased the likelihood of being awarded a penalty but also led to the award of a lower score.

The post-trial qualitative report and the analysis of the variance in participant trust in the Pilot Study had both indicated that in the presence of the audio-voice communication, participant trust improved. Thus, the observation was that the participants' trust of the autonomous synthetic agent had advanced so much that they were prepared to take advice and take actions that under other circumstances they would not contemplate. This increase in trust had been expected (and hypothesized) as other researchers such as Waytz, Heafner and Epley (2014) had previously identified a link between providing a synthetic agent (an autonomous driving assistant) with a voice and an increase in participant trust. However, it was surprising that the increase in trust persuaded participants with decision support advice to take risks that others without that advice would not.

The immediate interest was whether the provision of the recommendations using audio-voice had caused the participant trust in the synthetic to increase too much, a situation identified in early research into human interaction with automation as over-trust (Parasuraman and Riley 1997). The thought was, if the trust was over-trust it could lead to participants over-using automation advice, abdicating decision-making to the synthetic agent and being prepared (on

advice) to take potentially unnecessary risks. However, as discussed by Hoff and Bashir (2015), the link between trust, risk and uptake of use of automation is not straightforward with some research indicating that even in situations where participants have high levels of trust in automation, the use of the automation can decrease when perceived risk is high (Rajanonah et al. 2008) and other research suggesting the opposite (eg Jones and Stokes 2012). Thus, the literature indicates that the presence of high levels of trust would not necessarily lead to increased use of automation, and perhaps of more relevance, that the acceptance and implementation of risk tolerant advice was not necessarily a consequence of high levels of trust. As Wiegmann, Rich and Zhang (2001) observed, a high uptake or use of automation is more an indicator of reliance on automation than trust in automation. However, the uncertainty of the presence of over-trust remained, and it was determined that this study would attempt to evaluate whether the improvement in trust brought about by the presence of the audio-voice communication was in fact over-trust.

Ironically, as a mitigation for over-trust, Endsley (2018) recommends that the automation should be transparent, which Chen et al. (2016) had previously shown to positively effect and create a balanced or calibrated trust, that calibrated trust included knowing when to reject synthetic agent recommendations. When combining this advice with our earlier observation on how the presence of audio-voice communication can improve trust, the question is raised, "would providing an audio-voice communication with additional transparency advice lead to a calibration of trust that was matched by an increase in Situation Awareness (SA) and performance?"

The expectation was that the answer to the question would be to compare the behaviour of the participants in the Reliable and Uncertain conditions of Automation Degradation not only to each other, but also to the "ideal" standard of only following synthetic agent recommendations. If participant's over-trust, then irrespective of whether the automation was reliable or making

errors, the participants would largely abdicate decision-making to the synthetic agent, would limit themselves to just following the recommendations provided by the synthetic agent and would have reduced performance and SA.

6.2 Study Aim and Hypotheses

6.2.1 Aim

The original aim of the study was to determine if providing the HAT with an audio-voice communication capability that provided reasoning and explanations of deductions and recommendations would improve the overall SA of the participant and the teaming relationship between the participant and synthetic agent. The general hypothesis was that if the presence of the additional reasoning information led to improved operator SA and trust in the synthetic agent, the operator would enter into a closer teaming relationship and the subsequent HAT would be more aligned and efficient in their decision-making. This in turn would lead to more effective and efficient performance and task completion. In summary the general hypotheses was:

> *Would providing additional reasoning detail to synthetic audio-voice messages improve the decision-making, task performance, SA, and teaming of the human operator of a HAT.*

Following the observations from the Pilot study and the of the introduction above, an additional general hypothesis was added:

> *Would any trust developed from the implementation of the audio-voice communication and the inclusion of additional transparency information in the messages be calibrated or would it tend to develop as over-trust.*

6.2.2 Hypotheses

From the original overarching hypotheses four testing hypotheses were proposed against which to evaluate the effect of the Reasoning Transparency on human participant behaviour in the periods where automation is Reliable:

> *Hypothesis 1 – The decision-making of human operators in a HAT will be more efficient and more closely aligned with synthetic agent recommendations when the synthetic teammate audio-voice communications provides detail of reasoning over when the synthetic teammate provides simple vocal messages or is silent.*

> *Hypothesis 2 – Human operators in a Human Autonomy Team will demonstrate more effective task completion and more efficient task processing when the synthetic teammate audio-voice communications provides detail of reasoning over when the synthetic teammate provides simple vocal messages or is silent.*

> *Hypothesis 3 – In a Human Autonomy Team, the human operators will demonstrate improved SA when the synthetic teammate communications provides detail of reasoning over when the synthetic teammate provides simple vocal messages or is silent.*

> *Hypothesis 4 – Human operators in a Human Autonomy Team will demonstrate improved Teaming and Trust in the synthetic teammate when synthetic audio-voice communications provides detail of reasoning over when the synthetic teammate provides simple vocal messages or is silent.*

In addition, to specifically test whether trust is calibrated or over-trust, four hypotheses that are variants of the Hypothesis were used to test variance in participant behaviour for each Reasoning Transparency condition when the synthetic agent has suffered a form of degradation and is Uncertain about the reliability of its recommendations.

Hypothesis 5 – Human operators in a Human Autonomy Team will be more independent of synthetic teammate recommendations when the Synthetic Agent indicates uncertainty in its cognitive reasoning in a period of announced Automation Degradation over conditions where the synthetic teammate provides simple vocal messages or is silent.

Hypothesis 6 – In periods of Automation Degradation, the performance of the human operators in a Human Autonomy Team will degrade less when the Synthetic Agent indicates uncertainty in its cognitive reasoning compared to when the synthetic teammate provides simple vocal messages or is silent.

Hypothesis 7 – In periods of Automation Degradation, the human operators in a Human Autonomy Team will increase the number of Situation Assessments when the Synthetic Agent indicates uncertainty in its cognitive reasoning, compared to when the synthetic teammate provides simple vocal messages or is silent.

Hypothesis 8 – In periods of Automation Degradation, the subjective trust of the human operators in a Human Autonomy Team will degrade less when the Synthetic Agent indicates uncertainty in its cognitive reasoning compared to when the synthetic teammate provides simple vocal messages or is silent.

The expectation was that during a period of Automation Degradation the human with calibrated trust would be better able to identify erroneous recommendations, would likely make some errors (as predicted by Endsley and Kaber 1999) but would be able to minimise the negative performance effect of those errors, would have more capacity to work harder to get a better SA and would not suffer from the catastrophic failure in trust predicted by Weigmann et al. (2010).

6.3 Methodology

6.3.1 Experimental Apparatus

The experimental apparatus used for this study was an implementation of an UTM Simulator in which the participant and an autonomous synthetic agent carried out a Counter Unmanned Aircraft System (C-UAS) task of detecting, managing and controlling drones that appear on the screen. In the UTM simulation, participants working in a HAT with a synthetic team member were presented with a number of Unmanned Aviation Vehicles (UAVs) overflying the west midlands area of the United Kingdom and had to monitor aircraft for illegal flight behaviour. If the HAT identified a UAV behaving or about to behave illegally (an unregistered drone or a registered drone without an approved operations plan attempting to overfly any controlled airspace) they could take three actions: either attempt to Contact it, task a security services drone to Intercept it, or warn all security services and other users about it by marking it on the screen with a GeoFence. Participants could take one, two or all three actions as they determined fit for the situation. Further details of the experimental apparatus can be found in the Methodology (Chapter 3.4.2.2).

6.3.2 UTM Simulator Scenario Design

As this study was primarily to evaluate participant reaction to audio-voice messages and any included recommendations, it was essential to ensure that all trial scenarios had similar and comparable sets of messages and recommendations from the autonomous synthetic agent of the HAT. The synthetic agent's interactions and speech messages were generated dynamically by the UTM Simulator in reaction to the flight profiles of "illegally behaving" UAVs, the presence of the flight profile triggering the synthetic agent to pass through a risk evaluation decision tree that generated a recommendation for action. Therefore, to ensure that each scenario was comparable it was essential that they included similar numbers of illegally behaving UAVs that caused the generation of a similar and comparable collection of synthetic audio-voice messages.

Randomly generating a series of UAV illegal behaviour flight profiles would not guarantee that each trial scenario would generate similar numbers of illegally behaving UAVs and would therefore not guarantee that participants would have similar and comparable experiences between trial scenarios. To overcome this and ensure equality and comparability of participant experience between scenarios, three scripted scenarios were developed for use in each condition, those scripts being derived from a standard scenario template of premeditated encounters. To ensure that the participants remained unaware that the scenarios were based upon a standard template, the sequence of incidents was arranged to ensure each scenario had a unique flow of incidents and the exact start and end locations of each UAV flight path were moved so as to avoid visible similarity (although the start and end points of each flight path were engineering so that the UAV was airborne for the same period and generate the same risk analysis).

The standard template was designed to provide an incident from each branch of the synthetic agent risk evaluation decision tree, thus ensuring that no two incidents of a scenario would follow the same progress path, nor have the same decision-making logic (and thus eliminating opportunity to "learn" responses). Incidents were delivered at a rate of slightly more than one incident per minute, with the result that there were 11 incidents during a normal 10 minute trial, and 12 incidents during the trials extended to include a period of automation degradation. An example of an incident can be seen in Figure 6.1. In this incident an Unregister UAV coloured orange is about to enter Restricted Airspace over HMPO Onley Prison.
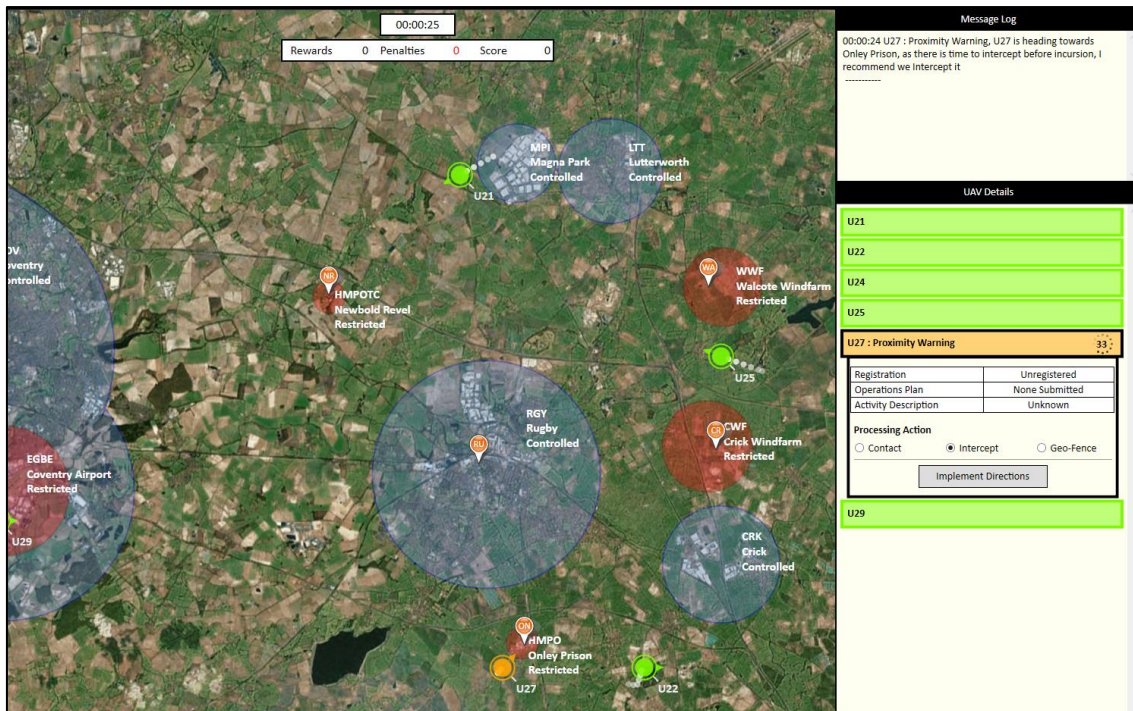
Figure 6.1: UTM Simulator Interface with Incident UAV (U27) and Anthropomorphic Message

The result of this scenario engineering was that each scenario condition provided the participant with an overall similar experience in terms of number of incidents and the number and complexity of synthetic agent communications, whilst at the same time ensuring each trial appeared to the naive participant, subjectively dissimilar and therefore unpredictable.

If the participant was "compliant" and only followed the synthetic recommendations, the 11 unique incidents would each generate a predictable and thus standard number of interactions against which behaviour could be evaluated. Three of the incidents would each generate two synthetic agent interactions (eg a recommendation to contact that fails followed by a recommendation to intercept) and the other eight would each generate only one interaction (eg a recommendation to place a geofence or a recommendation to contact that was successful).

However, participants were not limited to following the recommendations of this "standard profile" of interactions with the synthetic agent, nor in fact obliged to wait for synthetic agent recommendations at all; participants were free to choose when to make an action decision and were free to select any one of the three action options available at all time, before or after a

174

synthetic recommendation. Furthermore, participants could carry out multiple attempts at the same interaction, for example attempting to contact a UAV more than once. They could even attempt to carry out an action on a UAV that the synthetic agent had not identified as presenting a risk.

Participant engagement was encouraged through the implementation of a scoring mechanism that rewarded goal-achievement and penalised goal-failure, the size of those scores determined by the apparent "risk" value of the action. A successful contact (taking an average of 25 seconds to complete) was worth 60 points, a successful intercept (taking between 5-30 seconds to complete) was worth 50 points, and a successful geofence (instant) was worth 40 points. This biasing was established to encouraged participants to considered undertake the more "risky" actions that could be recommended by the synthetic agent rather than simply always placing a geofence around all targets. Failure to "prevent" a UAV conducting an illegal action would result in a penalty, as would carrying out a destructive action on a legal UAV (eg intercepting a UAV that was registered and conforming to its submitted operations plan).

6.3.3 Research Conditions

The study was conducted as a 3 x 2 combination of Repeated-Measures study with two independent variables: Reasoning Transparency and Automation Degradation. The Reasoning Transparency was tested using a Within-Groups evaluation across three conditions. A trial was implemented for each of the Reasoning Transparency conditions:

- **Silent**: a baseline condition in which the participant received information, warnings and recommendations from the synthetic agent as graphical messages and images.

- **Agentic**: an audio-voice communication condition in which the participant received all acknowledgements, information, warnings and recommendations as succinct factual communications that provided no reasoning information on synthetic agent decision-making.

- **Anthropomorphic**: an audio-voice communication condition in which the synthetic agent messages, where applicable, included a short explanation or reason for the cognition that led to a deduction or recommendation.

The independent variable of Presence of Voice was by default directly correlated to the Reasoning Transparency and was thus tested by the Reasoning Transparency conditions.

The Automation Degradation was tested as a Between-Groups factor. Two conditions were set for Automation Degradation:

- **Reliable**: a baseline condition in which the synthetic agent performs without error or malfunction. In this condition all recommendations were the "ideal" derived from the standard synthetic agent decision tree. This condition was implemented for the full 10-minute duration of the first two trials (irrespective of Reasoning Transparency condition) and for the first 7½ minutes of the final trial.

- **Uncertainty**: a degradation condition in which the synthetic agent suffers a sensor failure and warns the operator that decisions contain a level of uncertainty and could be erroneous. In this condition 50% of recommendations were deliberately incorrect and if followed would lead to the award of Penalty points. This condition was implemented for the last 3 minutes of the final trial and during that period participants would experience 4 incidents (2 in which the synthetic agent gave good advice and 2 in which the agent gave faulty advice). Immediately prior to the period of automation degradation the synthetic agent gave the following message:

  o **Silent**: A warning box was displayed over the top centre of the screen with the following text message "WARNING - Drone Speed Calculation Failure. I will continue to provide decision recommendations but I suggest you review each carefully".

176

- **Agentic**: The same message and warning box was displayed, but the message was also said aloud by the synthetic as audio-speech.

- **Anthropomorphic**: The warning box was displayed but the message was increased in length by the addition on an explanation "WARNING - Drone Speed Calculation Failure. The data I receive on drone movement appears to have mistakes - this could result in me making incorrect estimates of drone speed and time to incursion. I will continue to provide decision recommendations but I suggest you review each carefully". As in the Agentic condition this warning was said aloud.

For this study, the independent variables of Teaming Structure was kept constant at LOA5 Decision Support. As many participant PCs did not have voice recognition that was compatible with the C# software used, the Operator Speech was set at Quiet, with the synthetic agent speaking to the human operator and the human operator interacting with the agent through the mouse.

6.3.4 Participants

Thirty-six voluntary participants aged between 20 and 54 (M=35.72, SD=10.625, 14 female and 22 male) completed the study. The experimental study was conducted during the winter 2020 period of Covid-19 lockdown and limited social distancing in the United Kingdom with the result that all participants completed the study without the researcher present. Participants were volunteers responding to social media advertising and were from a range of countries, primarily the United Kingdom, but also Italy, New Zealand and the United Arab Emirates. Twelve participants had experience of the Aviation industry, five as pilots and the remaining seven in ground-based occupations such as Engineering or Human Factors. Participants were not provided with any financial incentive to participate. Participant education levels ranged from High School graduate to Post-Doctorate.

Strenuous efforts were made to maintain the optimal Latin-Square study design by allocating participants equitably to Latin-Square groups. However, it was not possible to rigorously enforce this allocation as many of the original volunteers did not complete the study due to factors outside of our control, mainly due to simple non-completion. As the study was conducted during the lockdowns of 2020/2021 the participants carried out the study at home in a time that was convenient to them, often requiring a reminder to complete them within a month of volunteering. This meant that participants were often allocated a trial sequence to meet the Latin-Square up to a month in advance of taking the trials, and of the total 70+ volunteers who were allocated to a Latin-Square group, only 36 completed the study, with a consequential impact on adherence to the optimal Latin-Square design.

Whilst efforts were made to redress the un-balance in the Latin-Square through careful allocation to groups of the later volunteers (some of whom also did not complete the study), the final result was a slightly un-balanced allocation of participants with just three of the 36 in total being in a different group to the optimal. It is believed that this will have minimal effects on results but was worthy of note.

Participants undertook three trials, one for each implementation of the Reasoning Transparency condition. To implement evaluation of Automation Degradation without providing the participant an opportunity to learn about and anticipate the onset of the degradation state, the Automation Degradation condition of Uncertainty was only implemented in the final trial of each participant experience, and then only for the final period after the third SA measurement interrupt (see Figure 6.2). Because of the use of the Latin-Square sequencing, it was possible to compare the data gathered from the UAV incidents in the Uncertainty period against the same UAV incidents in the Reliable periods of other participants. For example, for Participant 1, Trial 3 was Silent, whereas for Participant 2, Trial 1 was Silent, therefore, to evaluate variance between Reliable and Uncertainty conditions within the Silent condition, the last four incidents

in the Unreliable period of Trial 3 for Participant 1 could be compared to the same four incidents in the Reliable period of Trial 1 for Participant 2. Thus, the use of a Between-Groups analysis of the Reliable and Uncertainty conditions.
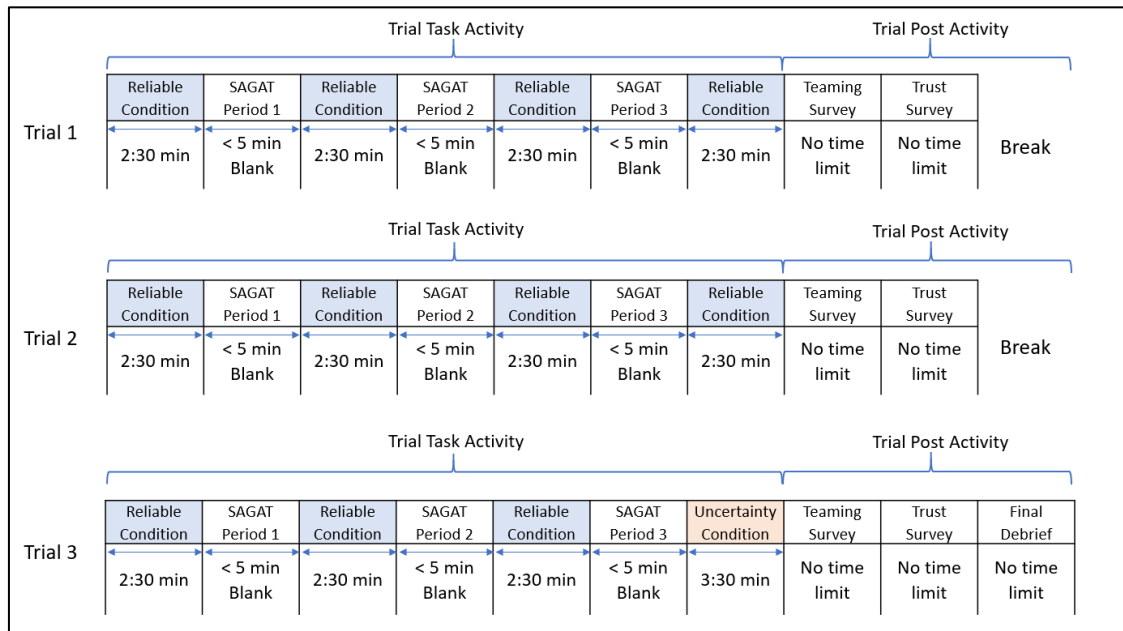
Figure 6.2: Trial Schedule Including SAGAT Freezes

Due to the global 2020 Covid-19 lockdown restrictions, plans to host the UTM Simulator at Coventry University had to be cancelled and instead participants were required to undertake the trials in their home accommodation, using their own IT equipment, setting their own starting time, and maintaining their own environmental conditions. Participants were given guidance on ideal environmental and IT requirements through email communication and an installation document. Participants were advised to use personal headphones rather than speakers to hear the synthetic agent.

Participants downloaded a copy of the UTM Simulator from a safe Coventry University data repository. The UTM Simulator was coded to run on any version of the Windows 10 Operating System. The UTM Simulator software was also engineered to configure and scale its graphical window presentation to the screen of the participant, ensuring that each participant had as similar an experience as was possible to achieve. The UTM Simulator would not run if key

software elements were not present, such as the speech synthesis capability. The researcher made themselves available to provide virtual advice on UTM Simulator installation and to assist with the trial training and practice, although only two participants took this offered assistance.

The UTM Simulator software was built with an embedded five-minute tutorial and 10-minute practice session that all participants were required to take before they could access the trials. Trials were run automatically as per the schedule in Figure 6.2, with questions for the SA measurement interrupts constructed dynamically and presented by the UTM Simulator software, and all trial timing managed by the software. After each trial participants were asked to complete two short five question surveys.

After the final trial, participants were asked to complete a short open-ended questionnaire on their experience that gathered qualitative information of their perception of teaming and SA. The participants were asked one simple question on SA "What information did you want from the screen and automation and why?" and two perception of teaming questions "Which of the three conditions (Silent, Agentic or Anthro) did you prefer and why?", and "How did the failure of the automation in the final trial effect your trust of the automation and why?".

Finally, the UTM Simulator software would generate an encrypted .csv file containing all the data samples for each study which it would save on the participant's home computer. Once the participant had completed all three trials, they would email back the encrypted data files and ethics participant consent forms to the researcher for processing.

6.4   Results

6.4.1   Data Collection

Data collection for the dependent variables of this study was different to the Pilot Study (Chapter 4) and Shisa Kanko Study (Chapter 5) as the primary hypothesis was to evaluate decision-making behaviour and the UTM Simulator scenarios had been engineered to provide all participants

with ample time to manage each incident independently. The following section details the data to be collected to measure variance in the dependent variables and test the study hypotheses.

6.4.1.1 Decision-Making

The directions that participants gave the synthetic agent were used as the primary source of evidence of decision-making. From the hypothesis it was expected that if the operator were more efficient, they would make fewer decisions overall, and if they were aligned with the synthetic agent, they would tend to accept more recommendations, and make more decisions after receiving a recommendation. The expectation Measures used were:

- Total Directions – the average number of action directions given per UAV;

- Follow Recommendations – the number of participant directions that complied with (Followed) a synthetic agent recommendation;

- Independent Directions – the number of directions made either instead of a synthetic agent recommendation or in addition to a recommendation;

- Before Recommendation – the number of directions made before a synthetic agent warning and recommendation; and,

- After Recommendation – the number of directions made after a synthetic agent warning.

For these measures, data was only collected on interactions for incident UAVs as the synthetic agent was only designed to provide warnings and action recommendations for these UAVs (and thus the standard profile only included the incident UAVs).

The measures were also used to test for calibrated trust in the uncertainty periods of automation degradation with an expectation that if the participants over-trusted then they would not change their decision-making behaviour and would tend to accept all recommendations even if

faulty. Conversely, if they had more calibrated trust then during automation degradation uncertainty periods they would behave more independently.

6.4.1.2 Performance

As identified in the Methodology (Chapter 3.4.4.1), the UTM Simulator scoring mechanism provided a direct evaluation of the success of the completion of task goals (Task Completion) and provided three of the four primary measures of performance outcome. In the Pilot Study (Chapter 4) and Shisa Kanko Study (Chapter 5) work rate had been used as a Task Completion measure; however, as the scenarios for this study were scripted and designed so that participant interaction could not affect scenario incident flow rate, rate of work or processing could not be used to evaluate performance. As an alternative, the relationship between time to impact with the edge of the control zone and time taken for a participant to react to a warning was used as a measure of work rate urgency. Thus, the four Task Completion measures were:

- Total Score (cumulative reward and penalty);

- Reward Score;

- Penalty Score;

- Reaction Time;

As decision Selection Behaviour data was being used to test Hypothesis 1, the alternative measures of Processing Efficacy were used to provide additional observations on performance quality. The Processing Efficacy measures taken were:

- Change Of Mind – the frequency that participants changed the action to take for an incident;

- Errors – all forms of mistakes and errors, including asking for impossible actions;

- Contact Interruptions – the frequency of interrupting or overriding a direction to the synthetic agent to Contact a UAV before the synthetic agent has received a reply from the UAV.

- Repeated Directions – the frequency of making additional nugatory instructions to carry out an action that the synthetic agent is already conducting; and,

Data was collected for all UAVs for all performance measures as an action could be carried out on any UAV.

6.4.1.3 Situation Awareness

As discussed in the Literature Review (Chapter 2.2.1.2.3) the robust debate on whether Endsley's SA model identifies product or process ended with Endsley concluding that SA requires both. Observing this conclusion, it follows that measures of both process and product are required to make an accurate estimation of the quality of an individual's SA.

The method for measuring the product of SA employed in this study was the Situation Awareness Global Assessment Technique (SAGAT). Three SAGAT interrupts were conducted per trial at approximately 2:30 minutes, 5:00 minutes and 7:30 minutes of the trials, with a simple correlation expected between SAGAT score and explicit SA. The SAGAT interrupts were timed to run for a maximum of 5 minutes. Six SAGAT questions were asked; two for each of the three levels of SA of the Endsley (1995a) model. SAGAT questions were developed and piloted in collaboration with a Counter-UAV expert (see Appendix A). The questions asked were:

- Location – recognition of the location of each active UAV on the map, see Figure 6.3, participants had to match the coloured UAV image to a grey UAV location (SA Level 1);

- State – recall of the registration and operations plan status of up to four UAVs (SA Level 1);

- Interactions – recall of the number of geofence, intercept and contact directions given (SA Level 2);

- Timing – recall of the time before incursion UAVs were cleared (SA Level 2);

- Destinations – recall of the Zones that up to four UAVs were heading towards (SA Level 3);

- Restricted – recall of the Restricted Zone nearest to up to four UAVs (SA Level 3);

Figure 6.3 SAGAT Recognition question for SA Level 1 Perception.

To measure the processes of producing SA, the frequency with which participants took situation assessments for each UAV was evaluated by assessing:

- Ops Plan Requests – frequency of requests to reveal the drawn operational plans (normally hidden) of UAVs.

- Data Page Requests – frequency of requests to reveal the details page of UAVs. In the Silent condition when a Data Bar was clicked the Data Page was shown which detailed whether the UAV was Registered and had an Operations plan. When the synthetic agent was able to speak it also said this information as well as showed it as text;

Furthermore, as per the conclusion on individual SA provided in the Literature Review (Chapter 2.2.1.6), that measures of decision-making quality if seen collectively across a cohort, provide secondary evidence of quality of SA, the processing performance measures taken for Hypotheses 2 were also used to provide supporting evidence to the primary SAGAT and situation assessment results.

6.4.1.4 Perception of Teaming

Three sets of measures were used to evaluate the participants' subjective perception of teaming across all conditions:

- Teaming – the teaming section of the Collaborative Adaptability Proficiency Test Evaluation Assessment Methodology (CAPTEAM) was used to measure subjective perception of teaming.

- Trust – the Dstl Human Autonomy Teaming Trustworthiness Assessment Protocol was used to measure trust;

- Opinions – qualitative comments collected from the participant debriefs conducted after all trials were used to give context to the survey results.

In addition, just for the two audio-voice conditions, measures of the variance in time duration of synthetic agent speech were taken to identify the participant's tolerance of the speech duration. This information was not used to test the hypothesis but rather to provide insight into the scope of any observed variance in the participant's trust and their tolerance for long message speech.

6.4.1.5 Automation Degradation

To test whether automation degradation caused variance in behaviour the samples of incident responses for the four UAV incidents from the Uncertainty period in each participant's final Trial 3 were compared to incident responses for the same four UAVs in Trial 1 or Trial 2 of other

participant's when the automation would have been Reliable using Between-Groups analysis. The effects of Automation Degradation on decision-making and performance were evaluated using this comparison method. The SA was evaluated through measurement of situation assessments, however, the product of SA as measured by SAGAT was not evaluated as no extra SAGAT interrupts were provided after the Uncertainty period. Perception of Teaming for Automation Degradation was gathered through the use of the qualitative question "How did the failure of the automation in the final trial effect your trust of the automation and why?" rather than from the CAPTEAM and Trust surveys.

6.4.2 Data Sampling

Interaction and performance data were collected continuously over the full duration of each trial. As interaction with the synthetic agent and task scoring was all generated to deal with individual UAVs, the interaction and performance data for Hypotheses 2 − 4 and 5 − 8 was collated by all UAVs and that collated data then used for analysis of variance. However, for Hypothesis 1 and 5, which evaluated participant responses to synthetic agent recommendations, participant decision-making data was collated by response or interaction to each of the synthetic agent recommendations rather than by UAV. SAGAT data was sampled three times per trial as per the schedule in Figure 6.2, and Teaming and Trust data was collected at the end of each trial condition.

6.4.3 Statistical Analysis

All data sets were tested for normality using the Shapiro-Wilk test, which the majority (86%) failed. Therefore, all variance evaluations for all hypotheses were conducted using non-parametric tests. For Hypotheses 1 to 4, within-Groups tests with multiple related samples were tested using the Friedman test and the Wilcoxon signed-rank test was used for pairs of related samples. As recommended by Fritz, Morris and Richler (2012:12), Cohen's r was calculated from Z and N values to provide numerical and relative effect size (Cohen 1992: 157, Table 1) for

Wilcoxon signed-rank tests and for Pearson correlation tests. Between-Groups tests for Hypotheses 5 to 8 were evaluated using the Mann-Whitey U test. For all non-parametric tests, the significance level was set at 5% (.05).

The results are presented in two sections. The first section covers analysis of variances in human participant behaviour and performance purely for the Reliable condition of Automation Degradation. The second section compares decision-making behaviour and performance in the Uncertainty periods of each Reasoning Transparency condition with the Reliable periods.

6.4.4 Section One – Reliable Automation

6.4.4.1 Hypothesis 1 – Participants Will Be Biased Towards Recommendations.

The results of the analysis of variance for each measure show significant variance for three of the five measures: Total Directions (χ^2(2, 455)=9.660, p=.008), Follow Recommendations (χ^2(2, 455)=17.917, p<.0005), and Independent Directions (χ^2(2, 455)=17.833, p<.0005). No significant variance was found for the measure of Before Recommendation (χ^2(2, 455)=1.211, p=.546) and After Recommendation (χ^2(2, 455)=2.818, p=.244).

Total Directions: The pairwise comparisons and profile plots (Figure 6.4, plot 1) show the significant variance was a decrease between the Silent and Anthropomorphic (detailed reasoning message) condition (Z=-2.914, p=.004) found to be a small effect size (N=455, r=.153). The descriptive statistics indicate that in all the conditions participants gave more directions than the recommendations asked for, but that this excess of directions reduced first when the audio-voice was introduced in the Agentic Condition (short synthetic agent message), and then reduced further in the Anthropomorphic Condition when the synthetic communication contained reasoning data. In the Silent condition participants gave on average 141% (M=1.41 SD=.969) of the number of directions that would have been expected if they had limited themselves to only following synthetic agent recommendations (following the "standard

profile"). In the Agentic condition this reduced to 133% (M=1.33 SD=.666) of the standard profile and reduced again in the Anthropomorphic condition to 126% (M=1.26 SD=.640).

Follow Recommendations: The pairwise comparisons for directions followed shows the significant variance to be between the Silent Condition and the Agentic condition (Z=-3.701, p<.0005) and the Silent and Anthropomorphic Condition (Z=-4.014, p<.0005) and the Cohen's r showing this variance to be small for both (Agentic N=455, r=.174, Anthropomorphic N=472, r=.185). The descriptive statistics and profile plots (Figure 6.4 plot 2) show that the variance tended towards an increase in number of recommendations followed (Silent M=.75 SD=.433, Agentic M=.85 SD=.353, Anthropomorphic M=.85 SD=.358).

Independent Directions: The profile plot (Figure 6.4 plot 3) show that participants steadily reduced the number of independent directions from the Silent condition (M=.66 SD=1.062) to the Agentic speech (M=.48 SD=.796) and then again from the Agentic Speech to the Anthropomorphic speech (M=.41 SD=.711). The pairwise comparisons show that there was significant variance between the Silent and Agentic speech condition (Z=-2.924, p=.003) of small effect size (N=455, r=.137) and a significant variance between the Silent and Anthropomorphic conditions (Z=-4.797, p<.0005) of small to medium effect size (N=472, r=.221).

Before Recommendation: Although there was no statistically significant variance, the profile plot (Figure 6.4 plot 4) indicate a small decrease in the number of directions given before a recommendation in the Anthropomorphic condition compared to the other two conditions.

After Recommendation: The profile plot (Figure 6.4 plot 5) show that the number of directions after a recommendation decreased, although there was no statistical significance to this graphical variance.
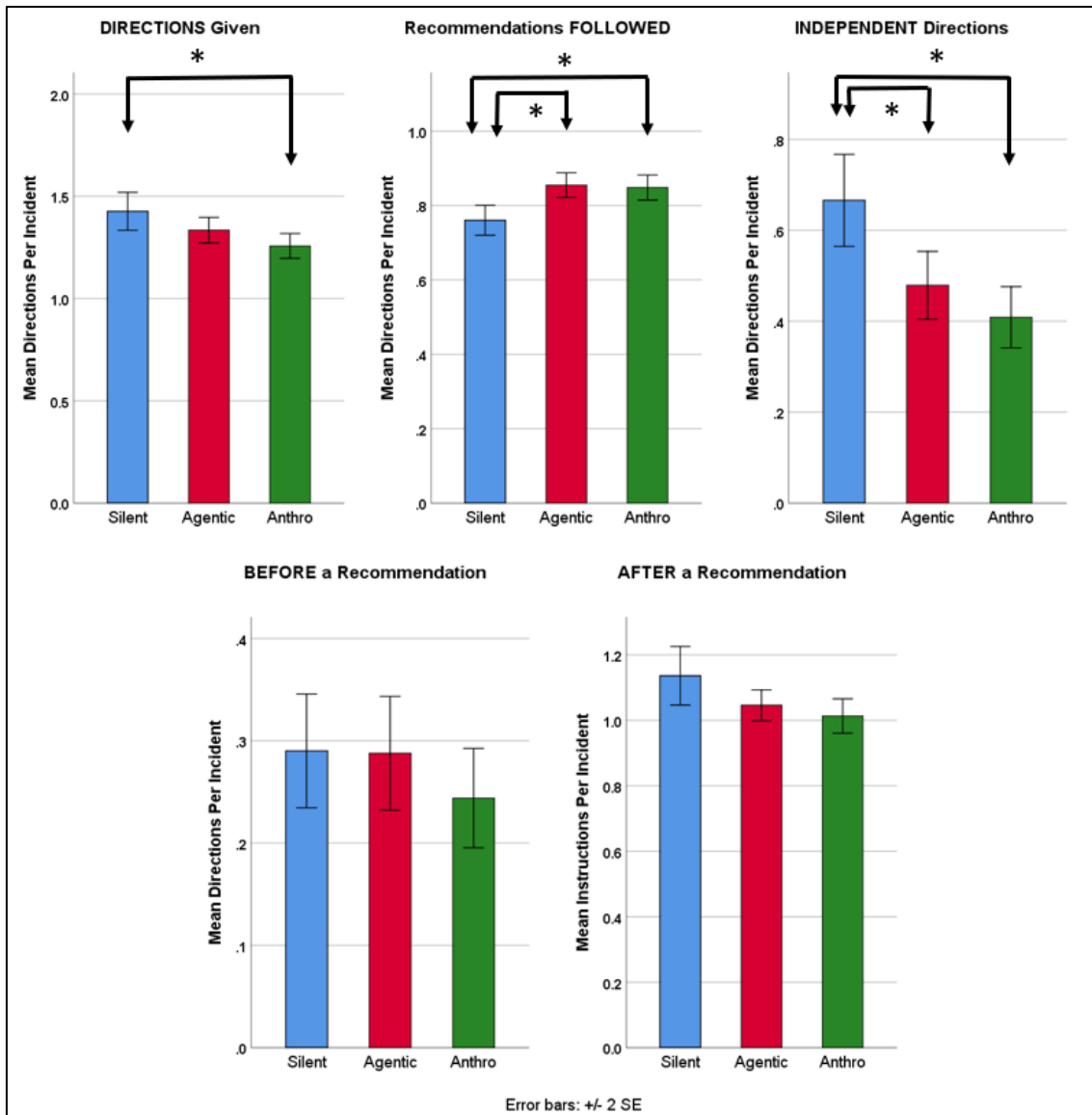
Figure 6.4: Decision-making Performance Between Conditions Statistically significant effects are indicated by asterisks*

The results of the first three measures show that in the anthropomorphic condition the participants reduced the overall number of decisions being made and thus became more efficient. They also more closely aligned their decision-making to the recommendations of the synthetic agent, increasing the number of decisions followed, and decreasing the number made in additional or independent of a recommendation. All these changes support the Hypothesis that the participants would become more efficient and more closely aligned with the synthetic agent recommendations. The fourth and fifth measures showed no statistically significant variance, indicating that participants did not vary when they reacted to an incident, sometime reacting before, but more often waiting until after they had received a recommendation before

determining what action to take. On balance, with three of the five measures showing variance as predicted, the hypothesis is supported.

6.4.4.2 Hypothesis 2 – Participant show improved performance

Evaluating the primary measure of Task Completion performance, there was no significant variance found between average Total Scores for any condition (χ^2(2, 855)=.203, p=.904) or between the Reward Score (χ^2(2, 855)=.145, p=.930). In fact, the profile plots show that the average Total Score (figure 6.5 plot 1) and Reward Score (Figure 6.5 Plot 2) remained very similar between conditions. This is confirmed by the descriptive stats for the average Total Score (Silent M=22.43 SD=31.308, Agentic M=22.89 SD=30.287, and Anthropomorphic M=22.78 SD=31.374) which were only marginally different to the average Total Score that would have been achieved if they had only followed the synthetic recommendation (M=22.48 SD=25.399). Thus, the Total Score and Reward score appeared, in all conditions, to be the "standard" expected from the scenario script.

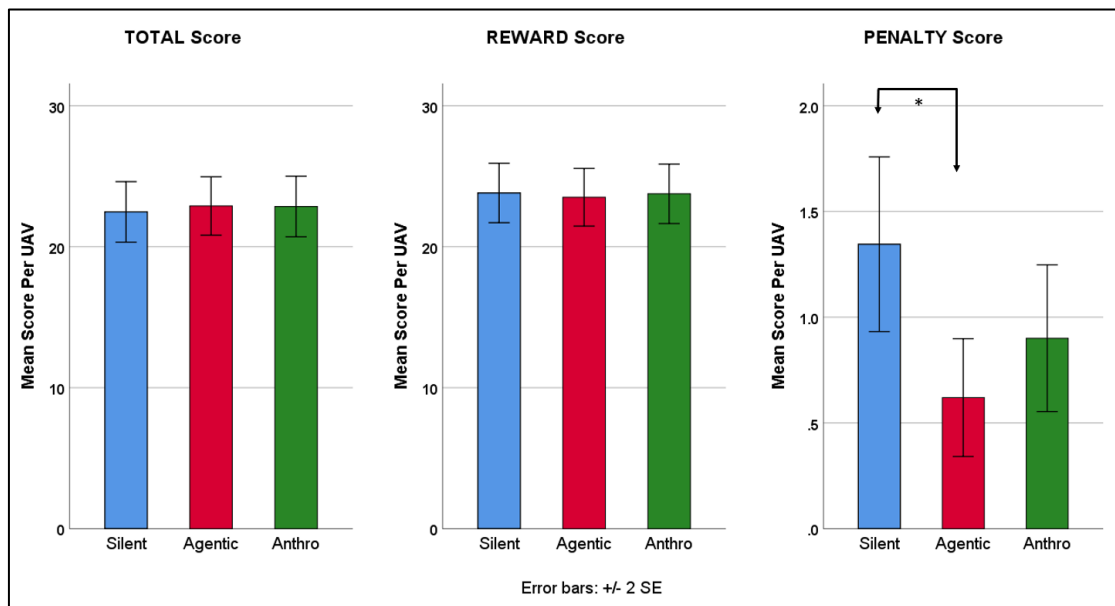Figure 6.5: Task Completion Performance Between Conditions. Statistically significant effects are indicated by asterisks*

Significant variance was found for Penalty Score (χ^2(2, 855)=8.693, p=.013) which the pairwise comparison demonstrates was between the Silent and Agentic conditions (Z=-2.924, p=.003) and the Cohen's r indicates was a small effect size (N=855, r=.009), see Figure 6.5 plot 3.

To evaluate variation in reaction time a Pearson product-moment correlation coefficient was conducted for each condition to determine if there was any relationship between the time the participants took to react (Reaction Time) and the time before the target was calculated to reach the controlled zone (Time To Incursion) which was displayed on the right end of the data bar for each UAV, see Figure 6.6 below.
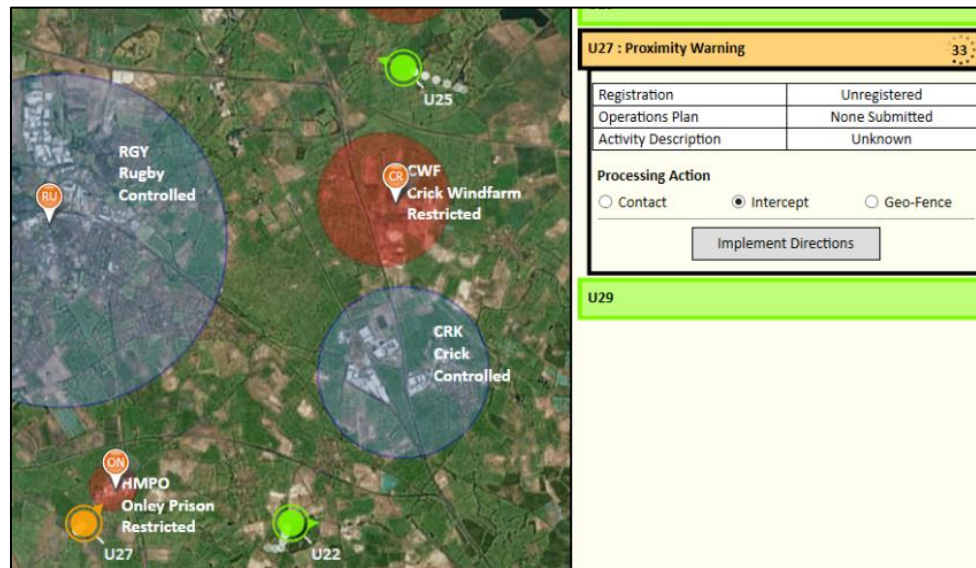
Figure 6.6: Example of Time To Incursion – U27 Data Bar shows 33 Seconds

There was a medium Pearson product-moment correlation that was statistically significant between Reaction Time, and Time To Incursion for each condition with the best correlation occurring in the Agentic condition (Silent r=.303, n=366, p<.0005, Agentic r=.354, n=351, p<.0005, and Anthropomorphic r=.272, n=363, p<.0005). However, when comparing the Pearson's r of each correlation using a Fisher's r-to-z transform for each condition pair (as per advice from Weaver and Wuensch (2013)) there was no significant difference found between the correlations of any condition (Silent to Agentic Z=-.762, p=.223, Silent to Anthropomorphic Z=.454, p=.325, Agentic to Anthropomorphic Z=1.21, p=.113). The scatter plots for the correlations are shown in Figure 6.7 below. Thus, the addition of the audio-voice messages in either format did not appear to affect the correlation between reaction time and time to incursion.

Figure 6.7: Correlation of Effect Time To Incursion has on Participant Incident Reaction Time

Of the four Processing Efficacy measures, three showed significant variances between conditions and one showed no variance. The three Processing Efficacy measures showing significant variance were: Changes of Mind (χ^2(2, 855)=13.657, p=.001), Contact Interruptions (χ^2(2, 855)=15.951, p<.0005), and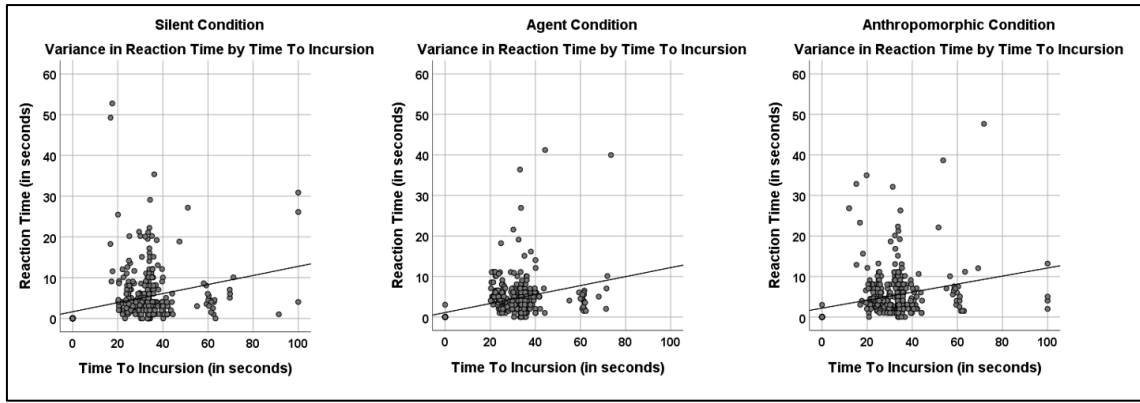 Repeated Directions (χ^2(2, 855)=12.834, p=.002). There was no statistical evidence of variance for Errors between conditions (χ^2(2, 855)=4.075, p=.130); however, worthy of note is that whilst the variance is not statistically significant, the profile plots do show a trend to reduce the number of errors (see Figure 6.8, plot 2) in the Agentic and Anthropomorphic conditions.

Change of Mind: The pairwise comparison and profile plots (Figure 6.8 plot 1) show that the significant variation was a reduction in Changes of Mind between the Silent and Agentic condition (Z=-2.680, p=.007) with a small effect size (N=855, r=.092) and the Silent to Anthropomorphic condition (Z=-2.557, p=.011) is also with a small effect size (N=867, r=.089).

Contact Interruptions: There was significant variation between all three conditions, a reduction between the Silent and Agentic conditions (Z=-2.013, p=.044) of small effect size (N=855, r=.069), a reduction between Silent and Anthropomorphic conditions (Z=-4.206, p<.0005) of small effect size (N=867, r=.143) and reduction between the Agentic and Anthropomorphic conditions (Z=-2.223, p=.026) of small effect size (N=855, r=.076).

Repeated Directions: The variance was a significant reduction between the Silent and Agentic condition (Z=-2.336, p=.019) and Silent and Anthropomorphic (Z=-3.156, p=.002). Again, in both Repeat cases the effect size was small (Silent to Agentic N=855, r=.079, Silent to Anthropomorphic N=867, r=.107). Profile plots for all Processing Efficacy is show in Figure 6.8.

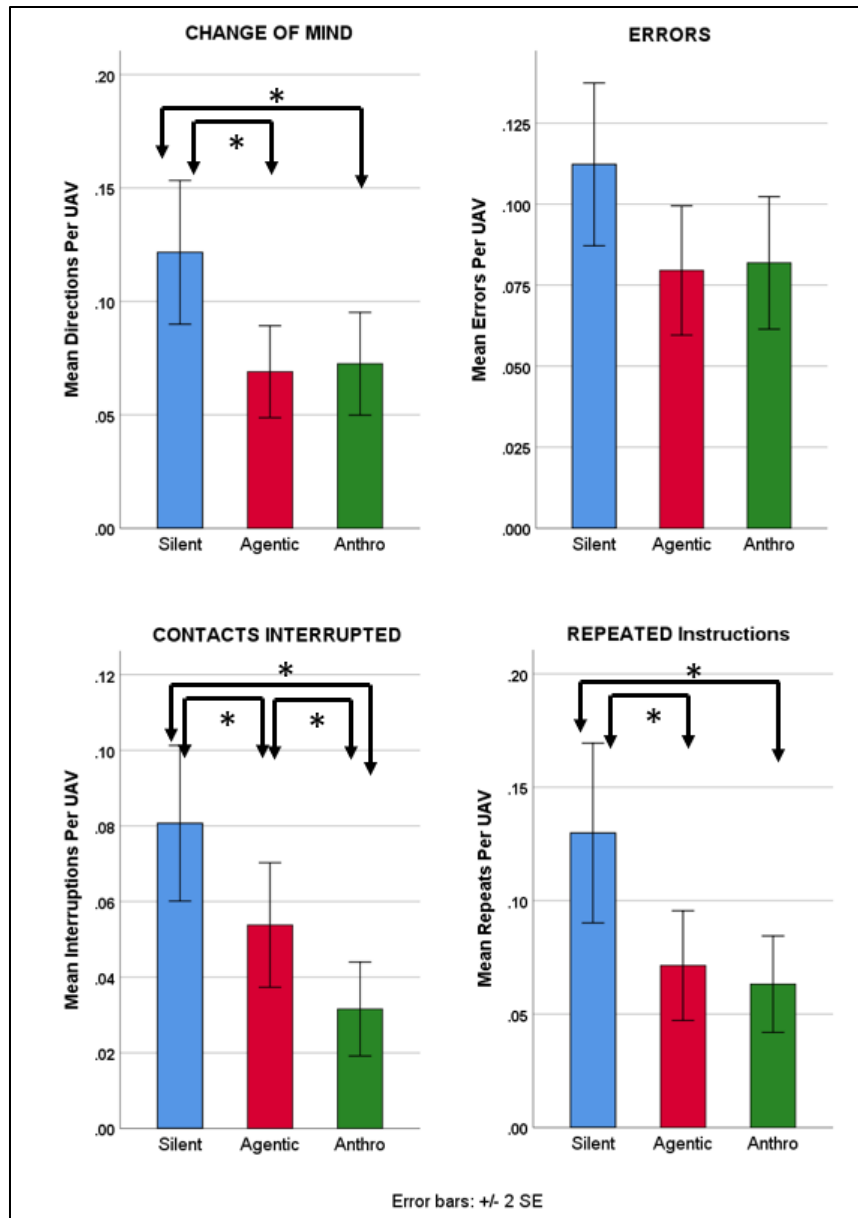Figure 6 8: Processing Efficacy Performance Between Conditions. Statistically significant effects are indicated by asterisks*

Overall, the results indicate that there was no statistical variance between conditions for three of the four Task Completion measures (Total Score, Reward Score, Reaction Time). However, there was a drop in the Penalty Score awarded in both audio-voice conditions although only the

variance between the Silent and Agentic condition was significant. Thus, there was no Task Completion improvement to support the Hypothesis.

In contrast to Task Completion, there was significant improvement in Processing Efficacy for both the Agentic and the Anthropomorphic condition against the Silent condition for three of the four measures, and in one of those measures, Contact Interruptions, significant improvement in the Anthropomorphic over the Agentic, meaning that overall Processing Efficacy improvement was greater in the Anthropomorphic condition than the Agentic condition.

Therefore, when taking Performance in the UTM Simulation as a whole, there is evidence to support the Hypothesis that participant performance would improve in the Anthropomorphic condition compared to the other two conditions. Interestingly however, and warranting of future research, this improvement in observed performance was displayed predominately in Processing Efficiency measures rather than Task Completion measures.

6.4.4.3 Hypothesis 3 – Participants Will Have Improved SA.

6.4.4.3.1 SAGAT Interrupts

Overall, the total scores for the SAGAT interrupts show that participant recollection was within the scope of that predicted by current Working Memory and Short Term Memory theories (eg Baddeley 2010), with participants able to recollect an average of approximately 5 ± 3 items of information per SAGAT interrupt (Silent: M=5.15 SD=3.485, Agentic: M=5.68 SD=3.801, Anthropomorphic: M=5.46 SD=3.858). Thus, we are confident that in this study the SAGAT interrupts were successful at sampling participant recognition and recollection.

The SAGAT results show that there was no significant variance in overall SAGAT scores between conditions (χ^2(2, 108)=1.479, p=.477), or SAGAT scores for any of the three Endsley SA Levels: L1 (χ^2(2, 108)=2.030, p=.354); L2 (χ^2(2, 108)=2.522, p=.283); or, L3 (χ^2(2, 108)=3.344, p=.188), nor was any variance found by individual question: Q1 (χ^2(2, 108)=.395, p=.821), Q2 (χ^2(2)=4.558,

p=.102), Q3 (χ^2(2, 108)=.937, p=.626), Q4 (χ^2(2, 108)=1.483, p=.476), Q5 (χ^2(2, 108)=3.638, p=.162) and Q6 (χ^2(2, 108)=3.799, p=.150).
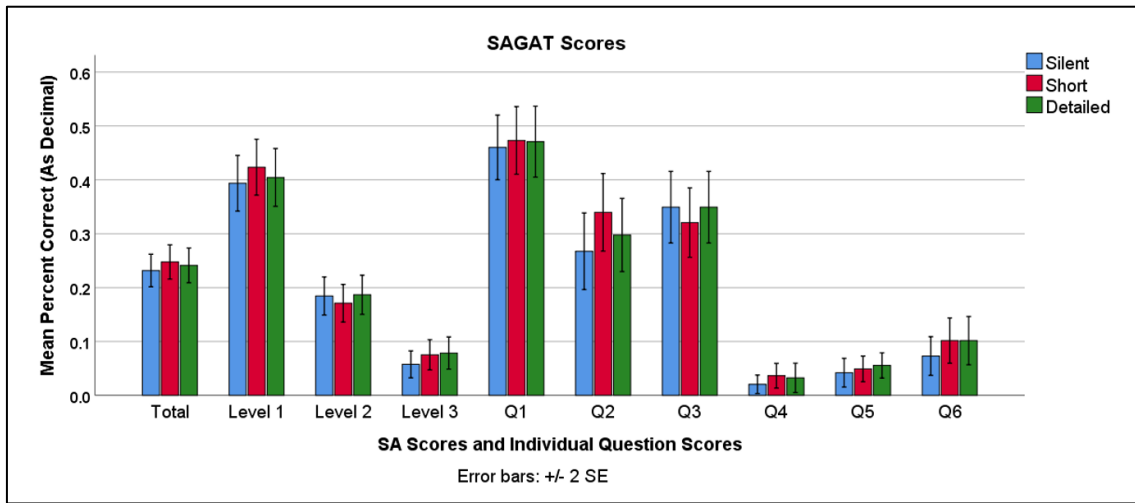
Figure 6.9: SAGAT Scores Between Conditions

6.4.4.3.2   Situation Assessments

Collectively examining data on the total number of comparable team SA interactions there was significant variance in the Ops Plan Requests (χ^2(2, 855)=12.354, p=.002) and Data Page Requests (χ^2(2, 855)=23.327, p<.0005). Examining the two measures with significant variance in more detail, the pairwise comparison shows a significant variance in the Ops Plan Requests (Z=-2.506, p=.012) with a small effect (N=855, r=.086) between the Silent and Agentic condition showing that the average number of requests to view an Ops Plan increased in the Agentic condition. The variance in the Data Page Requests was much more widespread, occurring between all three conditions: Silent to Agentic (Z=-3.201, p=.001) small effect (N=855, r=.109), Silent to Anthropomorphic (Z=-5.395, p<.0005) small effect (N=867, r=.183) and Agentic to Anthropomorphic (Z=-2.271, p=.023) small effect (N=855, r=.077). The profile plot for Data Page Requests (Figure 6.10 plot 2) shows that the number of requests decreased when an audio-voice was added (Agentic condition) and decreased further again when the audio-voice messages contained reasoning data (Anthropomorphic condition).
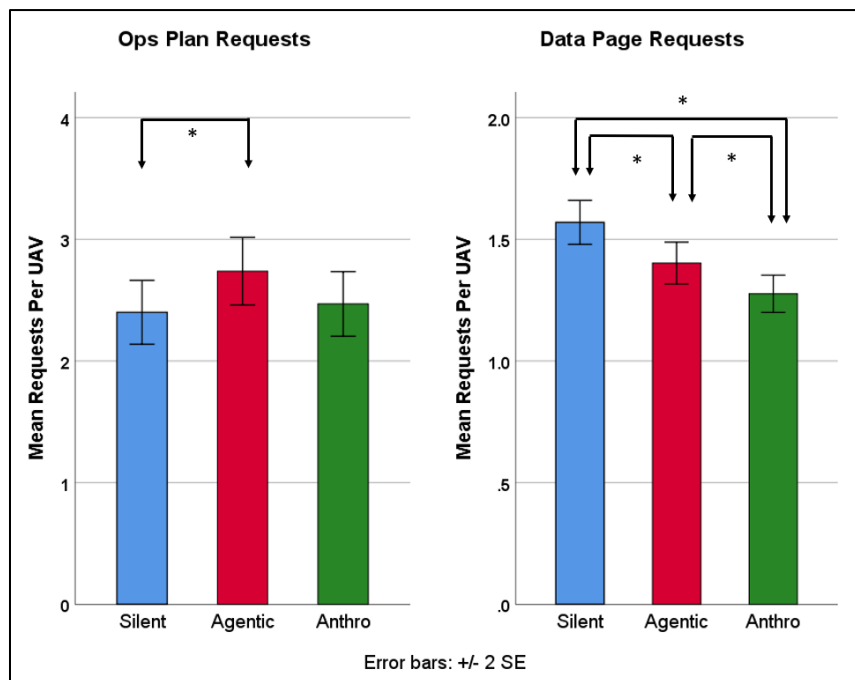
Figure 6.10 Frequency Of Situation Assessment. Statistically significant effects are indicated by asterisks*

If only using the results of the SAGAT which indicates that the product of SA, SA knowledge, did not vary, the hypothesis for an improvement in SA is not supported. However, the reduction in requests for SA information in the Anthropomorphic condition indicates an improvement in SA processing in that less effort was required to collate SA data to build that SA. Furthermore, the secondary processing performance results from Hypothesis 2 that show participants improved their Processing Efficacy would also appear to indicate an improved SA (as predicted by Endsley 1995a, an improved SA being necessary for an improvement in decision-making efficiency and effectiveness).

Answers to the post-study debrief question on SA ("What information did you want from the screen and automation and why?") were generally inconclusive, as many participants interpreted the question as a request for suggestions for UTM Simulator interface improvement with 14 participants indicating what additional information they would have liked to see, and the remaining equally divided amongst the key graphical abstractions of the interface (the UAV Icon colour, the UAV flight path, the UAV Details Page and the Recommendation), meaning each aspect had only 4-8 "votes" which was considered too small a number from which to make valid

estimates of whole of cohort SA attention. Interesting, unlike the previous two studies, in this study participants made positive comments about the audio-voice communications, with eight participants saying they wanted the information provided by the audio-voice messages, either as audio or as a text message.

With the SAGAT results showing no variance in the product of SA, but the measures of SA processing and the secondary evidence of decision-making showing a variance and improvement in SA, the hypothesis is supported.

6.4.4.4 Hypothesis 4 – Participants Will Have Increased Trust And Teaming.

Significant variance was found for the CAPTEAM measure of teaming (χ^2(2, 36)=46.567, p<.0005), with the pairwise analysis and profile plots (Figure 6.11) showing the variance to be between all conditions: an increase between the Silent to Agentic (Z=-4.950, p< .0005) of large effect size (N=36, r=.825), an increase between the Silent to Anthropomorphic (Z=-5.111, p<.005) of large effect size (N=36, r=.852) and an increase from the Agentic to Anthropomorphic (Z=-2.677, p=.007) of medium effect size (N=36, r=.446). However, the descriptive stats appear to show that the perception of Teaming score was quite close between the Agentic (M=.776, SD=.114) and Anthropomorphic (M=.811, SD=.111).

Examining the individual questions, variance was found for four aspects of teaming: Communication (χ^2(2, 36)=39.361, p<.0005), Shared Situation Awareness (χ^2(2, 36)=32.919, p<.0005), Synthetic Teammate Leadership (χ^2(2, 36)=35.548, p<.0005), and Synthetic Teammate Support (χ^2(2, 36)=36.527, p<.0005). Only Team Workload did not vary (χ^2(2, 36)=1.187, p=.552). The significant pairwise comparison for each aspect are:

- Communication: Silent to Agentic (Z=-4.826, p<.0005) and Silent to Anthropomorphic (Z=-4.866, p<.0005).

- Shared SA: Silent to Agentic (*Z*=-4.150, *p*<.0005), Silent to Anthropomorphic (*Z*=-4.910, *p*<.0005) and Agentic to Anthropomorphic (*Z*=-2.559, *p*=.011).

- Team Leadership: Silent to Agentic (*Z*=-4.210, *p*<.0005), Silent to Anthropomorphic (*Z*=-4.707, *p*<.0005) and Agentic to Anthropomorphic (*Z*=-2.331, *p*=.020).

- Team Support: Silent to Agentic (*Z*=-4.748, *p*<.0005) and Silent to Anthropomorphic (*Z*=-4.888, *p*<.0005).
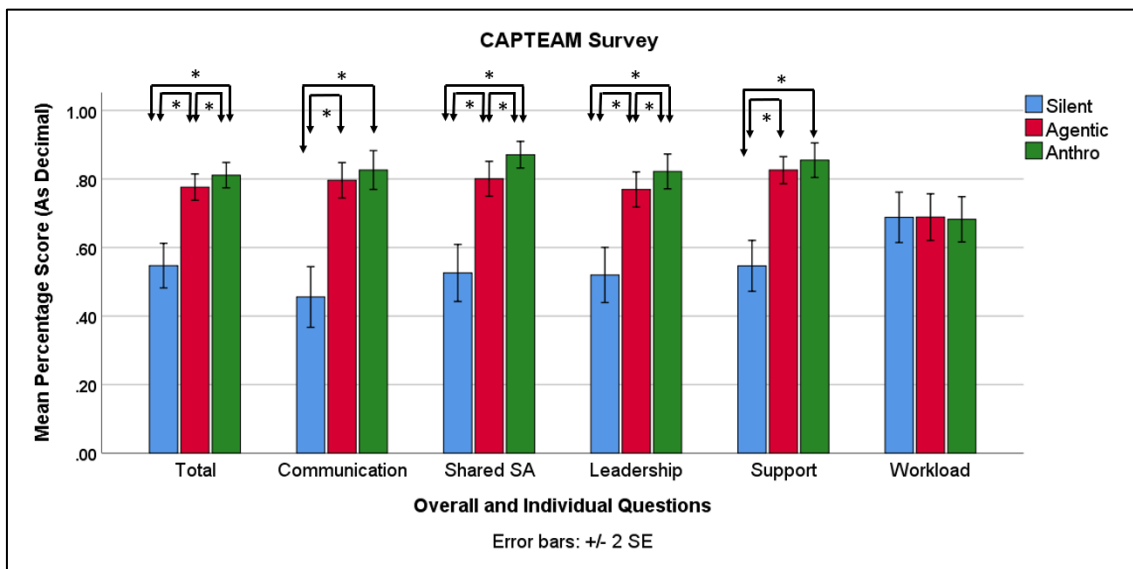
Figure 6.11 CAPTEAM Survey of Perception of Teaming. Statistically significant effects are indicated by asterisks*

There was significant variance found for the overall results of the Human Autonomy Teaming Trustworthiness Assessment Protocol (χ^2(2, 36)=19.504, *p*<.0005) with the pairwise test showing that the variance was between the Silent and Agentic conditions (*Z*=-3.251, *p*=.001) and between the Silent and Anthropomorphic condition (*Z*=-3.859, *p*<.0005). The profile plot (Figure 6.12 plot 1) shows that trust increased from the Silent to Agentic and from Silent to Anthropomorphic condition, with a small but not statistically significant increment seen between the Agentic and Anthropomorphic. The improvement was a large effect size between the Silent and Agentic (N=36, *r*=.542) and even larger effect size between the Silent and Anthropomorphic (N=36, *r*=.643).

When examining the individual questions, significant variance indicating improved trust was found for all five aspects tested: Reliability (χ^2(2, 36)=7.339, p=.025), Dependability (χ^2(2, 36)=12.741, p=.002), Predictability (χ^2(2, 36)=21.020, p<.0005), Availability (χ^2(2, 36)=19.036, p<.0005), and Resilience (χ^2(2, 36)=19.709, p<.0005). The significant pairwise comparison for each aspect are:

- Reliability: Silent to Agentic (Z=-2.121, p=.034) and Silent to Anthropomorphic (Z=-2.794, p=-.005).

- Dependability: Silent to Anthropomorphic (Z=-3.559, p<.0005).

- Predictability: Silent to Agentic (Z=-3.064, p=.002) and Silent to Anthropomorphic (Z=-3.403, p=.001).

- Availability: Silent to Agentic (Z=-3.152, p=.002) and Silent to Anthropomorphic (Z=-3.492, p<.0005).

- Resilience: Silent to Agentic (Z=-3.806, p<.0005) and Silent to Anthropomorphic (Z=-3.368, p=.001).
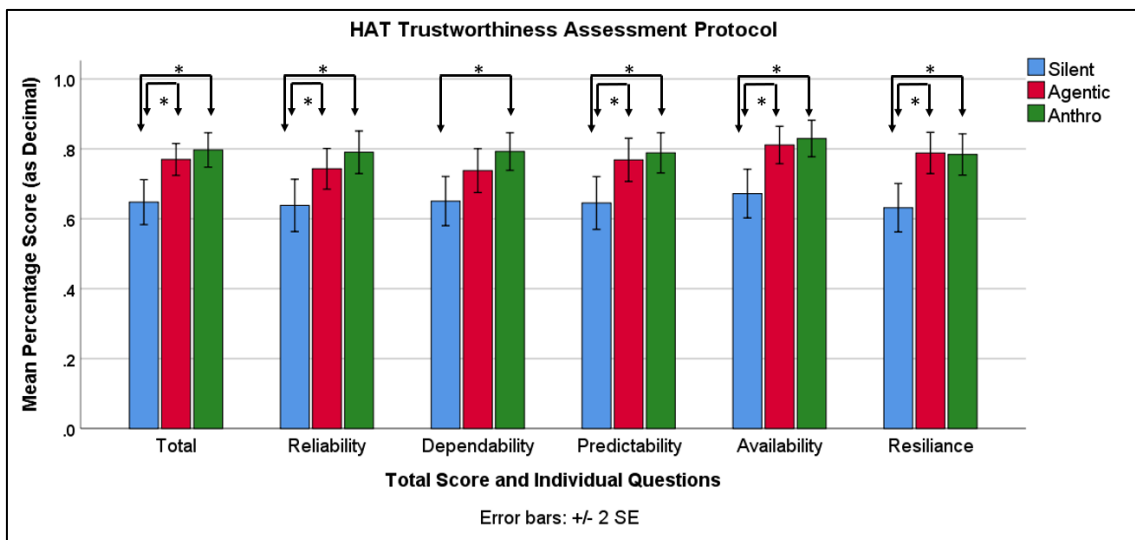

Figure 6.12 Measure of Participant Trust in the Synthetic Agent. Statistically significant effects are indicated by asterisks*

The profile plots for all the teaming and trust measures indicate a consistent rise in perception of teaming between the conditions from Silent, to Agentic, to Anthropomorphic, with the greatest variance between the Silent and Anthropomorphic conditions.

In response to the qualitative question of the final survey "Which of the three conditions did you prefer and why?" 35 of the 36 participants expressed a preference for the audio-voice condition. However, more of those participants expressed as preference for the Agentic condition (17 participants) than the Anthropomorphic (13 participants), and 5 were undecided. Interestingly, the balance of preference between the conditions was closer for participants employed in Aviation (6 preferring Agentic, 6 preferring Anthropomorphic and 1 both equally) as than for participant employed elsewhere (11 preferring Agentic, 7 preferring Anthropomorphic, and 4 both equally), and more of the Pilots preferred the Anthropomorphic (3) over the Agentic (1). This would be worth investigating in further research with a participant audience consisting of more individuals employed in safety critical industries such as Aviation.

Of the three measures, two provide support for the hypothesis (Teaming and Trust surveys) and one does not, with the qualitative results quite closely balanced between the two audio-voice conditions. Therefore, the hypothesis that the addition of the audio-voice capability with reasoning content would improve teaming is supported.

6.4.4.5 Speech Duration Statistics.

To provide context to the participant's tolerance and use of the synthetic audio-voice messages descriptive stats were obtained on the duration of each of the four types of message that were provided by the synthetic agent: Acknowledgement; Warning + Advice (a recommendation); Information + Advice (a recommendation) and Information (UAV Status). The statistics are provided in Table 6.1

Table 6.1 Descriptive Statistics for Speech Duration. All times are in seconds

Message	Agentic			Anthropomorphic		
	Expected Duration	Average Duration	Speech Interrupted	Expected Duration	Average Duration	Speech Interrupted
Acknowledgement	6-7	M=5.657 SD=1.143	16 %	6-7	M=5.737 SD=1.245	15 %
Warn + Advice	6-7	M=4.100 SD=1.967	65 %	10-12	M=5.029 SD=3.271	87 %
Inform + Advice	4-7	M=4.816 SD=1.916	21 %	4-10	M=5 607 SD=2.879	24 %
Information (UAV Status)	4-5	M=3.572 SD=1.313	45 %	4-5	M=3.572 SD=1.519	41 %

The results show that the participants tended to listen for longer to the synthetic speech messages in the Anthropomorphic condition than to the short messages of the Agentic condition. This is perhaps not surprising as in the Anthropomorphic condition the Warning and Information messages could be 4-5 seconds longer than the shorter Agentic messages.

What is perhaps more interesting is how many of the messages were interrupted by the participant before the synthetic agent could finish speaking them. The results of Hypotheses 1 to 4 have demonstrated that the inclusion of the detailed reasoning information in the audio-voice messages changed and improved decision-making, performance, SA and teaming, yet the data on speech duration shows that some participants may not have listen to the complete messages. Further research is needed to understand the effect of these interruptions on the participants results and determine whether there is a correlation between the percentage of the messages heard and the measured improvements in decision-making, performance, SA and teaming. It would be interesting to know if the explanation element needs to be heard to improve SA and performance or whether the mere availability of the explanation can change behaviour and improve those measures.

The other effect of the interruptions is that in both the Agentic and the Anthropomorphic condition the participants did not always wait to hear the recommendation spoken (which would occur in the last two or three seconds of the message). Judging from the average durations it appears that quite frequently participants would had to have determined what the recommendation was from the graphical information on the Data Page. This behaviour seems to indicate that many of the participants in the audio-voice conditions, when building SA for an incident, were simultaneously reading the synthetic recommendation whilst listening to the situation brief and explanation of the message. This provided some support for the proposition that advantage could be gained from cognitive timesharing by splitting and delivering information as part audio-voice and part visual graphics simultaneously as suggested by multiple-resource theory (Wickens 2008).

6.4.5 Section 1 – Reliable Automation: Discussion

The aim of the study was to determine if the implementation of an audio-voice communication capability with voice messages that included additional Reasoning Transparency information would positively affect the decision-making behaviour of the participants, leading to an improved performance, SA and perception of teaming. Four hypotheses derived from this overarching aim were tested. Overall, the results directly supported all four Hypotheses: Hypothesis 1 that participants would be more accepting of recommendations, Hypothesis 2 that participants would have improved performance, Hypothesis 3 that participants would have improved SA, and Hypothesis 4 that participants would have improved teaming. As with the Pilot Study (Chapter 4), these findings are consistent with those of Chen et al. (2016) that an improvement in synthetic agent transparency leads to an improvement in performance and operator subjective trust.

The secondary aim of the study was to evaluate whether any trust developed was calibrated or instead over-trust. The results of Hypothesis 1 show that as the detail of the audio-voice

message increased, the volume of participant decision-making decreased and trended towards the standard profile of recommendations (and decision requests) provided by the synthetic agent. Furthermore, the participants tended to increase their compliance with the synthetic agent recommendations, slightly increasing the number of recommendations followed, and more noticeably reducing the number of decisions made independent of the synthetic agent recommendation. This does seem to indicate that the participants trended towards delegating responsibility for decisions to the synthetic agent with reasoning audio-voice messages. However, even in the Anthropomorphic condition the average number of decisions made leading to interaction with the UTM Simulator (126%) was still greater that the standard profile that would be achieved if only following synthetic recommendations (100%). This indicates that the participants, even though they were becoming increasingly compliant, were still implementing at least one additional action of their own for every three synthetic agent suggestions.

Furthermore, the participants never completely accepted and followed all synthetic recommendations. The mean percentage of recommendations followed was 75% in Silent, and 85% in both Agentic and Anthropomorphic. Even in the Anthropomorphic condition participants rejected at least one in seven recommendations even though all the recommendations were actually viable and "correct". Taking the lead from Chen et al. (2016) who identified the rejection of synthetic agent recommendations as an indicator of balanced trust, the willingness of the UTM participants to reject synthetic agent recommendations indicates that even in the Anthropomorphic condition trust was calibrated. Thus, whilst the participants were definitely affected and strongly influenced by the presence of the voice articulated recommendations (as expected from the Pilot Study results), they did not over-trust and abdicate to the synthetic agent, instead continuing to evaluate synthetic agent recommendations and retaining a level of independent decision-making.

Interestingly, this significant change in behaviour did not appear to lead to a similar change in task completion performance. The provision of Reasoning Transparency, achieved by the synthetic agent giving an explanation for the recommendations offered, appears to have done nothing to improve Task Completion performance, only improving the Processing Efficacy. Whilst this result appears to be at odds with the findings of the Pilot (Chapter 4) and Shisa Kanko (Chapter 5) studies, these results do have similarities with those of Chen et al. (2018) who found in a series of studies that improvements in transparency of a synthetic agent, whilst often leading to an improvement in operator SA did not always lead to an improvement in operator performance; in fact they reported that in one experiment, an increased level of synthetic transparency even lead to a degradation in operator performance.

The results also show that despite the significant change in decision-making behaviour between the three conditions there was no commensurate change in the SA measured using SAGAT. This finding is in contrast to the results of other HAT researchers like Chen et al. (2016) who had observed improved SAGAT scores with improved transparency. The immediate question is whether perhaps the SAGAT performed ineffectively as it did in the Pilot Study. The average number of questions answered correctly per SAGAT interrupt was consistently higher in this study (average 5 ± 3) than the first two studies of this research programme (Shisa Kanko 4 ± 3 and Pilot 2 ± 2) which would indicate that participant recognition and recall was an improvement on the other studies and within the expectations of Working Memory limits. Therefore, the evidence is that the SAGAT was able to successfully gather information from the participants. With the SAGAT apparently able to adequately sample knowledge, but unable to determine how SA was changing to drive changes in decision decision-making (or vice versa) the thought is that the SAGAT might simply have sampled the wrong elements of SA (there is, after all, a limit to the number of questions one can ask in a SAGAT and as Jones and Kaber (2005) observe the success of the SAGAT does depended upon the selection and efficacy of the questions) or as per Lo et

al. (2016) might not have been able to sample SA as the knowledge for the SAGAT questions was implicit.

Whilst there was no SAGAT measured evidence of SA improving, there was a marked and significant improvement in the efficiency of processing of data to generate SA. The reduction in requests for SA information indicates that the participants needed to make significantly fewer situation assessments in the Anthropomorphic condition. As perhaps expected by Demir, McNeese and Cooke (2017), in the two audio-speech conditions (Agentic and Anthropomorphic) when receiving audio-voice communications that effectively pushed information (eg getting a verbal Warning, or having the Data Page information read to them) rather than them having to pull information (by reading the graphical map to see a Warning of a UAV changing colour, or reading the Data Page) the participants did not need to work as hard to build SA. These findings also appear to provide some support for the advantages of cognitive timesharing predicted by Multiple Resource Theory (Wickens 2008). The use of the auditory channel to transfer visually available SA information resulting is less effort to build SA, suggesting that overall cognitive processing was either more efficient, or more effective (or both), with less frequent visual accessing of information needed to create and maintain the SA.

Worthy of note was that the frequency of situation assessments in the Anthropomorphic condition was considerably lower than the Agentic condition. This is interesting as the only difference between the two audio-voice conditions was the presence of the explanation, and that explanation did not provide SA information that was evaluated in the SAGAT questions (an example of a warning explanation was "because we have time I recommend we contact it"). Thus, the presence of the explanation could not have directly contributed towards the building of SA. This apparent and unusual relationship between anthropomorphism, non-SA communication and more efficient situation assessment processing is certainly worth further research.

Irrespective of the reason behind it, the fact remains that the participants did not take as many situation assessments in the Anthropomorphic condition, which shows a reduction in the effort and cognitive workload require to build SA. Theoretically this reduction in cognitive workload should result in the freeing up of cognitive capacity normally occupied with building SA. That freed or spare cognitive capacity could be (again theoretically) used for other cognitive tasks to improve performance, or alternatively could be used to apply and use the SA to improve decision-making (remembering SA is primarily generated to support decision-making according to the Endsley 1995a model). The study results appear to support this theoretical supposition, as they show that both decision-making and performance improved the most in the Anthropomorphic condition when SA processing was at its most efficient. Ironically this could suggest that the observed improvements in decision-making and performance in the Anthropomorphic conditions was not actually a consequence of an improvement in the product of SA, but rather a consequence of the improvement in the processing and use of SA. This would also explain why there was an observed lack of variance for SAGAT measured SA, but an observed variance in decision-making; it seems that SAGAT might have been right and the product of SA did not vary, but the processing and use of it did.

Thus, acknowledging the deduction of the Literature Review (Chapter 2.2.1.2.3) that SA consists of both product and process and can be evaluated by measurements of both, the improvement in SA processing supported the Hypothesis that SA improved the most in the Anthropomorphic condition. Furthermore, it demonstrated that the improvement in SA was highly valuable and a likely primary contributor towards positive decision-making and safety-oriented task processing.

The likely improvement in cognitive spare capacity could be of particular importance for periods of unexpected automation degradation. The accident investigation observations from the AF 447 of 31 May 2009 (BEA 2012) demonstrate that when the automation failed the pilots became highly agitated and had to work hard to find information for their SA, a situation that the Yerkes-

Dodson Law indicates will likely lead to poor information processing, decision-making and performance (Hanoch and Vitouch 2004). The hypothesis generated by the result of this research is that this situation of over-arousal and over-work during unexpected automation failure could be moderated if the operator's workload for SA building was reduced and cognitive spare capacity was increased. If, and when, the participant arousal "spiked" due to the onset of unexpecting conditions, there would be more spare cognitive capacity to respond to the sudden increase in demand for cognitive resources to maintain SA and performance.

Finally, the results from the teaming and trust surveys show that the participants Teaming was greatest in the Anthropomorphic condition, although the qualitative results show slightly more participants appeared to prefer the shorter spoken Agentic messages over the Anthropomorphic conditions.

6.4.6 Section Two – Automation Degradation / Uncertainty: Results

A Between-Groups analysis of variance of between the two periods of Automation Degradation (Reliable and Uncertain) were taken for each of the three Reasoning Transparency conditions (Silent, Agentic and Anthropomorphic). The analysis sought to determine if and how the behaviour of the participant in either the Silent, Agentic or Anthropomorphic condition changed when the automation suffered a degradation. To allow analysis of variance, behaviour data collected on the incidents that occurred during the period of Uncertainty were compared with data collected on the same incidents (but for other participants) during the period of Reliability.

The expectation was that during the period of Uncertainty a participant with calibrated trust would be better able to identify erroneous recommendations and be less likely to accept those recommendations (Hypothesis 5), would likely made some errors (as predicted by Endsley and Kaber 1999) but would be able to minimise the negative performance effect of those errors (Hypothesis 6), would work harder to get a better SA (Hypothesis 7) and would not suffer from the catastrophic failure in trust predicted by Weigmann et al. (2010) (Hypothesis 8).

6.4.6.1 Hypothesis 5 – Participants Will Be More Independent.

The change of reliability from Reliable to Uncertain did not result in any significant between-groups variance for any of the decision-making behaviours measures taken for the Silent condition. However, it did lead to significant variance for most decision-making behaviours for the two audio-voice conditions.

In both the Agentic and Anthropomorphic conditions significant variance was found for Total Directions (Agentic: U=3463.500, *p*=.003; Anthropomorphic: U=3131.000, *p*=.013), Independent Directions (Agentic: U=3120.000, *p*<.0005; Anthropomorphic: U= 2429.500, *p*<.0005) and After Recommendation (Agentic: U=1456.500; *p*<.0005, Anthropomorphic: U=1411.000, *p*<.0005). In addition, significant variance was found for Follow Recommendations between the Reliable and Uncertain periods of the Anthropomorphic condition (U=2790.500, *p*<.005). The profile plots are below in Figure 6.13 and show that when receiving more detailed audio-voice messages participants tended to increase the total number of directions they gave, followed recommendations less, and carried out more additional independent activities generally after the recommendation had been provided. The descriptive statistics show that in the Uncertainty period of the Anthropomorphic condition, on average the participants only followed half of the synthetic recommendations (M=.51, SD=.504), which fits with the number of correct recommendations (2 of the 4 recommendations in the Uncertainty period would be faulty). However, in the Agentic condition, the number of recommendations followed still remained high at 78% (M=.78, SD=.419).

Figure 6.13: Variance in Decision-making Per Reasoning Transparency Due To Automation Degradation. Statistically significant effects are indicated by asterisks*

These results show that the period of Uncertainty only affected decision-making in the two audio-voice conditions, with the greatest variance towards independence between Reliable and Uncertain periods being seen in the Anthropomorphic condition. Therefore, the Hypothesis that the presence of detailed audio-voice messages would lead to greater independence is supported.

6.4.6.2 Hypothesis 6 – Participant Performance Will Not Change.

Significant variance in performance was found for the Penalty Score (Silent: U=1717.000, p=.010; Agentic: U=2022.500, p<.0005; and, Anthropomorphic: U=1924, p=.020) and the Errors measure (Silent: U=1753.500, p=.022; Agentic: U=2064.000, p=.001; and, Anthropomorphic: U=1932.000, p=.024) between the Reliable and Uncertain periods of all three Reasoning Transparency

conditions. The profile plot (Figure 6.14, plot 1 and 2) showing that irrespective of condition, when the automation degraded all participants made more mistakes and got more penalty points.

Further significant reductions in performance were found in the speech conditions for Change of Mind (Agentic: U=2076.000, *p*=.001; Anthropomorphic: U=1903.000, *p*=.039), Contacts Interrupted (Agentic: U=2160, *p*=.024; Anthropomorphic: U=1888.000, *p*=.005), and Repeated Directions (Anthropomorphic: U=1893.500, =.024). With all variance showing that performance reduced almost equally irrespective of whether the synthetic agent spoke, or whether it provided an explanation or not, the hypothesis is not supported.

These results are an interesting contrast to those of Hypothesis 5 above, as they show that although the participants in the Anthropomorphic condition attempted to be more judicious about their decision-selection, reducing the number of recommendations they followed, they were no more able to identify the correct action than the participants in the Silent and Agentic conditions.
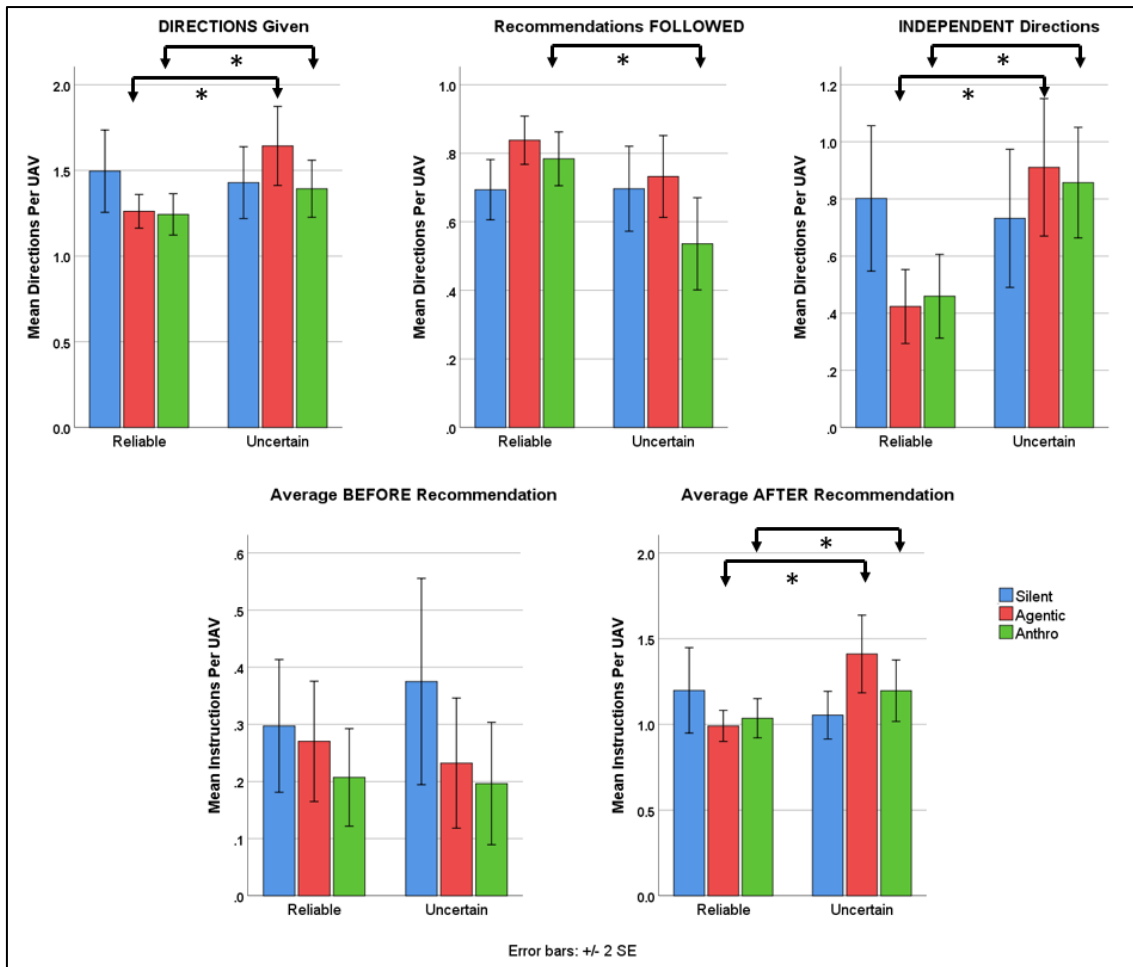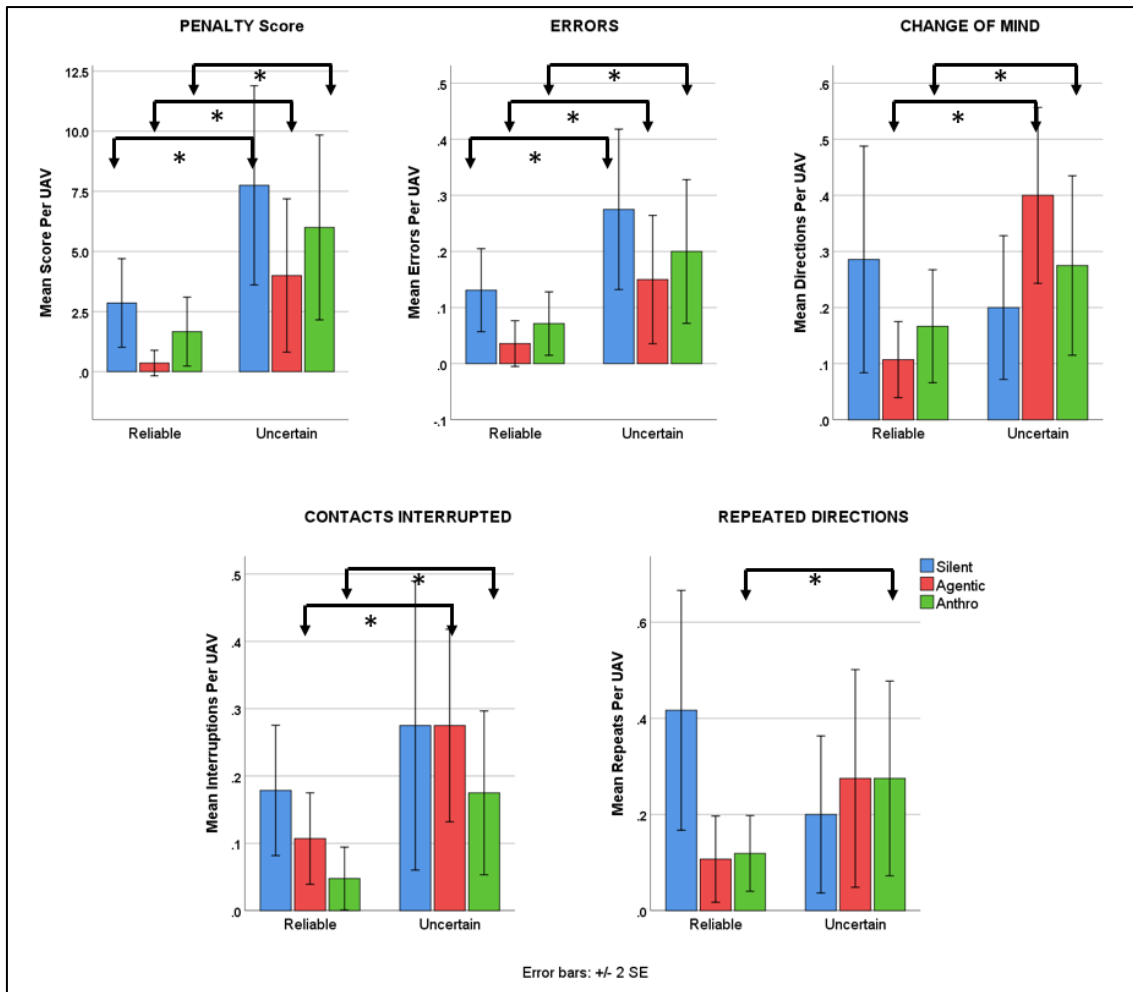
Figure 6.14: Variance in Performance Per Reasoning Transparency Due To Automation Degradation. Statistically significant effects are indicated by asterisks*

6.4.6.3 Hypothesis 7 – Participant SA Will Increase Situation Assessments.

Significant variance in the number of Ops Plan Requests was found for both the Agentic and Anthropomorphic conditions (Agentic: U=1772.000, p=.002; Anthropomorphic: U=1702.000, p=.022). The profile plot (Figure 6.16 plot 1) shows that the number of Ops Plan requests went down in the Uncertain period for the Agentic condition but went up in the Uncertain period for the Anthropomorphic condition. Thus, in the Uncertain period of the Agentic condition the participants decreased the overall number of situation assessments, whereas in the Anthropomorphic they increased the overall number of situation assessments. The profile plot (Figure 6.15 plot 2) for Data Page Requests shows a similar trend for a decrease in situation

assessments in the Agentic condition, although the variance was not significant. These results support the hypothesis.
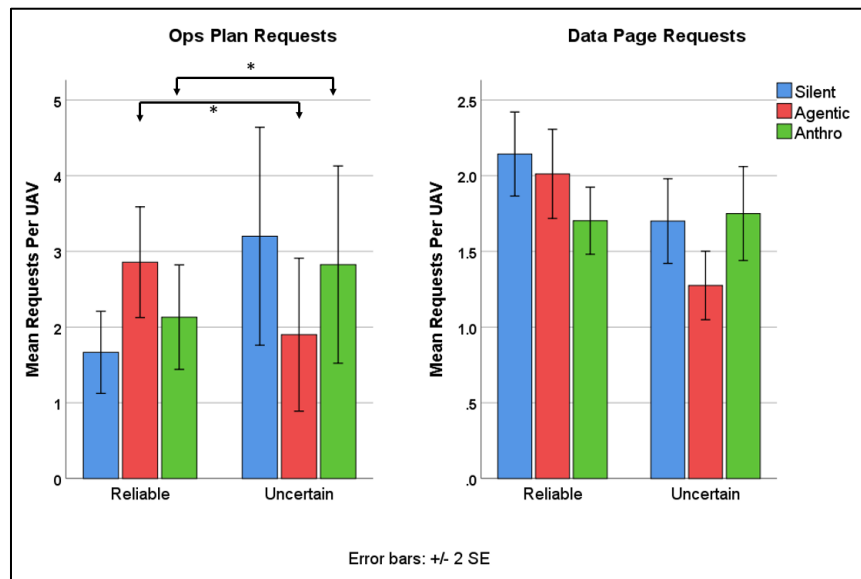
Figure 6.15: Variance in Situation Assessment Per Cognitive Transparency Due To Automation Degradation. Statistically significant effects are indicated by asterisks*

6.4.6.4 Hypothesis 8 – Participant Subjective Trust Will Degrade Less.

The results of the qualitative final survey question "How did the failure of the automation in the final trial effect your trust of the automation and why" were used to determine variance in trust; however, unfortunately not all participants actually answered the question with an indication of an increase or decrease in trust. Seventy percent of the participants who experienced the automation degradation in the Silent responded, with all 70% saying that their trust was decreased. Two thirds of the participants experiencing degradation in the Agentic condition responded, with the majority 47% saying their trust decreased and 20% saying it was unaffected. Eighty two percent of the Anthropomorphic condition responded, with 36% saying that their trust decreased, but the majority 45% saying it remained unaffected. Thus, in the Silent and Agentic condition more participants felt a decrease in trust than not, but in the Anthropomorphic condition this effect was reversed; more felt no change in trust than a decrease in trust. Therefore, with a much smaller relative reduction in trust in the Anthropomorphic condition, the Hypothesis is supported.

Interestingly, some participants answered the question with a statement on perceived workload; with participants explaining that they felt their workload had increased when the automation degraded. In the Silent condition 30% of the participants identified an increase in workload, 18% in the Anthropomorphic and only 13% in the Agentic.

6.4.7 Section Two – Automation Degradation / Uncertainty: Discussion

The primary aim of the second part of the study was to provide specific evidence to determine whether participant trust in the synthetic agent was calibrated or over-trust. Four hypotheses were used to test for calibrated or over-trust, and three of the four were supported (Hypothesis 5 – More Independent; Hypothesis 7 – More Situation Assessments; and Hypothesis 8 – Less Degraded Trust)

The results show that during the automation Uncertainty period, making the synthetic agent provide additional transparency and reasoning for uncertainty (the Anthropomorphic condition) resulted in the participants behaving more independently and rejecting more of the synthetic agent recommendations, thus being more selective of the recommendations they accepted.

However, participant's attempts to be more discerning and make better decisions did not translate into an improvement in performance, with the number of Errors and Penalties awarded significantly increasing in the Uncertainty condition irrespective of whether the synthetic talked or gave reasoned advice. As explained earlier in Section 6.2.2 when preparing the hypotheses, it was expected that participants in all three Reasoning Transparency conditions (Silent, Agentic and Anthropomorphic) would likely make some errors. However, it was expected that the presence of the explanation in the Anthropomorphic condition would assist participants to minimise the negative performance effect of those errors, but this did not happen. It appears that despite being more cautious and rejecting more recommendations in the Anthropomorphic condition, the participants were no better at discerning the correct action to take than participants in the other conditions; they did not necessarily select the "correct"

faulty recommendation for rejection. These results are in contrast to Chen et al.'s (2016) finding that improved transparency lead to better evaluation of errors and improved "correct rejection" of faulty recommendations. In fact, the results seem to indicate that performance was actually worse in the Uncertainty period of the Anthropomorphic condition.

Despite the lack of support for Hypothesis 6 (that performance would not degrade as much in the Uncertainty period if the synthetic provided detailed audio-voice messages), there was support for both Hypothesis 7 and 8. For Hypothesis 7, the number of situation assessments went up for the Anthropomorphic condition, did not vary for the Silent condition, and actually went down for the Agentic condition. This suggests that only in the Anthropomorphic condition (previously associated with the most efficient SA processing and performance in normal conditions) did participants have the spare capacity and awareness to take more SA evaluations in an attempt to identify faulty synthetic recommendations. It also provides support for a more calibrated trust as it shows that participants felt a requirement to gather more information to make their own evaluation of the situation. This finding on variance in rate of situation assessments warrants further research.

Finally, in support of Hypothesis 8, the percentage of participants who felt their trust in the synthetic decreased during periods of Automation Degradation was much smaller for the Anthropomorphic condition compared to the other two conditions (36% of participants said their trust decreased in the Anthropomorphic Uncertainty, verses 47% Agentic, 70% Silent). Furthermore, the ratio of participants who Trusted compared to those who Distrusted was also greater in the Anthropomorphic conditions compared to the other two conditions. Of the participants experiencing automation degradation whilst in the Anthropomorphic condition, 45% Trusted the automation and 36% Distrusted the automation. In the Agentic condition this changed to 20% Trusted vs 47% Distrusted, and, dropped again in the Silent condition to all 70% who answered the questions saying they Distrusted.

Interestingly, these ratios of trust to distrust indicate an unexpected positive outcome of the implementation of the transparency information in the audio-voice messages; in the Silent condition, when the automation degraded participant trust appears to have totally collapsed with all the participants who commented about variance in trust stating that their trust decreased. This aligned with the general expectation identified by Madhavan and Wiegmann (2007) that when automation makes errors operator trust in that automation can very swiftly decline. However, in the Anthropomorphic condition only 36% of participants said their trust decreased and in fact more, 45%, said that their trust did not vary.   Thus, whilst the presence of the audio-voice with additional transparency information does not seem to have prevented some participants losing trust, it seems to have prevented a loss of trust for more of the participants in the Anthropomorphic condition. This seems to accord with the observations made by de Visser et al. (2012) that human trust is more resilient when the failing automation is humanlike over when it is machinelike.

6.4.7.1 Is Over-Trust Present?

The significant reduction in the decision-making measure of Follow Recommendation between the Reliable and Uncertain periods of the Anthropomorphic conditions and the significant increase in the number of Independent Decisions provide direct support that over-trust was certainly not present when the synthetic agent provided explanation detail in the Anthropomorphic condition. The average number of synthetic recommendations accepted in the Uncertainty period of the Anthropomorphic condition (M=.51 SD=.504) does show that participants decreased the number of recommendations accepted in line with the error rate of the synthetic agent (50% of its recommendations in the Uncertain period were faulty and if taken would lead to error and a Penalty). The results for the Agentic condition are not as good. Whilst the participants did decrease the number of recommendations followed during the Uncertainty period of automation degradation from M=.84 SD=.370 to M=.72 SD=.451 the figure indicates that they still tended to follow most recommendations (72%). Thus, the addition of

the audio-voice alone can affect trust, but not sufficiently to significantly change behaviour, whereas the addition of explanation can.

Unfortunately, the performance results, which show an increase in the number of Penalties in all conditions, indicates that although the participants attempted to make better decisions and not take faulty recommendations, they were unable to do so. On reflection, it is perhaps not surprising that the participants were unable to determine which recommendations were faulty or continued to accept many of the synthetic recommendations. The automation degradation was completely unexpected (participants were not briefed that it could happen), and prior to the unexpected automation degradation, which occurred in the final 2½ minutes of the third 10-minute trial, for 27½ minutes (in three trials) the synthetic agent had been completely reliable always giving sound advice that resulted in participants getting Reward points. Furthermore, the majority of participants were naive to aviation and had received no training in how to cope with automation degradation.

Interestingly, whilst the presence of calibrated trust is relatively clearly demonstrated in the Anthropomorphic condition, the decision-making behaviour and trust results for the Silent and Agentic condition are not so clear and in fact seem to provide conflicting evidence. In the Agentic and even more so in the Silent condition, when automation degraded participants declared that they did not trust the synthetic agent and advice, but they still tended to use the recommendation provided. These latter results seem to align with early research into the relationship between automation failure and trust by Wiegmann, Rich and Zhang (2001), who determined that measures such as rate of use of automation are more likely to reflect reliance on automation than trust in automation. They observed users can distrust automation advice but still take it because aided diagnosis is better than unaided diagnosis; the results from the Silent and Agentic condition seem to fit this description.

This then raises the question that whilst perhaps not over-trusting of the automation, were participants over-reliant on the automation? Hypothesising, if participants were over-reliant, or alternatively, over-dependent on synthetic advice then in periods of both Reliability and Uncertainly their behaviour would trend towards not only relying upon the outcome of synthetic agent's decision, but also relying upon the synthetic agent to identify when a decision needed to be made. This is certainly the case for incident UAVs assessed in Hypothesis 5 where, irrespective of Automation Degradation and Transparency Reasoning the vast majority of directions were given after the synthetic agent gave a warning (the range of statistical means for After Recommendation measures was M=.99 SD=.477 to M=1.28 SD=.783, vs Before Recommendation M=.22 SD=.418 to M=.38 SD=.676). However, what was not discussed in the results was the participants actions towards the "distractor" UAVs.

Over the duration of each trial scenario, participants would be presented with 14 "distractor" UAVs that the synthetic would not make a recommendation towards. Despite the lack of recommendations, it was still possible to take an action on these distractors and potentially earn extra Reward Points (normally by attempting to contact registered UAVs with no Operations Plan). However, across the cohort, participants only gave directions for actions on just 21% of all distractor UAVs (20% in Silent, 24% in Agentic and 20% in Anthropomorphic), and even then, approximately 50% of those actions were towards UAVs that they could not get rewards for. Thus, it would appear that, in the absence of recommendations to interact with the distractors, the participants either paid the distractors no attention, or if they did pay them attention, did so rather haphazardly, as likely getting penalties from those interactions as rewards.

Adding these two sets of observations together what emerges is an apparent tendency of the participants to both rely upon the synthetic to identify when to take an action as well as determine what action to take. Thus, overall, the results seem to support the ad hoc hypotheses (derived from Wiegmann, Rich and Zhang's (2001) advice), that the continuing acceptance of

recommendations and apparently inability to identify faulty suggestions in the Uncertainty period despite degraded trust, does indicate an unusually high level of over-reliance. However, as explained earlier, this finding is entirely within expectation of participant behaviour, as in the 27½ minute before the unexpected synthetic failure the synthetic had been both highly reliable and in fact highly competent, always making suggests that lead to a reward.

6.5   Conclusion

The results of the study generally continue to support the core hypotheses of the research programme that adding an audio-voice communication capability to a Human Autonomy Team (HAT) will result in an improvement in human operator SA, performance and perception of teaming.

If only considering the Task Completion results for Hypotheses 2 (improved performance) and the SAGAT results of Hypothesis 3 (improved SA), it would appear that the results of the study lead to the conclusion that there is little to be gained by providing additional reasoning and transparency information in the audio-voice message; it only provides an slight improvement in processing efficacy performance over the short spoken Agentic and it does not lead to an improvement in SAGAT measured SA. However, this conclusion would be ignoring two important observations; the addition of the transparency detail in the voice had a significant effect on the decision-making behaviours of the participants, significantly reduced the workload necessary to build SA upon which to make decisions and significantly improved the perception of teaming. The presence of the additional information appeared to steady participants, encouraging them to make significantly fewer but more effective decisions to gain the same overall Total Score. It appears to have encouraged the participants to choose quality of decision-making over quantity.

Importantly the results also show that whilst the presence of the additional transparency information did lead the participants to heed and follow the synthetic recommendation more

assiduously, the participants did not totally abdicate their decision-making responsibilities. In the Anthropomorphic condition participants continued to make independent decisions, not always following the guidance of the synthetic agent, and not limiting themselves to just the number of actions recommended. This behaviour indicates that in the presence of transparency information the participants' trust was calibrated, not under or over trust. The results obtained during the period of automation failure at the end of the third trial also support this determination, with participants in the Anthropomorphic condition accepting significantly fewer recommendations, increasing the number of independent mitigation actions and reporting some decrease of trust. Interestingly, the results during automation failure indicate that not only was trust calibrated, it was also less fragile, with a considerably smaller proportion of participants reporting a decrease in trust due to automation failure than in the other two conditions.

However, whilst the results show that trust did not degenerate into over-trust, the same could not be said for reliance. The participants general lack of attention towards distractor UAVs in Reliable conditions (no automation degradation) and the habit of waiting for recommendations before taking actions indicates that the participants depended very heavily upon the synthetic agent to provide them with cues as to when to interact. That dependence on the synthetic agent leading the risk analysis seems to have been strong in all conditions, but particularly prevalent in the Anthropomorphic condition when the automation was Reliable, and oddly, in the Silent condition when the automation was Uncertain (suffering automation degradation).

Ironically, this increase in dependence may actually be providing evidence of teaming. As the synthetic agent advanced from Silent to Agentic to Anthropomorphic, the participants seemed to increasingly rely upon the synthetic to take the lead as a "look-out and advisor". They seemed to grant the synthetic agent with increasingly more autonomy, waiting for the synthetic to make a recommendation before selecting an action and more frequently just taking the advice given

and doing no more. This deduction is supported by comments made by the participants in the final survey such as "I much preferred having an audio guiding and advising me" and "It felt as if we were working together and were sharing information and decision-making". It is also supported by the teaming results which show that the participants had a significantly higher regard for the synthetic as a leading teammate in the Anthropomorphic condition (CAPTEAM "Teammate Leadership" score and "Teammate Support" score). The trust results also show that the participants considered it to be more Reliable and Dependable (dstl HAT Trustworthiness Assessment Protocol) when Anthropomorphic. All these results combined appear to show that the participants increasingly viewed the speaking and rationalising synthetic agent as a teammate that they could rely upon to carry out a specific lead role. The apparently high levels of reliance upon the synthetic agent's vigilance, analysis and recommendations may actual indicate (and is taken as) the very manifestation of co-operative teaming aimed for in Human Autonomy Teaming. This possibility is worth investigating through further research.

Perhaps the most notable of the results was the observation that the addition of the transparency detail reduced the cognitive effort (and potentially workload) required to build Situation Awareness (SA). This is particularly important for safety critical systems such as aviation flight control as it follows that the reduction in cognitive processing would likely result in an increase in cognitive spare capacity which can in turn be used for other cognitive functions to improve performance and safety, or for better applying the SA built to conduct decision-making. Or alternatively, it can be kept spare for use in response to unexpected situations that bring about a surge in demand for cognitive processing, as for example appeared to happen in the period of automation degradation in the study. The results show that in the period of automation uncertainty the participants receiving the longer audio-voice messages containing reasoning information were able to surge the frequency with which they took situation assessments, whereas participants in other conditions were not.

Taking the observations made on changed decision-making, calibrated trust and improved SA processing into consideration, a broader more complete conclusion drawn from the results can be made. The conclusion is that implementing an audio-voice communication capability in which the synthetic agent passes reasoning and transparency information, whilst not necessarily improving the task completion performance for normal conditions, does significantly affect processing efficacy performance. It also contributes significantly towards the efficiency of building SA, freeing up cognitive capacity and making decision-making more resolute, efficient, systematic and ultimately safer. This improvement in spare cognitive capacity could theoretically also be of use in abnormal conditions; freeing up cognitive resources to improve operator resilience to unexpected scenarios where both arousal and workload increase. This latter hypothesis offers an opportunity for further research.

This deduction on the value and structure of audio-voice speech is an important finding for safety critical industries where the avoidance of unforced errors and survivability in periods of automation degradation is paramount. These results would seem to indicate that to improve the safety performance of a Human Autonomy Team, the synthetic agent should be provided with an audio-voice communication capability that provides detailed messages with reasoning and explanation information that facilitates the creation of higher more abstracted levels of SA on both the team and the task.

However, before simply determining to add complex transparency detail to all audio-voice messages, it is important to recall that, whilst the greatest improvement in safety performance occurred in the presence of the detailed messages of the Anthropomorphic condition, more participants subjectively preferred the shorter messages of the Agentic condition over the Anthropomorphic condition. Also, the participants consistently interrupted more of the longer Warning (87%) messages of the Anthropomorphic condition, compared to the shorter messages of the Agentic condition (65%), see Section 6.4.4.5 above. From the average duration of the

Warning and Information messages in the Anthropomorphic condition, participants did not always heard the explanation of the Anthropomorphic messages, yet both Warnings and Information messages affected the participant's behaviour. This indicates that either the participants were affected by the presence of the explanation rather than knowledge of that explanation, or, they obtained their knowledge of the explanation from the text message window.

Thus, the ultimate conclusion from the study is that there is definite value in including additional transparency information in audio-voice communication as it is able to significantly improve SA and trust and consequently positively affect decision-making and performance. However, it would be advisable to make the audio delivery of the transparency information optional and under the control of the operator, and to provide a message board with a record of all messages, even those that were interrupted, complete with explanations and recommendations that the operator could read. Or alternatively, the ability of the operator to ask for transparency information could be provided.

Chapter 7 – Discussion

7.1 Review of Research Aim and Hypotheses

The aim of this research programme was to experimentally evaluate the effect of providing an autonomous synthetic agent of a complex (aviation) system with an audio-voice speech and voice recognition capability on the Situation Awareness (SA), performance and perception of teaming of the human operator of the system. The experimental studies were to provide answers to the general research questions and hypotheses:

Hypothesis 1:   In a Human Autonomy Team (HAT), the human operators will demonstrate improved SA when the human and autonomous synthetic agent communicate using a combination of audio-voice and graphics over when the human and autonomous agent communicate using graphics alone.

Hypothesis 2:   Human operators of a HAT will demonstrate improved task performance when the human and autonomous synthetic agent communicate using a combination of audio-voice and graphics over when the human and autonomous agent communicate using graphics alone.

Hypothesis 3:   Using a combination of anthropomorphic audio-voice communication and graphical communication between the human and autonomous synthetic agent will improve the overall teaming and facilitate the creation of a HAT in comparison to when the human and autonomous agent communicate using graphics alone.

The three experimental studies were designed to answer these hypotheses within the specific context of the Independent Variables discussed in the research questions of the Introduction (Chapter 1) and identified in the Methodology (Chapter 3):

- Presence of Audio-Voice Communication.   Would the presence of the audio-voice communication capability change SA, performance and teaming?

- Structure of Team. How would the structure off the team, which would affect the scope of communication, affect the SA, performance and teaming?

- Operator Speech. Would requiring the participant to speak provide any additional effect on SA and performance?

- Reasoning transparency. Would adding message content that would provide evidence of synthetic cognition affect the SA, performance and teaming?

- Automation Degradation. Would the receipt by the human of audio-voice uncertainty information before and during a period of automation degradation prepare them to successfully and safety take over tasks during the period of automation degradation?

7.2   Message vs Medium

One of the earliest concerns about the fundamental theory behind this research project was that when using the audio-voice (speech) medium to deliver a message, it could well be that the content of the message, by providing additional information, would actually be the cause of any observed change in SA, performance or teaming, rather than the combination of message and medium. To address this concern two activities were taken.

Firstly, as a preventative mitigation the experimental simulators and scenarios were designed so that the SA information being provided in the message was also readily available in graphical representation on the screen (eg in the UTM study a change in UAV colour as a visual warning given at exactly the same time as the verbal warning message).

Secondly, an additional "Text" condition was implemented in the Pilot study in which the participants received a copy of the audio-voice message, but had it displayed in a text-messaging window deliberately placed to be conveniently visible to the participants. Through this condition it was possible to test the effect of the message content alone, comparing it to receiving no messages and also receiving messages as audio-voice speech.

The results of the Pilot study showed that the delivery of information readily available in graphical format as a text message had no significant effect on either SA or performance. The message had to

be delivered as an audio-voice message for there to be a significant improvement in performance. Thus, it is the use of the auditory channel that has the determinant effect.

7.3   Performance

The interpretation of the performance results of the three studies requires a quick discussion on the definition of "good performance".   If viewing a positive change in performance from the perspective of simply an increase or decrease in the number of tasks completed, then the results indicated that performance and variance in performance across the three studies was quite inconstant, increasing in the Pilot study (Chapter 4), decreasing in the Shisa Kanko study (Chapter 5) and not varying in the UTM study (Chapter 6).   These finding are similar to those of Chen et al. (2018) who found in a series of studies that improvements in transparency of a synthetic agent, whilst often leading to an improvement in operator SA did not always lead to an improvement in operator performance.

However, focusing on quantity measures of performance such as task success and completion rate could be providing a skewed and even inaccurate assessment, as good performance is not always measured by frequency of task completions.   In fact, the practice of emphasising the quality over quantity of performance is apparently not usual in HAT research, with Chen et al. (2016) determining to evaluate performance in their "IMPACT" study through the measurement of acceptance or rejection of autonomous agent decision recommendations rather than task completion rates.

In the Shisa Kanko study, the consequence of introducing participant audio-voice communication was complex, affecting two aspects of performance but in opposite directions.   Whilst the rate of task completion, effectively the quantity of performance decreased, the quality of decision-making improved with an increase in the number of safe or at least more safety orientated decisions.   While on the surface these two variances in performance might seem to counter each other, the deduction made in the Shisa Kanko discussion was that these two changes in performance were, in hindsight, the change in behaviour the Shisa Kanko practice aims to achieve.   The change to a safety-oriented performance with safer decisions and a slower more deliberate and considered rate of work is the

goal of the Shisa Kanko practice. Thus, the deduction is that, if assessing the performance from the goal perspective of the Shisa Kanko method (eg Iwasaki and Fujinami 2012, Shinohara et al.2013), the introduction of operator audio-voice communication did not lead to a decrease in performance (as measured by just points scored) but rather an increase in performance (as measured by an improvement in safety behaviour and risk mitigating decision-making).

Furthermore, the lack of variance in task performance in the UTM study (Chapter 6) may also be providing an inappropriate assessment of performance in relation to overall HAT performance. In the UTM study, the simulator scenarios were fully scripted with the highly reliable automation providing recommendations for all obvious opportunities to obtain a score. Whilst it was possible for participants to increase their score by another 80%, to do so the participants would have needed to behave in a very independent, uncooperative and, ultimately, unteam-like manner to the synthetic, duplicating its risk-assessment work and over-ruling almost all of its recommendations. This would hardly be the coordinating and cooperating that improves team performance identified by Salas, Cooke and Rosen (2008), and is actually the type of task conflict behaviour that is associated with poor team work engagement that negatively impacts on trust and effective team function (Costa, Passos and Bakker 2015). Thus, ironically, obtaining a higher than normal score in the UTM study would indicate that the participants did not form a Human Autonomy Team (HAT) with the synthetic, and that the HAT did not perform as a cohesive pair, a result that could hardly be considered "good teaming performance".

Examining the goals of the HAT and simulation task it becomes apparent that perhaps viewing performance only as task-completion is inappropriate, as in all three of the studies the goals that the participants were set placed greater emphasis on safety than outcome: in the Pilot study and Shisa Kanko study the first two goals were to avoid collisions with the centre deadline and with other targets; in the UTM study the first two goals were prevent unauthorised UAVs flying over controlled

airspace and report and geofence those UAVs that cannot be prevented. Only the third goal of each study encourage improvement in task outcome; "obtain the maximum score".

Taking a goal centric view of performance in this series of studies places a greater emphasis on the safety behaviour as the primary value and measure of performance and places the measure of task completion in a secondary ranking. This approach is not unusual in safety critical industries like aviation where the safe completion of a task is significantly more important than the speedily completion of that same task. For example, in air traffic management it is more important to ensure aircraft keep a safe distance apart than it is to make sure that they pass through a region as fast as possible; it would hardly be good performance to increase transition speed but also increase the risk of mid-air collision.

Therefore, in this research project an improvement in safety behaviour (and thus performance) was taken as the overall primary indicator of performance and was evaluated using quality measures such as decision-making behaviour, risk management, and reduction in goal failures and errors. Task achievement was kept as a gauge of performance but was set as the secondary indicator, provided by such measures as rate of task achievement and relative task value (scoring) indicated an improvement in performance.

Taking all three experimental studies as a whole, performance hypotheses for each study were supported, with safety-orientated performance improving in two of the four Levels of Automation (LOA) of the Pilot Study and both the Shisa Kanko and UTM study. Thus, the performance results of all three studies combined supported the research project Hypotheses 2 that performance would improve in the presence of audio-voice communication.

Taking a closer look at the results from the individual studies, it can be seen that whilst in all three studies there was a general improvement in the quality of performance and the safety behaviour of the participants, the way that improvement manifest varied between each of the studies. This

indicates that the independent variables being tested in those studies also had a contributing effect on participant performance.

The independent variables with the greatest effect were the Structure of Team (or LOA), which broadly dictated whether there was or was not a recommendation present in the audio-voice messages, and the Reasoning Transparency which determined whether the recommendations included explanation or rationalisation information. When the teaming structure resulted in recommendations complete with explanation and justifications being present (eg the UTM Study and LOA 5 in the Pilot Study), the participants were highly influenced, changing their behaviour to match the synthetic recommendations.

In the UTM study, where synthetic advice was designed to be safe and valuable, the addition of justified recommendations led to the participants becoming more consistent in their behaviour, significantly reducing the overall number of additional or surplus decisions. This can be attributed to a significant improvement in their safety orientated performance, with the participants making fewer penalising mistakes, being more assured in the decision selection (changing their minds less frequently) and less likely to interrupt synthetic agent work or "nag" the synthetic agent (repeat directions).

Conversely, in the Pilot study at LOA 5 where the speaking synthetic agent gave recommendations that encouraged participants to take a potentially unsafe approach to improve their score, the participant behaviour changed to becoming more "risk tolerant". This change in behaviour was unexpected, as in the lower LOA of the Pilot study when the synthetic agent spoke and provided warnings and information but did not provide "risk-taking" recommendations the participant behaviour trended to becoming more "risk adverse", exaggerating behaviours that increased safety margins.

Taken together the results of the two studies indicate that providing synthetic agent recommendations with through the audio-voice can strongly influence participant behaviour in almost

any risk management direction and thus performance direction. Just as providing an anthropomorphic speech interface to can influence trust, performance and even buying habits (Large et al. 2019), so providing synthetic recommendations with explanations through the audio-voice can be used to influence decision-making behaviour, either encouraging safety behaviour, or conversely, it can also be used to encourage risk taking. This indicates care should be taken in safety critical systems when implementing voice interfaces, to ensure that positive, safety related voice messages are prioritised.

In hindsight perhaps this observed significant impact of synthetic spoken recommendations should have been anticipated as previous researchers (eg Schaefer et al. 2016) had identified that operator trust can improve when working with reliable systems that communicate using human speech. However, older research had also show that the uptake or use of automation decision support is not always directly related to the trust in automation, with Wiegmann, Rich and Zhang (2001) observing that even when participants trust automation highly, they do not always use it as frequently as expected to optimise performance.

The observed change in behaviour led to the question "was the trust over-trust leading to the participants effectively abdicating decision-making to the synthetic agent?' This would be of particular concern in safety critical systems such as aviation pilotage and traffic control, as decision abdication would be a highly unsafe behaviour especially in situations of automation degradation. However, the results of both the Pilot and UTM studies do provide some evidence that the participants, whilst strongly affected, did not over-trust and did still retain a level of independence in their decision-making.

In the Pilot Study at LOA5 when the synthetic agent provided recommendations as spoken messages, 75% of the risk-taking suggestions were accepted and actioned; however, conversely, 25% were not. Of the accepted recommendations, 66% were then overturned independently by the participants within an average of 8 seconds before any further recommendation was provided by the synthetic

agent. Thus, the participants would not always accept risk-taking suggestions, and when they did, they retained a level of attention and were frequently prepared to over-rule the suggestion and take their own counsel and manage risk as they preferred.

In the UTM study when the synthetic agent was Anthropomorphic, participants accepted 85% of recommendations, and they reduced the number of independent decisions from an additional 66% (in the Silent condition), to just an additional 41% of the standard profile of expected actions. Again, whilst these figures could be used to support the presence of decision-making abdication, these statistics show that even though participants accepted many spoken recommendations, they did not accept all of them. Results show participants reduced the number of independent actions, but they still continued to take independent actions at a rate of just over one independent action for every three recommendations provided. Thus, the recommendation acceptance rates show increase acceptance of advice, but not abdication.

The deduction that decision-making was not abdicated is also supported by the performance results from the period of automation degradation in the UTM study. When the speaking automation degraded and provided both a warning that its recommendations might be flawed along with an explanation for why, the participant rate of acceptance of the recommendations dropped significantly and there was an increase in the number of additional independent decisions taken after the synthetic warning.

Returning to the Independent Variables, in the Shisa Kanko study where the Operator Speech was being tested, the results show that when the participant was required to speak to the synthetic agent, the participant task completion rate slowed, and their decision-making became more safety orientated; more "risk intolerant". In the study, in order to allow for possible variance in rate of performance the Teaming Structure had been set at LOA3 where no risk-taking recommendations were provided. As a result, it is not known if the significant influencing effect of the spoken reasoned recommendations (from LOA5 and the UTM study) could have over-ridden this "risk intolerant"

behaviour observed when the participant spoke, or conversely, whether the participant speech may have moderated the significant influence of the recommendations. Further research is needed to study this potential interaction of participant speech and synthetic recommendations.

Overall, the collected results from all three studies support the general research project hypothesis that providing an audio-voice communication capability to the Human Autonomy Team (HAT) will improve the performance of the operator. However, the results indicate that the direction that the performance will improve is dependent upon three key factors:

- Recommendations: The provision of decision-support advice as a recommendation for action will strongly influence the behaviour of the participant either towards or away from risk.

- Reasoning Transparency: The provision of detailed reasoning providing an explanation or rationalisation for advice will act as an amplifier for recommendations, increasing the influence of those recommendations.

- Operator Speech: Requiring the participant to speak can strongly influence the participant, encouraging a slower more considered rate of work and risk-adverse safety behaviour.

The first two of these factors appear to work together to drive the performance in a direction heavily influenced by the synthetic, and third appears to work primarily towards safety and potentially offers a source of influence moderation.

7.4 Situation Awareness

Interestingly, the results on Situation Awareness (SA) produced by the Situation Awareness Global Assessment Technique (SAGAT) neither matched the general variation in performance nor the expectation from Endsley (1995a) that there would be a positive relationship between SA and performance; that for performance to have changed there must have been a change in SA.

7.4.1 SAGAT Evaluation of SA

The SAGAT results indicate that, despite the presence of the audio-voice communication causing significant changes in both task completion and safety behaviours, SA was affected by one independent variable alone: whether the participant spoke (Operator Speech). In both the Pilot Study and the UTM Study the participant SA as measured by SAGAT did not change. From the SAGAT results SA appeared to be unaffected by the presence of the message content (Pilot study Text condition) or the delivery of those messages as auditory speech, irrespective of the affect those had on performance. Only in the Shisa Kanko study, when the participant was required to talk to the synthetic agent, did the SAGAT register an improvement in SA. Interestingly, the results of the Shisa Kanko study show that the improvement in SA was not just for recall of information that the participant spoke but was also for recall of information that the participant heard. Thus, it appears that the act of speaking could have affected the participant's attention to both sides of the communication. This needs further research to clarify the effect of talking on alertness, attention and listening to the synthetic.

The generic observation from across the three studies, if using the results of the SAGAT alone, is that simply listening to audio-voice messages (Pilot and UTM study) is not sufficient to improve SA; it is necessary to fully engage in the communication, in the conversation, for that audio-voice conversation to have a positive effect on the explicit SA of the participant. This directly supports the UK Civil Aviation Authority (CAA) requirements of Crew Resource Management (CRM) and the requirements of Shively et al. (2017) for HAT to have bi-directional communications to be effective.

However, before accepting the deduction on a limited change in SA taken from the SAGAT results in isolation, it is worth re-appraising the raw SAGAT results, and reflecting on the capability of SAGAT to make a complete and thus systemic assessment of participant SA for this research project. Reviewing the Literature, other researchers have also had unexpected and perhaps unusual results when using the SAGAT. As observed by Stanton, Salmon and Walker (2015), in an earlier study (Walker et al. 2009)

SAGAT demonstrated an apparent inverted relationship between communication media complexity and SA; the participant SA was lowest when the abstracted media information available to them was greatest. This would appear to be counter intuitive but may explain why providing additional information (as either text or voice) had no incremental effect on volume of information recalled for the SAGAT even though it had an apparent effect on decision-making and behaviour. As suggested by Stanton, Salmon and Walker (2015) it may be that the participants did not need to remember (recall) information spoken as it was apparently held by the system (the synthetic agent) and therefore only needed to be applied (and promptly forgotten!).

As well as the unexpected lack of variance between conditions, the other anomaly in this research was the systemic poor SAGAT results. In all three of the studies conducted the absolute values of the SAGAT interrupts appeared to be very low with participants consistently unable to answer the majority of questions, even though they knew in advance what the questions would be and had between 10 - 13 separate exposures to the questions to learn them (there were 4 interrupts in training and practice and 3 interrupts per condition).

In the Pilot study, the best rate of SAGAT recall response was 2 ± 2 questions answered correctly (out of between 21-25 questions asked). In the Shisa Kanko study this best rate of response rose to 4 ± 3 (out of 21-25 questions), and in the UTM study peaked at 5 ± 3 (out of between 19-23 questions which incidentally included 5-7 recognition questions). Thus, even in the best SA performance study, participants were only likely to answer 42% of questions (8 out of 19 in UTM study). All rates of response seem within the expected scope of instantaneous recollection as predicted by current Working Memory theories (eg Baddeley 2010); however, the percentage scores that they generate seem at odds with those reported by other researchers. For example, Endsley and Kaber (1999) using the same experimental apparatus and methodology as the Pilot and Shisa Kanko study reported SAGAT scores of between 48% and 71% whereas the range of scores in this research was between 4% and 33%.

Lo et al. (2016) also reported low absolute SAGAT score when attempting to measure the SA of expert Train Traffic Controllers and came to the conclusion that their participants' SA must have been largely implicit and therefore unmeasurable by SAGAT. The reflected deduction made during the progress of the studies and discussed in the Pilot study (Chapter 4.4.4) and the introduction of the Shisa Kanko study (Chapter 5.1), was that the SAGAT methodology which primarily uses recall questions generally samples explicit knowledge from conscious processing. However, for the participants in the first two studies, the task was easy and highly automated and relied heavily upon visual processing of information, with the result that much of their cognitive processing would likely have been unconscious (automated processing and some elements of visual processing are unconscious). As it would likely be largely unconscious, it would also likely have been largely non-declarative and their SA knowledge of "what was going on" with the task would have been implicit. This agrees with Lo et al.'s (2016) deduction that for their study the SAGAT worked poorly because the task was highly automated and the majority of the information that the participants were collecting for their SA was therefore largely non-declarative implicit knowledge.

This deduction that SA was largely implicit also offered an explanation for the only noticeable variation in SA as measured by SAGAT; the variance in knowledge spoken by the participant and synthetic in the Shisa Kanko study. The explanation was that speaking observations and SA information aloud, forcing it into conscious memory, likely assisted participants convert implicit knowledge to explicit knowledge. That explicit knowledge could then be used to answer SAGAT questions and thus gain a higher SA score.

Accepting SA could have been largely implicit, following advice from Endsley (2004) graphical recognition questions were included in the SAGAT interrupts of both the Shisa Kanko and UTM studies to attempt to sample implicit SA. In the Shisa Kanko study (Chapter 5) these were graphical multi-choice questions where participants selected a visual match for the screen immediately prior to the interrupt (see Figure 7.1 below), and in the UTM study required participants to identify visually on a

map where UAVs had been immediately prior to the interrupt. These contrasted significantly from the standard recall questions which were text multi-choice asking for example "What Colour was Target 3" or "How far from the Centre was Target 1" or "How many Intercepts have you called".
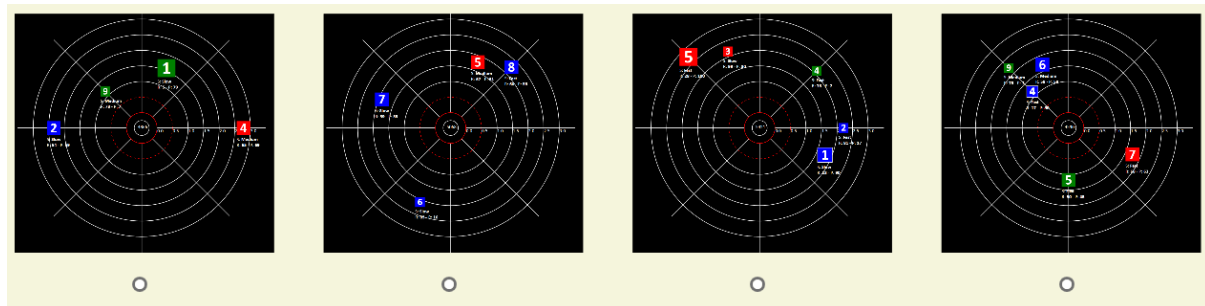

Figure 7.1 Example of SAGAT Multi-Choice Recognition Question

Both sets of recognition questions obtained consistently higher scores than recall questions. In the Shisa Kanko study recognition questions scored between 84%-86%, but recall questions scored between just 4%-33%. In the UTM study recognition questions scored 46%-47%, but recall questions score between 6%-41%. These recognition results demonstrated that a proportion of the participants' SA, and likely a considerable proportion of the SA, was implicit, non-declarative and would not be measurable using SAGAT recall questions. However, despite these high absolute scores, there was no significant variance in recognition scores between any conditions in either study. Despite attempts to make the SAGAT as inclusive as possible the results from it remained unyielding, meaning a lack of difference in SA was likely not a measurement artefact, but a genuine lack of difference between the SAGAT measured SA of participants, conditions and studies.

7.4.2 Decision-Making and SA

As explained earlier, the generally low absolute SAGAT results and lack of variance in those results indicates that the study participants had a poor SA that was un-affected by the presence or absence of message content or audio-voice synthetic speech. However, this observation is completely at odds with the often significant change in performance and decision-making observed in all three studies. It is understood that there is no guarantee of a direct correlation between "good" SA and "good"

235

decision-making, with Endsley's (1995: 36) observing "even the best-trained decision makers will make the wrong decisions of they have inaccurate or incomplete SA .. a person who has perfect SA may still make the wrong decision". Nevertheless, it would seem highly improbable that a large cohort (of say 24-36 participants) would all suffer from the same disconnect between SA and decision-making. As Endsley (2000: 7) explains "Decisions are formed by SA and SA is formed by decisions". Following this logic, if a change in decision-making is observed then that change is either a result of a change in SA, or is about to cause a change in SA. Therefore, it would seem safe to assume that any observed change in performance, across a cohort, would indicate a generic change in SA across that same cohort.

If performance, and more specifically the decision-making, of the participants is used as an indicator of SA, the general trend of participants in the audio-voice condition of all three studies to change their decision-making behaviour indicates that the presence of the audio-voice messages inevitably led to a change in SA, but that change was not measured by the SAGAT. However, how that SA changed was very much affected by the content of the audio-voice message.

In the Pilot study, in the lower LOA when participants heard the audio-voice warning and information messages their behaviour change indicates that they became more aware of threats to goal failure, and consequently became more alert, working harder and changing their target selection criteria. In the higher LOA, when recommendation messages were included, the participants appeared to become more aware of alternative but riskier options for improving their score that they were prepared to take.

In the Shisa Kanko, when participants also spoke as well as heard messages, their behaviour and thus SA changed again, with participants slowing down the rate of processing, and their behaviour indicating that they took longer to appreciate the situation and became more aware of likely goal failure threats. This in turn suggests that they became more aware of the time available for decision-making, making them less hurried and more selective.

Finally, in the UTM study, in normal conditions in the presence of detailed audio-speech messages the participants again slowed down decision-making and became simultaneously more resolute and less error prone, indicating a more safety conscious but confident SA. In the abnormal condition of automation degradation in the UTM study, the presence of the reasoning explanation in the synthetic speech messages lead to them being significantly more selective in the recommendations they accepted, although they did still tend to wait for the synthetic agent to initiate decision-making by providing a warning.

Reflecting back on these variances in decision-making and thus SA, it appears that the independent variables of Teaming Structure and Reasoning Transparency both had an effect on the SA being constructed, as did the Operator Speaking. Obviously, from the SAGAT results, the Operator Speech had the greatest and only measured effect. However, the Teaming Structure, through the provision (or not) of recommendations, and Reasoning Transparency, through the provision (or not) of explanation for recommendations, appeared to be able to direct or drive how the participants formed SA. The results indicate that if the synthetic agent provides recommendations emphasising safety, then SA becomes safety orientated. Alternatively, if it provides recommendations that emphasize taking a risk and justify why to take that risk (eg Pilot Study LOA 5) then SA becomes more risk tolerant. Or, if it were to provide recommendations that demonstrate the competence and reliability of the synthetic (eg UTM study Anthropomorphic) the SA directs increased trust and reliance on those recommendations.

7.4.3 Situation Assessments and SA

The conclusion of the Literature Review of Endsley's Model (Chapter 2.2.1.2.3) and the conclusion of individual SA (Chapter 2.2.1.6) was that SA consists of both process and product. Therefore, to measure SA it is not sufficient to simply measure the product of SA using SAGAT alone. If possible, the process of building SA which Endsley's (1995a) model of SA identifies the process of SA as the taking of situation assessments, should also be evaluated. Unfortunately, due to practical limitations in the

design and implementation of the "multi-task" simulator apparatus this was not achieved in the Pilot and Shisa Kanko study. However, an evaluation of the frequency and type of situation assessment could be taken by measuring the frequency with which participants requested "additional" information about the UAVs in the UTM Simulator study.

The SA results of the UTM study show that in the Anthropomorphic condition when the synthetic agent provided detailed spoken messages the frequency of requests to view operational plans and UAV details decreased. At the same time, the product of SA either remained constant (if using only SAGAT results) or improved (if using decision-making and performance as an indicator of SA). Thus, the results of the UTM study (Chapter 6) show that when provided with Anthropomorphic speech messages the participants appeared to expend less effort to build the same if not better SA. This effect could have been a consequence of the UAV details being spoken as well as displayed in the two synthetic speech conditions; hearing the information on the data page could have made it easier to remember longer and thus assisted make SA more robust. However, that does not explain the variance between the two speech conditions; it does not explain why there was further reduction again in making situation assessments in the Anthropomorphic condition as the only difference between the Agentic and Anthropomorphic conditions was the explanation and that provided no information tested for in the SAGAT. This is worth further research to attempt to identify the source of SA processing improvement.

Interestingly the feedback from the participants in the post-activity debrief and the reduction in SA data sampling, offer some evidence for the anticipated cognitive timesharing (Wickens 2008) discussed in the Introduction (Chapter 1.1) and the Literature Review (Chapter 2.2.1.2.3). Some participants complained that in the Silent condition, without audio-voice SA information participants had to work hard, dividing their attention between evaluation of the graphical map and reading the Data Pages. However, when receiving the Data Page information as speech in the Agentic and Anthropomorphic condition, participants indicated that they did not need to turn away from the map

but were able to take in information visually (icon colour and movement on the map) and auditorily (information on registration and operations plan state), with the apparent consequence of significantly reducing the number of samples needed to maintain the same level of SA.

Perhaps as important as the implementation of cognitive timesharing and the reduction in situation assessments is the affect this reduction in effort to build SA could have on cognitive workload. The reduction in frequency of situation assessments indicates a likely reduction in effort to process and build SA, which will result in the freeing up of cognitive capacity, creating cognitive spare capacity. This finding is particularly valuable as it demonstrates that adding an audio-voice capability to a synthetic agent through which it can provide detailed and reasoned SA messages, could reduce cognitive workload and increase the amount of cognitive capacity available to then apply the SA that has just been built. As discussed, the UTM study (Chapter 6.4.5) this spare capacity could then be put to use to improve decision-making and improve performance, both improvements being observed in the Anthropomorphic condition of the UTM study (see Chapter 6.4.4.1 and 6.4.4.2). Alternatively, it could be kept in reserve and used to surge effort to build SA in unusual or abnormal periods of, for example, automation degradation. This is effect was also observed in the UTM study where participants who were receiving detailed audio-voice messages from the synthetic agent when automation degraded were able to increase the frequency of situation assessments (see Chapter 6.4.6.3).

7.4.4 Summary on SA

The determination at the beginning of the research project, as identified in the Methodology (Chapter 3.3) was to use both SAGAT and, where possible, measurements of situation assessments as the primary and quantitative measure of SA and to only use observations on changes to decision-making and performance to provide insight on how any observed changes in SA likely came about.

With only SAGAT results for the Pilot Study, despite observations that the decision-making varied across the LOA, the conclusion is that in the Pilot Study SA did not vary. However, in the Shisa Kanko

study, the SAGAT results show that SA did vary, improving with the addition of audio-voice communication. Finally, in the UTM study, although the SAGAT did not find variance in the product of SA, the product remaining constant, the measurement of situation assessments did show an improvement in the process of SA. Thus, the conclusion of the UTM Study was that the presence of audio-voice speech with detailed messages lead to an improvement in SA.

Thus, with two of the three studies showing an improvement in SA, the overall conclusion of the research project is that the introduction of an audio-voice speech capability, if used to transfer SA information, can improve SA. However, how that SA is improved is linked to how the communication interface is implemented. Add a communication channel from the synthetic agent to the human and make the synthetic agent provide detailed explanations of SA information (ie provide not just Perception but also higher levels of abstracted SA in the form of Comprehension and Projection) and the process of building SA can be made less onerous. Add a communication channel from the human to the synthetic agent where the human articulates essential SA information and the human's declarative product of SA will expand and improve. Thus, the deduction is that add both together an SA will be improved as both a product and a process, providing the greatest change in SA.

Interestingly, all the SA results discussed above are very much in alignment with the expectation of Transactive SA (Salmon et al. 2010), where the act of deliberately transferring SA observations from one member of the team to another results in expansion and changes in the recipients individual SA.

7.4.5 The Final Word on SA – Reflections On The Use Of SAGAT

For integrity as SAGAT was selected as primary measure of SA in the methodology, the determination in 7.4.4 above was to continue to use the SAGAT results to evaluate variance in SA. However, the other results in this project, in particular the observation that variance in decision-making indicated variance in SA (that the SAGAT failed to detect) does raise a question over whether there is an issue with SAGAT that needs discussion before a decision is made to use it the future.

There are two main issues of concern. Firstly, the SAGAT results of this project and other researchers (eg Lo et al. 2016) show that SAGAT, by only measuring the conscious and therefore declarative product of SA, is providing an unnecessarily limited view of SA. It is overlooking valuable information on SA that could have an equally significant impact on performance such as the non-declarative product of SA and the process of building SA through repeated situation assessments. Secondly, and potentially of greater concern, is that the results this project and those of other researchers (eg Lo et al. 2016, Chen et al. 2018), have provided results that appear to show that variations in SAGAT scores of SA and performance are not always directly correlated as SA theory suggests they should be. Both Lo et al. (2016) and Chen et al. (2018) have reported improvements in SAGAT lead to an improvement in performance but have also observed that an improvement in SAGAT can be associated with a decrease in performance. To add further uncertainty over the relationship, the results from this project demonstrated an improved performance with no associated change in SA. In fact, the results when viewed collectively seem to indicate that variations in SAGAT results and variations in decision-making are often hard to consistently correlate. These two observations and concerns do raise questions on the actual value and even validity of SAGAT.

7.4.5.1 Is SAGAT Complete and Sufficient?

The first concern for discussion is that the results of this study do indicate that SAGAT, by only measuring the declarative part of SA, is providing a partial and incomplete observation.

As discussed above in 7.4.1, it was difficult to get SAGAT to provide a view of non-declarative SA, and even those results obtained in the Shisa Kanko and UTM studies were inconclusive. Furthermore, as discussed in the Literature Review (Chapter 2.2.1.6) SAGAT does not provide a view of the efficacy of the process of building SA. Finally, by attaching a score to SA it is attaching a value to an observation; how 'correct' the SA observed is. The outcome of this potentially myopic view is a temptation to regard the scope of SA set by SAGAT questions to be a complete and accurate definition of SA for the research studies conducted. As a consequence, even amongst SA experts there could be (and arguably

has been) a tendency to limit research into evaluating the impact of factors on SAGAT scores, rather than attempting to obtain and explore broader views of SA and the impact the factors have on the effort to build SA.

Interestingly, Endsley provides a case study that illustrates the issue and concern. In a study into team and shared SA, Bolstad and Endsley (2003) observed that the results of a SAGAT test identified that the Commander of an Army Headquarters (HQ) in a simulation had a lower SA than the other members of the HQ. The apparent gap was in knowledge on tactical details such as which of the Commanders units were having difficulties with their task and which units had changed their mission parameters. Conversely, and perhaps unsurprisingly, the Commander had superior strategic knowledge on "the location of friendly and enemy units, which friendly units are firing weapons, enemy's force capabilities, enemy objective's, and the impact of terrain on friendly unit's missions" (Bolstad and Endsley 2003:372). The conclusion of the research was that "SA was not distributed amongst the cells, demonstrating less than optimal levels of shared SA on information that should have been shared between different positions" (Bolstad and Endsley 2003:373).

The problem here is that the above SAGAT-centric interpretation of the results was that because the Commander could not answer the tactical questions he had a 'low' SA, and that this could only be because he was not provided with all the information being tested for in the SAGAT. The assumption was SAGAT was correct, and the Commander's SA was poor. If following that determination to its logical conclusion the designer would then attempt to build an information system that provided the Commander with the 'missing' tactical information. However, at no point in the research paper was there any discussion on whether the lack of knowledge was an issue or not. There was no discussion on whether the Commander wanted or needed that information. In fact, there appeared to be no check if the assessment of SA obtained from the SAGAT was valid or complete. There was no assessment to see if the Commander had SA not covered by SAGAT questions, nor if his 'low' SA was because he wanted to keep spare cognitive capacity so as to wrestle with complex problems.

Unfortunately, by focusing solely on SAGAT measurements and making value judgements from those measurements, the study failed to consider that the Commander may not have needed to build the same SA as other members of his HQ. In dynamic situations many individuals will actually store information externally to release cognitive capacity (Chiappe, Strybel and Vu, 2015) which the Commander could have done by tasking his sub-ordinates to hold tactical knowledge so that he could focus on generating SA on how his overall strategy and plan was progressing.

Similar issues can be seen in other research. The research described by Chen et al. (2018) focused on whether providing specific information would lead to an improvement in SAGAT scores and only rarely discussed the effect of providing that information. Interestingly, in one of the three studies the researchers found that adding information to attempt to improve SAGAT scores resulted in a degradation in performance that they struggled to explain, eventually putting it down to increased participant complacency induced by the extra knowledge. It would thus appear that they admitted that adding information to improve SAGAT scores lead directly to a reduction in performance, which does raise the question did the new information really improve SA or did it make it worse?

Of course, these are only examples and does not mean all researchers using SAGAT will automatically limit their analysis to the effects provided by SAGAT nor make value judgments on SA observed, but it does demonstrate the traps of using the outcome of SAGAT to evaluate the efficacy of an information system and of focusing research towards measuring how to affect SAGAT scores. The main issue with only focusing on the product of SA when designing a system is that the design then becomes circular. The goal of research become solely focused on improving SA as reported by SAGAT purely for its own ends, not to improve SA for the benefit of improving performance.

SA theory is clear, the value of SA is in how it is used to affect and drive decision-making and performance; therefore, the only true measure of the quality of SA is what affect that SA has on the behaviour and competence of the person it resides in. Without measuring performance effect there is no way to determine if the changes to the SA are beneficial, nugatory or worse harmful. It is not

sufficient to measure accuracy of knowledge against a limited view of SA, it is necessary to explore the full stretch and shape of SA without making judgement on it. It is more important to know how changes to SA, especially the higher abstractions, are driving changes to goal prioritisation and other factors that drive performance. It is impractical (and likely impossible) to prepare sufficient SAGAT questions to cover all (appropriate) SA eventualities sufficient to provide a complete view of SA and even then, those questions would likely be unable to evaluate cognitive effort and the relative (subjective) priority that participants placed on each piece of information.

Therefore, accepting that SAGAT is incomplete and insufficient the priority should be to evaluate what is done with SA and how it reflects changes in the key factors of SA such as Goals, Schema and the very temporal process of building SA and the workload of that process. Ironically, it would seem, measurements of what SA affects, decision-making, are more likely to provide an accurate view of SA and its variance as can be seen from the results of this project.

7.4.5.2 Is SAGAT Valid?

In addition to raising questions on scope, the discussion and conclusion above in 7.4.5.1 also leads to the questions over whether SAGAT is even providing a valid assessment of SA. The definition of validity as provided by Boslaugh and Watters (2008: 12) is "how well a test or rating scale measure what it is supposed to measure". The primary validity issue emerging from this project is that the results provided by the SAGAT, whilst fairly consistently able to provide a measure of declarative memory, did not detect a variation in SA that a range of non-SAGAT methods were able to clearly identify (and in fact provide meaningful detail on). In fact, the only variance in SA detected by the SAGAT was the ability to recall specific words that were the answers to two of the SAGAT questions, that the participants were required to repeatedly say aloud and hear, a process that sounds interestingly like a memory retention 'trick'.

Conversely, the non-SAGAT results of the three studies provided both quantitative and qualitative evidence that SA varied: the quantitative evidence from observations in changes to decision-making

behaviour and target selection and frequency of situation assessments; and, the qualitative evidence from reports from the participants on how hearing the synthetic agent helped them change their locus of attention, their scope of knowledge and comprehension of the screen activity, and their anticipation of how the synthetic agent would behave and provide support.

7.4.5.2.1 Playing the SAGAT Memory Game?

In fact, the qualitative results of the studies, particularly those from the Pilot Study, provide the strongest argument against the validity of the SAGAT measurement. When being interviewed many participants complained that they struggled to naturally recall the SA information asked in the SAGAT interrupt, and a third of them explained that they gave up attempting to remember the information. In the recorded interviews one participant admitted "I was more involved in making sure things didn't crash than remember the 'SA' data", while another admitted to having to resort to the use of a memory retention 'trick' they knew of to hold onto information to answer SAGAT Situation Awareness questions: to quote the participant "I recited its numbers in my head so I could answer at least one question on the survey". Two thirds of participants (16 of 24) in the Pilot Study indicated that their focus of attention was on evaluating the relative positioning and movement of targets, not the colour, size, penalty score, reward score, absolute location, absolute speed, and time to impact gathered in the SAGAT. The indication here is that the knowledge being sampled by the SAGAT questions was not pertinent to the majority of participants. To answer SAGAT questions the participants had to make additional effort to remember specific information they appeared not to use. Of course, it could be argued that the issue was potentially inappropriate selection of the SAGAT questions (see the discussion in section 7.4.5.2.2). However, in this particular study the SAGAT questions were supplied in the method description used to generate the simulator for the study (Kaber and Endsley 1997), one of the authors happening to be the creator of the SAGAT process.

These qualitative observations indicate there was a disconnect between the SA reported and the knowledge that the SAGAT questions measured. When coupled with the non-SAGAT quantitative

results (showing significant variation in SA dependent decision-making) the researcher is lead to the conclusion that, in this project at least, the SAGAT questions were not necessarily measuring what they were supposed to be measuring. They were measuring declarative knowledge, but the results indicate that the declarative knowledge may not have been related to operational SA.

7.4.5.2.2 Its the question's fault!

Of course, the primary proponents of the methodology do provide caveats that the SAGAT method is highly dependent on the quality and appropriateness of the probe questions, explaining that "the foundation of a successful SAGAT data-collection effort rests on the efficacy of the queries" (Jones and Kaber 2005: 42-2). However, in this project two distinctly separate sets of SAGAT questions were asked, the first set generated in part or at least with approval of the author of the SAGAT methodology (Endsley and Kaber 1999), and the second set provided by a UTM and Counter Drone subject matter expert currently employed as lead advisor to a major arm of the UK government. Unfortunately, neither set of questions was able to discern a change in SA that performance measures were able to clearly identify. Despite the researcher's best efforts to generate high quality and pertinent SAGAT questions, the SAGAT results were consistently unable to demonstrate a variance in SA that correlated with observations raised by almost all other measures.

Ironically the reliance upon quality raised by Jones and Kaber (2005) may in fact present the greatest challenge to the generic validity of SAGAT; that the questions of SAGAT represent the subjective and individual SA of the examiner, meaning SAGAT only measures the extent to which an individual can recall knowledge determined by the examiner to be pertinent to SA. It does not provide an answer to what else the participant may also have been thinking about and thus cannot be guaranteed to detect the full scope of an individual's SA. Like any educational examination it can only test against a standard, it cannot answer the question "yes, but what does their SA really look like?". As a result of its examination approach all it actually tests is "is your SA as good as mine?"

The non-SAGAT results of this project indicate that the participants were not thinking what the examiner preparing the SAGAT questions expected they would be thinking. The results indicate that is it likely that the participants had an SA that was different to that predicted by the examiner and tested for with SAGAT.

It is important to observe that the arguments above are all made against validity not success. The problem is not that SAGAT did not give a measure, SAGAT definitely gave a measure of some knowledge being held. As discussed above in 7.4.1. the SAGAT did successfully sample participant's knowledge. The problem is the non-SAGAT results suggest that SAGAT measured knowledge might not be the knowledge being used for SA. Simply put, SAGAT clearly measures something, the results just suggest that something is unlikely to be SA.

Before finishing the discussion, it must be cautioned that the results of must be taken in context and it is accepted that they provide only limited evidence against the use of SAGAT. However, whilst the results of this project are not sufficient to categorically determine that SAGAT is flawed as it is consistently incomplete and can easily provide invalid data, they are perhaps sufficient to raise questions in researchers over the value of its use and its focus on 'chasing the product' that are certainly worth further post-project study (theoretical or practical).

The caution does not mean that the researcher has determined that there is no value in the freeze and interrupt probe approach that SAGAT is based upon. The primary issue raised against SAGAT is the validity of its examination approach to establish SA, not its reliability to interrupt without (overly) affecting the SA it is sampling. In fact, it is suggested that the freeze probe practice could be used as an excellent method to very quickly gather short samples of qualitative data in a "hot debrief" using open ended questions to attempt to identify exactly what knowledge individuals are paying attention to and are using to drive their decisions. Freeze probes used in this way could provide a method to sample the scope of SA being used rather than attempting to measure the absolute accuracy of SA.

7.5 Teaming

Whilst performance and SA were tested in all three studies, the Perception of Teaming and Trust were only tested in the Pilot study and the UTM study. The results of the participant Perception of Teaming and Trust taken in those two studies were much clearer and consistent than the results for performance and SA, with participants almost universally and overwhelmingly viewing the synthetic agent as a better teammate when the agent spoke to them, supporting the general teaming Hypotheses 3 and showing that the presence of the audio-voice messages had a consistently significant impact on teaming and trust.

The quantitative measures and various qualitative comments from the participants about the synthetic agent provide assurance that the design and programmatic implementation of the audio-voice communication discussed in the Methodology (Chapter 3.2) was sufficient to engage the participants in the studies and allow them to identify the system automation as an intelligent synthetic entity with its own agency. Although this synthetic agent was extremely limited in its capability, only able to provide messages constructed from the short category of speeches given in the Methodology (Chapter 3 Table 3.1) and only messages that were directly related to the team goals, its capability was sufficient to generate a positive view of it in all audio-voice speech conditions.

The results are very much in line with the results of Waytz, Heafner and Epley (2014) that the more human-like speaking agent was trusted more by participants than non-speaking. In fact, the results show that the more talkative and transparent the agent became the more the participants subjectively trusted and valued it. The results of the survey of Perception of Teaming in the Pilot study show that the highest average Teaming score was given for the synthetic agent in LOA9 when the synthetic agent was almost constantly providing transparency information on its actions. In the UTM study both the Teaming (CAPTEAM) and Trust (HAT Trustworthiness) scores were highest for the Anthropomorphic synthetic agent. In that study, where the synthetic agent was both relatively high (LOA5) and provided abstracted SA information as explanations (SA L2 Comprehension) and recommendations (SA L3

Projection), the participants behaviour indicates a particularly high level of reliance on the synthetic agent, appearing to act as if they expected the synthetic agent to fulfilling a specific team role. Thus, the results show that the two independent variables of Teaming Structure and Reasoning Transparency have the greatest effect on Perception of Teaming and Trust. Simply put, the more work the synthetic agent was assigned, the more support the synthetic agent gave, and the more transparent the synthetic agent was, the more that the participants viewed it as a teammate and trusted and relied upon it.

The qualitative post-exercise debriefs also provide near universal support for the audio-voice communication conditions over the non-verbal graphical conditions. In the Pilot study, of the 24 participants, 21 said that they preferred the speaking synthetic agent over the non-speaking agent, and 16 said they trusted the speaking synthetic agent more than the non-speaking agent. In the UTM study, 35 of the 36 participants preferred the speaking agent; however, surprisingly, more of the UTM participants preferred the short spoken Agentic synthetic agent (17) over the longer spoken Anthropomorphic agent (13).

This last expression of preference for the Agentic agent is interestingly, as the results of the CAPTEAM and HAT Trustworthiness surveys showed that both Perception of Teaming and Trust measures were highest for the Anthropomorphic agent. The participants seemed to objectively report and view the longer spoken Anthropomorphic agent as a more trustworthy and generally better teammate, yet subjectively preferred to engage with the shorter spoke Agentic agent. The effect of this apparent conundrum can be seen with the statistical analysis of performance and warning response times in the UTM study. The overall performance of the participants was at its best (safest) when the synthetic teammate audio-voice messages contained a detailed explanation; yet the average reaction time to the synthetic warning indicates that some participants may not have actually heard that explanation. Thus, it would seem that knowing that the synthetic agent could provide an explanation for its recommendations could be sufficient to increase participant trust and affect performance, especially

decision-making performance and acceptance of recommendations. This effect warrants further research to determine whether it is the presence of the transparency information, or the availability of the transparency information that is improving teaming and trust.

The increase in perception of teaming and trust with Team Structure and Reasoning Transparency also provides an explanation for the observed changes in decision-making behaviour in the performance discussion above (section 7.2) that raised the question of whether the participants were abdicating decision-making to the synthetic agent. The question is whether the participants were over-trusting the synthetic agent.

In the Pilot study, Perception of Teaming at LOA5 was higher that the two lower LOA, and within LOA5, was greater for the Voice condition than the other two conditions. In the UTM study, trust was highest for the Anthropomorphic condition. In these two study condition pairs (Pilot study at LOA5, and UTM study in Anthropomorphic) participants were highly influenced by the synthetic agent, more readily accepting recommendations than in any other conditions. Thus, as participant trust is highest, so is acceptance of recommendations, a perhaps not unsurprising correlation as identified by Large et al. (2019).

The results show that when the trust is at its highest, the synthetic recommendations appear to be able to encourage participants to accept risks and carry out actions that they would not normally consider (eg taking a risk by prioritising high scoring targets in the Pilot study LOA5). The danger is of course that this could demonstrating the presence of over-trust. However, the statistics used in section 7.2 show that even in the Pilot study the participant trust was calibrated and did not degenerate into over-trust, and in fact the addition of explanation information assisted calibrate that trust even further.

However, whilst the results show that trust may have been calibrated across the studies, the results of the UTM study indicate that reliance upon the synthetic was possibly not so calibrated and remained consistently high even in the periods of automation degradation. This is a potential concern

as, when conditions are wrong, it could lead to the very "out-of-the-loop" issue that the addition of synthetic speech is hoped to mitigate (eg consistently accepting faulty recommendations without understanding what is flawed with them could lead to being out of the loop).

However, before making the judgement that high levels of reliance is a totally negative effect, it is worth remembering that the aim of introducing the audio-speech voice was not just to improve SA but was also to attempt to improve the teaming. In their model of HAT, Battiste et al. (2018) indicated that the human should be able to delegate tasks to the synthetic agent. The results of studies showed that in the UTM study participants had high levels of trust and teaming, and high utilisation of the synthetic warnings and recommendations, appearing to rely upon it to carry out tasks with relatively low oversight and interference, an approach that sounds exactly like trusted delegation. Hence the conclusion of Chapter 6 that the high level of reliance in the speaking synthetic could be viewed as positive evidence that the participants had accepted the synthetic agent as a teammate and had granted it with appropriate autonomy to carry out its "delegated" tasks. Furthermore, the fact that the participants continued to monitor and occasionally over-ride should be taken as evidence that the reliance was itself headed towards calibrated reliance.

Thus, the fact that the participants relied upon the speaking synthetic agent is not totally taken as a concern but instead also taken to provide reassurance that the presence of the audio-voice capability assisted participants enter into a teaming like relationship with the synthetic. It is suggested that further research is conducted to evaluate calibrating reliance to achieve an optimum teaming inter-dependency.

Thus, overall, the results of the studies all appear to provide strong support for the generic HAT hypotheses, drawn from research from the likes of Wynne and Lyons (2018) and Przegalinska et al. (2019), that when automation appears to have agency and is recognisably anthropomorphic the human operator is more inclined to identify the automation as a synthetic teammate. Furthermore, the results directly support the Hypothesis 3 of this research, that providing the automation with an

audio-voice speech capability will provide both that apparent agency and the anthropomorphic communication and will facilitate the creation of a HAT in the view of the human operator.

7.6 Limitations

Whilst every effort has been made to ensure that the research conducted for this project was done so as robustly and extensively as possible, it has been impossible to avoid some practical compromises that placed limitations on the research project:

- Synthetic Agent Technology. The automation implemented in the studies, whilst appearing to the participants in the study to be intelligent, was in fact extremely limited in capability, especially scope of speech and teaming. If intending to undertake future research where the synthetic agent is required to conduct longer more complex conversation exchanges, especially counter factual "what if" situations or task sharing cooperation activities, consideration should be given to either obtaining a "chatbot" conversational AI for synthetic speech or implementing a scripted Wizard of Oz methodology.

- SA Data In Communication Messages. In order to ensure that the impact of the auditory channel was being assessed rather than the impact of the content of message content being sent, the SA information in the messages was limited to the information available from the graphical display. This limited the potential scope of taking advantage of cognitive timesharing identified by Multiple Resource Theory, which would have been achieved by using the audio-voice channel to communicate information that was not currently displayed on the screen. The expectation is that if the channel is used to provide novel information (rather than just replicating current visual information) a greater increase in the effects on SA, performance and teaming observed in this research could be achieved without necessarily any commensurate significant increase in cognitive workload or reduction in cognitive spare capacity. In future research consideration should be given to using the auditory channel to provide novel SA information and instead of measuring the resultant SA product, measure the

cognitive workload associated with building SA especially when the visual and auditory channel are both providing different (but not conflicting or contrary) information.

- Participants. The majority of participants were recruited from Coventry University and did not have an aviation background or extensive aviation knowledge so may have found the tasks quite unusual and unfamiliar. This may have made them more prone to reliance upon the advice and recommendations provided by the synthetic agent. For future studies, especially those planning on implementing more capable synthetic agents delivering more complex conversations, a more specialised and skilled participant group should be recruited.

- Laboratory Conditions. The studies were all carried out in a quiet room using a highly abstracted simulator with a controlled rate of work and controlled incidents of automation degradation which may have amplified the size of effects observed. Future studies should be conducted in a more applied or realistic setting, or if kept in the laboratory, provided with more non-linear scenarios and more realistic simulators.

- Overly Dynamic Short Duration Scenarios. The study scenarios and conditions were all very short (only 10 minutes) and highly dynamic resulting in an unrealistically high work rate and little opportunity to conduct longer conversations. It would be appropriate to design future study scenarios and simulators of longer duration that have considerably fewer but more complex tasks that require concentration and team coordination to achieve. This will allow the implementation of speech messages containing more complex SA data that would allow the evaluation of the achievement of more complex and abstracted SA using methods other than simple SAGAT recall questions.

- Performance Measures. The task completion performance measures in the studies were not ideal as the set processing speed of the automation effectively heavily constrained the maximum rate of task at the middle to high LOAs implemented. Future studies should reduce the number of tasks, but increase their complexity with sub-tasks that can have a variable

level of task completion success thus giving a measure with more capability for measuring variance.

Whilst it is accepting that due to these limitations care must be taken when applying the results to systems design, the thesis results and conclusions have highlighted that the provision of a speech communication capability is a valid tool for achieving HAT and gaining the advantages in human SA and performance and can be taken forward for further research.

7.7   Future Research

This research programme has provided results that demonstrate that there is definite value to be gained for human operator SA and performance and for HAT by implementing an anthropomorphic auditory speech communication capability. However, as well as providing results the research has also provided further observations of knowledge gaps that offer an opportunity for further research. These include:

- Research into measuring the implicit SA of the human members of a team, especially when that team is engaged in a visually demanding dynamic activity.

- Research into the increased and improved use of graphical recognition questions in SAGAT.

- Research into alternative measuring systems or methodologies for evaluating SA through measurement of ancillary or secondary effects (eg decision-making, communication quality and quantity).

- Research to determine whether, when the human speaks to the automation, there is any "masking" of SA information not explicitly articulated by either human or automation.

- Research to determine whether the Independent Variable of Direction of Communication implemented as bi-directional communication could act as a safety-orientated moderator to the strong influence of the synthetic Detailed Recommendation.

- Research to evaluate whether there is any correlation between the incident of synthetic speech interruption and operator (and team) performance.

- Research to evaluate the impact of the auditory channel providing novel SA information simultaneous to the graphical channel providing different novel information on SA and cognitive workload.

- Research to evaluate the relationship between frequency of situation assessments and performance in periods of automation degradation.

- Research to evaluate the relationship between previous cognitive spare capacity and performance in periods of automation degradation.

- Research whether there is a relationship between the reliability of automation and the operator's willingness to grant it autonomy and whether it is possible to achieve calibrated reliance.

7.8   Summary

Overall, the results of the three studies provide evidence to support all three of the project hypotheses and provide general support for most aspects of the fundamental research question of the research project. The results of the three studies provide evidence that the addition of the audio-voice communication led to improvements in the participants' Situation Awareness (SA), performance and the reported perception of teaming and trust in the synthetic agent. Furthermore, the results also demonstrate that the content of the messages, particularly any recommendations provided, can have a significant and fundamental impact upon the SA and performance allowing the synthetic agent able to become highly influential and even able to lead the participant like any trusted advisor.

Chapter 8 – Conclusion

8.1   Introduction

There were two major sets of findings and conclusions that emerged from the research project. The first set are of conclusions are those emerging from the data on the human operator responses in the experimental studies of the study, the findings obtained from the observations and deductions on implementing the automation speech capability.   In addition, an unexpected and potentially significant second set of findings emerged from the experience of attempting to use the Situation Awareness Global Assessment Technique (SAGAT) methodology of measure Situation Awareness (SA). Both sets of findings and discussions on their contributions are included below.

8.2   Implementing A Speech Capability for Automation Systems

The goal of the research project was to determine whether providing an auditory speech communication capability to an autonomous synthetic agent would provide a channel through which key SA information (readily available graphically) could be transferred to improve human SA. That improved SA could then positively affect performance sufficiently to mitigate those humans becoming "out-of-the-loop". The project also aimed to evaluate whether implementing an audio-voice speech communication capability could contribute significantly towards the creation of the teaming effect required for Human Autonomy Teaming (HAT).

The results of the studies, discussed in length in Chapter 7, conclusively demonstrated that providing the automation of a transport management system with an auditory speech capability with which to provide timely and meaningful SA led to significant improvements to the SA, performance behaviour, trust and perception of teaming of a human operator of that system. Furthermore, the results were able to demonstrate that the improvements could be affected, almost tuned, by manipulating factors such as the Level of Automation (LOA) and the presence or absence of an explanation in the speech messages.

These findings by themselves are meaningful and have impact, demonstrating that implementing a speech interface, which is now common in consumer electronics and home automation, is more than just a gimmick and can deliver enhanced benefits beyond simple voice commands for everyday tasks. Implementing an auditory speech capability with which automation can provide key SA information can change and improve the human operator's perception of the system, calibrating their trust and making them more willing to permit the system to act autonomously. The results show that the speech capability can help generate the teaming relationship theorised in HAT literature (eg Demir, McNeese and Cooke 2018a) and reap the expected performance benefits of that relationship.

However, and perhaps even more meaningful, the results also demonstrate that all this can be achieved with an automation agent that does not meet all of the behaviour and processing requirements expressed by many researchers (see the Literature Review Chapter 2.3.4 for a detailed explanation of expected capabilities). The autonomous synthetic agents implemented in this project were not trivial, being able to provide timely observations, warnings and advice and ask for operator input and decisions. However, they were generally not able to work with the participant to co-ordinate work effort and dynamically transfer work between synthetic agent and human; nor where they able to engage in problem solving exchanges. This means that they were sufficiently limited in scope that they would be considered to fall short of some of the key the requirements of HAT such as Team Interdependence (Goodman et al. 2017) and Coordination and Cooperation (Demir et al. 2015, Wynne and Lyons 2018) and ability to solve counterfactual "what if" questions (Battiste et al. 2018).

Furthermore, the speech engine prepared for the synthetic agent was also limited in capability by design, unable to facilitate exploratory and planning conversations back and forth between human and synthetic. The speech engine was designed to mimic the highly standardised and limited conservation flow found in aviation challenge and response flight check sheets and standard operating procedures. The automation was not capable of, nor intended to replicate "free-speech".

Despite this, the results, both quantitative and qualitative, provide robust evidence that the participants believed themselves to be engaging with a synthetic team-mate. The fact that the synthetic agent was not fully autonomous and could not have long and detailed conversations did not affect their relationship with the synthetic agent. The provision of the auditory speech from the automation was sufficient to create a relationship, calibrate trust and reap the improved SA and performance benefits expected of a human entering into a HAT.

It is this latter observation that shows possibly the greatest impact of the findings of this project. The results indicate that it is not necessary to wait for some highly complex future AI capable of free speech to be able to gain benefit from the concepts and tenets of HAT. Improvements in SA, performance and trust can all be achieved immediately with current technology through the implementation of a limited capability synthetic agent that has a voice and uses it to provide essential SA observations, advice and even guidance. Furthermore, the speech generated by that synthetic does not have to achieve the full scope of human conversation. Rather it can, and should, follow the tenets of the Crew Resource Management and be limited to clear, concise and unambiguous communications designed to specifically improve SA (CAA 2014a). This will ensure human teaming best practice for safe and reliable teaming communication is applied whilst at the same time significantly reducing the software development complexities that could hold back implementation.

It is relatively easy to envisage how such a synthetic agent could be implemented to provide a teaming interface in future vehicle automation systems currently being researched such as the single pilot commercial flight deck. Providing the automation with an anthropomorphic voice that communicated the range of information that a human Co-Pilot would be expected to deliver during routine checks and when monitoring instruments would help the Pilot perceive the automation as an independent human-like agent and help the pilot to believe themselves in a HAT. However, the same speech capability could be used for the automation to highlight and explain changes in automation mode as

the aircraft follows its flight plan that are so essential to Pilot SA, thus helping keep the human "in-the-loop" on automation activity.

An appropriate method of demonstrating the application of these conclusions is to return to examining the example of the Air France 447 air crash cited in the Introduction. The results of this project indicate that a solution to attempt to prevent the two pilots becoming out of the loop would have been for the automation of the aircraft to be embodied as a speaking and apparently autonomous synthetic agent that they were used to frequently obtaining SA information from. Then, when the automation had to disengage, it would have spoken and announced that it was disengaging because it was unable to obtain a unified reading from speed instruments. If the Pilots were used to the system speaking and had improved SA, performance and perception of teaming (as this research indicates they would), then whilst the announcement of partial disengagement might not have prevented the pilots entering into the reaction states and undertaking their subsequent actions, the results indicate that they would likely have had more trust in the system and more cognitive spare capacity to deal with the situation and would have likely been on the right side of the Yerkes-Dodson Law curve. The speech from the autopilot would have left the Pilots in a better position of SA and thus in a better position to make an informed decision on what action to take.

8.3   SA, SAGAT and Decision-Making

Using SAGAT in the Studies

As explained in the Literature Review (Chapter 2.2.1.6), partly because of its recognition and partly because of the availability of the SA measurement technique SAGAT, the Endsley (1995a) model of SA would be given primacy for this research project, and the SAGAT methodology would be as the primary measure of SA.

However, as discussed in the Literature Review, it was apparent that the model of SA might have some limitations. Specifically, it was apparent that Endsley's (and indeed Smith and Hancock's 1995) determination that SA and decision-making were separate seemed incongruous. Endsley's theory of

SA made it clear that the very creation of SA involved making decisions, and that decision-making and SA were highly dependent and almost inseparable. Yet they were declared as separate.

In addition, concern was raised that SAGAT, by focusing in only on the resultant outcome of building SA, the product of SA (knowledge in the head), did not gather evidence of the other key half of SA, the process of collecting and generating SA (also identified as the process of taking situation assessments). These issues are discussed in the Literature Review (Chapter 2.2.1.6).

Cognisant of these reservations, but nevertheless aware of the precedence of previous (very well cited and senior) researchers' experience, every effort was made to utilise SAGAT to draw observations about SA. When, in the first study the SAGAT seemed to underperform providing very low absolute scores and results that indicated SA did not vary, efforts were made to account for this lack of variance and changes made to participant behaviour for the second study to attempt to improve SA results.

These efforts did appear to partially work, with the Shisa Kanko study providing results that showed a variation in SA. However, absolute scores remained extremely low and later reflection on the new results lead to the uncomfortable observation that the variation in SA could in fact be attributed to the new participant behaviour being remarkably similar to a 'memory retention trick' (repeatedly saying SAGAT answers out aloud until the SAGAT interrupt). This behaviour was abandoned in the final UTM study, with an unsurprising return to low absolute results that indicated SA, as inferred by SAGAT, did not vary.

At the same time, the more objective decision-making measures taken in all three studies provided overwhelming evidence that the participants' perception of what was going on, comprehension of what were the threats were, and projections of which action to take were constantly changing between conditions. The details of this discord and determination of what it meant are covered in length in the Discussion (Chapter 7.4.5) and led to the conclusion that SAGAT was incomplete, and worse, prone to being invalid. This leads to the conclusion that, in our studies, SAGAT did not measure all of SA, and what it did measure was not necessarily what it was supposed to measure.

Is SA Fit for Purpose?

This observation and finding now forces attention to the very models of SA from which SAGAT is derived and raises the question, is the model of SA also flawed? Endsley (1995a: 36) very clearly states that "SA is explicitly recognised as a construct separate from decision-making and performance" whilst simultaneously arguing that they are inexorably intertwined (Endsley 2000: 7 "Decisions are formed by SA and SA is formed by decisions"). As explained in the Literature Review (Chapter 2.2.16) it is difficult to imagine forming the two top levels of SA, comprehension and projection, without making a decision. In fact, it could be argued that both are actually decisions: comprehension being a decision on what is happening; and, projection being a decision on what is likely to happen next

This of course raises the even more extreme question, is in fact the concept of SA even valid? Does SA even exist? Some of the explanations of SA seem so ubiquitous ("knowing what is going on around you") that they could be mistaken for definitions of consciousness. Although the concept of SA has been in scientific publication for over 26 years, it appears modern cognitive psychology does not recognise SA, as neither of the two dominant cognitive psychology texts consulted (Sternberg and Sternberg 2012, Eysenck and Keane 2015) mention it. They mention attention, schema, sight and decision-making, but not SA. Conversely, the models of SA often explain SA in terms that encapsulate a number of facets of cognitive psychology. The model describes the process of making SA as the comparison of observations to schema to reach conclusions and generate a living schema; so, even SA theory appears to consider SA to be a combination of decision-making and memory. Which of course then leads to the logical deduction that there is no SA, there is just decision-making and memory.

SAGAT – A Warning

Whilst this research project can only open the door to make the latter challenge of whether SA is valid, it does perhaps provide sufficient evidence to provide some health warnings on SA to other researchers, especially those determining to utilise SAGAT to measure SA. The warning is simple; taking a view SA is independent of decision-making and then using SAGAT to measure SA could easily

lead to incomplete and even invalid results. SAGAT encourages users to limit their view of SA to the product of knowledge, and then only to a facet of the product of SA that corresponds to the SAGAT question-maker's determination of what is 'good' SA. If researchers use SAGAT and focus on product not process and exclude decision-making from consideration, they run the risk of focusing too entirely on how to generate knowledge for its own sake, rather than evaluate whether the presence of that knowledge leads to an improvement in performance and even a reduction in workload. The danger is that the research could then lead to the determination to prioritise delivery of information without consideration for the effort used to collect it and the impact that information has on decision-making and performance.

Acknowledging that researchers may be reluctant to simply discard the SAGAT approach entirely, the advice to complement the above warning is that if a researcher must make use of a freeze and interrupt probe, then at least make the questions open ended and exploratory (but remember to make the interrupt as short as possible). Use the probe to ask two simple questions:

- What goal is your current priority?

- What information are you focusing on and why?

The Value of Decision-Making

The position of this project is that, as per SA theory, SA is meant to be key to decision-making, it is meant to drive decision-making, therefore its value can only be determined in relation to the effect it has on decision-making. In the Pilot study participants were presented with goal specific information on reward and penalty (presented using colour and figures) that were expected to influence behaviour. Yet, more often than not, the target selection data indicates that the size, location and relative movement of the targets were what most affected decision-making. The results show us that even though the designers (and researcher) thought reward and colour information would be valuable to SA, it was in fact almost useless as most participants simply wanted to calculate which target was most a risk of crashing. The target selection results demonstrate that it is only when researchers

observe decision-making that it is possible to identify whether information is valuable to SA, as that is the only way to determine whether a specific piece of information is used to drive a decision.

Thus, it would seem that the argument on whether and how to measure SA returns to the relationship between SA and decision-making. It is the conclusion of this project, based upon results and discussions, that instead of focusing on simply recording the current state of memory to determine what an individual thinks of the current situation, it would be far more valid to measure how individuals collect data and how that data affects their decision-making.

In fact, continuing that line of reasoning, the overarching conclusion of this project, accepting the earlier contention that SA and decision-making cannot be separated, is that measuring decision-making will actually provide a direct line to evaluating the scope and direction of SA. Moreover, it will provide a much more detailed and complex view of SA, beyond a simple test against a predetermined standard. The results of this project provide direct evidence to support this contention, with the measures of decision-making taken (eg participants target selection preferences, participants acceptance of recommendations) providing insight into which goals participants were prioritising, what risks participants were willing to take, and how the presentation of information could influence participants' projection of what was right and wrong to do. The decision-making results don't just show that SA changed, they don't just identify what more or less was known, they provide insights into how and why SA changed and what the effect was of that SA change.

The conclusion that measuring decision-making provides a more complete and more valid measure of SA should be of value for research beyond the scope of this project, especially research into complex systems of multiple agents both human and synthetic. It could be used to evaluate how the data flowing around a complex system of human and synthetic agents affects the SA of each agent and influences their behaviour. It may even provide a method by which it is possible to gain a view, even quantify, the SA of an information system (a synthetic agent), by for example evaluating the structure and complexity of the synthetic decision-making code.

Process over Product

What has not yet been considered in this project and should be investigated next is how to carry out this measurement task with pure objectivity and with sufficiency to encapsulate and measure all facets of decision-making. In this project decision-making was not simply taken as outcome of selecting a single option, but rather as the process of generating options for consideration and then subsequent selection. Rather than measuring the outcome of decision-making, the measures examined decision-creation behaviour. In the first two studies this was achieved by comparing characteristics of the target selected to those of other targets available in an attempt to understand what goal objective lead the participant to select that target above the others. To this latter process holds the key to evaluating "SA" for human factors research; attempting to evaluate what situation assessment processes and decision options generation processing were occurring and how did they relate to goals? What do those processes tell us about what the individual thought was going on, what goal they had as priority and what they considered important to do?

Reflecting back on the measures taken of decision-making, and in fact on the other measures of performance, teaming and SA (non-SAGAT) taken throughout this project (Table 8.1 overleaf) it can be seen that many of them provide some degree of information about an individual's SA and offer an alternative approach to obtaining a view of the dynamic SA of a participant. The recommendation for those attempting to measure SA without, or in addition to, SAGAT is to create a suit of measures designed to meet one or more of the types of measure listed under the "Evaluating" Column of Table 8.1:

- Decision Selection Comparison. Compare the decision option selected against other options available. This measure should provide insights into what participant goals were, and which were a priority.

- Participant Knowledge Gathering. Measure how much knowledge has to be gathered and the effort (time and difficulty) of getting that knowledge. This will give a view of the cognitive workload and efficacy of the processes of making Situation Assessments.

- Task Information Requirement. Measure what the participants want/need to know, not what they actually are able to remember. The applied knowledge requirement and value is more pertinent to the task and is thus applied SA rather than generic memory.

- Decision Processing Efficacy. Measure the number of errors, mistakes and corrections in decision-making. This can give valuable insight into potential gaps or weaknesses in the SA, or participant under-confidence in that SA, both of are important when studying human factors of system interfaces or teaming interaction.

Furthermore, if the study involves working in a team, irrespective of whether the teammate is human or synthetic:

- Decision Making in Presence of Recommendation. Measuring how recommendations are accepted and rejected provides insight into SA of and across the team.

These measures may not provide an evaluation of SA "correctness" (but there again is there such a thing) but they will provide a view of the scope of an individual's SA and the impact a system or team on that SA; they will provide a valuable insight into the needs and process of building SA.

In summary, SA research needs to fundamentally re-direct itself; to stop focusing on determining what information is known and instead focus on how its presentation is used to generate and then take decisions. The value in SA research is not to determine what information to provide, but to determine what value is attributed to it, how is it used and, above all else, how the medium of presentation of that information will affect when and how individuals are able to find it and use it. That approach speaks back to the ergonomics history of human factors. This researcher certainly intends to take this approach towards future SA research and would encourage others to do the same.

Forget product, focus on process.

Measure	Dependent Variable	Evaluating	Study	Data Type Source?	Significant Results?	SA Product?	SA Process?	Contri provic
Score	Performance	Task Completion Success	1,2,3	Quantitative	Y	N	N	Overa
Reward Score	Performance	Task Completion Success	1,2,3	Quantitative	Y	N	N	Overa
Penalty Score	Performance	Task Completion Success	1,2,3	Quantitative	Y	N	N	Overa
Reaction Time	Performance	Task Completion Success	1,2,3	Quantitative	N	N	Y	Attent
Relative Target Size	Performance	Decision Making - Target Selection	1,2	Quantitative	Y	N	Y	Goal F
Relative Target Distance	Performance	Decision Making - Target Selection	1,2	Quantitative	Y	N	Y	Goal F
Relative Target Reward	Performance	Decision Making - Target Selection	1,2	Quantitative	Y	N	Y	Goal F
Relative Target Penalty	Performance	Decision Making - Target Selection	1,2	Quantitative	Y	N	Y	Goal F
Absolute Target Distance From Barrier	Performance	Decision Making - Target Selection	1,2	Quantitative	Y	N	Y	Goal F
Absolute Target Distance From Targets	Performance	Decision Making - Target Selection	1,2	Quantitative	Y	N	Y	Goal F
Directions Given	Performance	Decision Making In Presence of Recommendation	3	Quantitative	Y	N	Y	Rate o
Recommendations Followed	Performance	Decision Making In Presence of Recommendation	3	Quantitative	Y	N	Y	Confic
Independent Directions	Performance	Decision Making In Presence of Recommendation	3	Quantitative	Y	N	Y	Confic
Directions Given Before an Alert	Performance	Decision Making In Absence of Recommendation	3	Quantitative	N	N	Y	Confic
Directions Given After a Recommendation	Performance	Decision Making In Presence of Recommendation	3	Quantitative	Y	N	Y	Confic
Decision Change of Mind	Performance	Decision Processing Efficacy	3	Quantitative	Y	N	Y	Volati
Decision Selection Error	Performance	Decision Processing Efficacy	3	Quantitative	Y	N	Y	Locus
Ongoing Process Over-ride	Performance	Decision Processing Efficacy	3	Quantitative	Y	N	Y	Volati
Repeated Instructions	Performance	Decision Processing Efficacy	3	Quantitative	Y	N	Y	SA of
Flight Plan Reads	SA	Participant Knowledge Gathering	3	Quantitative	Y	N	Y	Workl
Flight Card Reads	SA	Participant Knowledge Gathering	3	Quantitative	Y	N	Y	Workl
Screen Information Focus Question	SA	Task Information Requirement	1,3	Qualitative	N/A	Y	Y	Focus
Reported Human Autonomy Teaming	Teaming	HAT Requirements	1	Qualitative	Y	Y	N	Confic
CAPTEAM	Teaming	Teaming Behaviours	3	Qualitative	Y	Y	N	Confic
Configuration Preference Question	Teaming	Subjective Teaming Preference	1,3	Qualitative	N/A	Y	N	Subjec
Trust Preference Question	Teaming	Subjective Teaming Preference	1,3	Qualitative	N/A	Y	N	Confic
HAT Trustworthiness	Trust	Trust	3	Qualitative	Y	Y	N	Confic

Bibliography

Altitude Angel (2021) *GuardianUTM O/S* [online] available from <https://www.altitudeangel.com/solutions/guardianutm-os/>

Atrash et al (2009 Atrash, A., Kaplow, R., Villemure, J., West, R., Yamani, H., and Pineau, J. (2009) 'Development and Validation of a Robust Speech Interface for Improved Human-Robot Interaction.' *International Journal of Social Robotics* 1, 345-356. available from <https://doi.org/10.1007/s12369-009-0032-4>

Avutu, S. R., Bhatia, D., and Reddy, B.V. (2017) 'Voice Control Module for Low Cost Local-Map Navigation Based Intelligent Wheelchair' in Sai, Y.P., and Garg, D. (eds.) *2017 IEEE 7th International Advance Computing Conference (IACC)* held 5-7 January 2017 at VNR Vignana Jyothi Institute of Engineering and Technology. [online]: IEEE, 609-613 [1 July 2021]. available from <https://doi.org/10.1109/IACC.2017.0129>

Baddeley, A. (2010) 'Working Memory'. *Current Biology* 20 (4), R136-R140. available from <https://doi.org/10.1016/j.cub.2009.12.014>

Bainbridge, L. (1983) 'Ironies of Automation' in Johannsen, G., and Rijnsdorp, J.E. (eds.) *Proceedings of the IFAC/IFIP/IFORS/IEA Conference,* 'Analysis, Design and Evaluation of Man-Machine Systems'. held 27-29 September 1982 in Baden-Baden. Oxford: Pergamon Press Ltd, 129-135. available from <https://doi.org/10.1016/B978-0-08-029348-6.50026-9>

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. Cham: Springer, 479-493. available from <https://doi.org/10.1007/978-3-319-92046-7_40>

Bedny, G., and Meister, D. (1999) 'Theory of Activity and Situation Awareness', *International Journal of Cognitive Ergonomics* 3 (1), 63-72. available from <https://doi.org/10.1207/s15327566ijce0301_5>

Billings, C.E. (1995) 'Situation Awareness Measurement and Analysis: A Commentary'. In Garland, D.J., and Endsley, M.R. (eds.) *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness.* held 1-3 November 1995 at Embry-Riddle Aeronautical University Daytona Beach. Daytona Beach: Embry-Riddle Aeronautical University Press, 1-5. available from <https://apps.dtic.mil/sti/citations/ADA522540>

Bolstad, C.A., and Endsley, M.R. (2000) 'The Effect of Task Load and Shared Displays on Team Situation Awareness'. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44(1), 189-192. available from <doi:https://doi.org/10.1177/154193120004400150>

Bolstad, C.A., and Endsley, M.R. (2003). 'Measuring shared and team situation awareness in the army's future objective force'. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 47(3), 369-373. available from <doi.org/10.1177/154193120304700325>

Boslaugh, S., and Watters, P.A. (2012) *Statistics in a nutshell: A desktop quick reference*. Sebastopol: O'Reilly Media, Inc.

Bradshaw, J. M. Hoffman R. R., Woods D. D., and Johnson, M. (2013) 'The Seven Deadly Myths of "Autonomous Systems'. *IEEE Intelligent Systems* 28 (3), 54-61. available from <https://doi.org/10.1109/MIS.2013.70>

Bogg, A., Birrell, S., Bromfield, M.A., and Parkes, A.M. (2020) 'Can we talk? How a talking agent can improve human autonomy team performance'. *Theoretical Issues in Ergonomics Science*, 22 (4) 488-509. available from <doi.org/10.1080/1463922X.2020.1827080>

Bogg, A., Parkes, A., and Bromfield, M. (2020) 'Can we talk?–the impact of conversational interfaces on human autonomy teaming perception, performance and situation awareness'. in Ahram,

T., Karwowski, W., Vergnano, A., Leali, F., and Taiar, R. (eds.) *International Conference on Intelligent Human Systems Integration* held 19-21 February 2020 at Modena, Italy. Cham: Springer, 938-944. available from <doi.org/10.1007/978-3-030-39512-4_143>

Brindley, P.G., and Reynolds, S.F. (2011) 'Improving verbal communication in critical care medicine'. *Journal of Critical Care* 26, 155-159. available from <doi.org/10.1016/j.jcrc.2011.03.004>

Bruder, C., and Hasse, C. (2020) 'What the eyes reveal: investigating the detection of automation failures'. *Applied Ergonomics*, 82, 102967. available from <doi.org/10.1016/j.apergo.2019.102967>

Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (2012) *Final Report on the accident on 1ts June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro – Paris.* [online] available from <https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf> [1 July 2021]

Bunz, M. (2019) 'Conversational Interface'. in *The Oxford Handbook of Media, Technology, and Organization Studies.* ed. by Beyes, T., Holt, R., and Pias, C. Oxford: Oxford Univerity Press, 149-159. available from <https://doi.org/10.1093/oxfordhb/9780198809913.013.14>.

Calhoun, G.L., Ruff, H.A., Behymer, K.J. and Frost, E.M. (2018) 'Human-autonomy teaming interface design considerations for multi-unmanned vehicle control'. *Theoretical issues in ergonomics science*, 19 (3), 321-352. available from <doi.org/10.1080/1463922X.2017.1315751>

Chakrabarty, A., Ippolito, C., Baculi, J., Krishnakumar, K., and Hening, S. (2019) 'Vehicle to Vehicle (V2V) communication for Collision avoidance for Multi-copters flying in UTM-TCL4' in *AIAA Scitech 2019 Forum*. held 7-11 January 2019 in San Diego. [online] available from <https://doi.org/10.2514/6.2019-0690> [1 July 2021]

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. (2014). *Situation awareness-based agent transparency (Report No. ARL-TR-6905)*. Defence Technical Information Centre [online]. available from <http://www.dtic.mil/docs/citations/ADA600351> [1 July 2021]

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. [online]. available from <https://doi.org/10.1109/SMC.2016.7844505> [1 July 2021]

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 <https://doi.org/10.1080/1463922X.2017.1315750>

Chen, Y., Tong, Z., Wu, W., Samuelson, H., Malkawi, A., and Norford, L. (2019) 'Achieving natural ventilation potential in practice: Control schemes and levels of automation'. *Applied Energy* 235, 1141-1152. available from <https://doi.org/10.1016/j.apenergy.2018.11.016>

Chiappe, D., Strybel, T.Z., and Vu, K-P.L. (2015) 'A Situated Approach to the Understanding of Dynamic Situations'. *Journal of Cognitive Engineering and Decision Making* 9(1), 33-43. available from <https://doi.org/10.1177/1555343414559053>

Cohen, J. (1992) 'A Power Primer'. *Psychological Bulletin* 112 (1), 155-159. available from <https://doi.org/10.1037//0033-2909.112.1.155>

Cooke, N., Demir, M., and Huang, L. (2020) 'A Framework for Human-Autonomy Team Research'. in: Harris D., and Li WC. (eds) International Conference on Human-Computer Interaction, 'Engineering Psychology and Cognitive Ergonomics. Cognition and Design'. held 19-24 July 2020 in Copenhagen. Cham: Springer, Cham. available from <https://doi.org/10.1007/978-3-030-49183-3_11>

Corus (2019) *U-space Concept of Operations.* [online] available from <https://www.sesarju.eu/node/3411> [1 July 2021]

Costa, P.L., Passos, A.M. and Bakker, A.B. (2015) 'Direct and contextual influence of team conflict on team resources, team work engagement, and team performance'. *Negotiation and Conflict Management Research*, 8 (4), 211-227. available from <doi.org/10.1111/ncmr.12061>

Dekker, S.W.A. (2015) 'The danger of losing situation awareness'. *Cognition, Technology and Work* 17**,** 159–161. available from <https://doi.org/10.1007/s10111-015-0320-8>

Demir, M., Cooke, N.J., and Amazeen, P.G. (2018) 'A conceptual model of team dynamical behaviors and performance in human-autonomy teaming' *Cognitive Systems Research* 52, 497-507. available from <https://doi.org/10.1016/j.cogsys.2018.07.029>

Demir, M., McNeese, N.J., Cooke, N.J., Ball, J.T., Myers, C., and Frieman, M. (2015) 'Synthetic Teammate Communication and Coordination With Humans'. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 59(1), 951-955. available from <https://doi.org/10.1177/1541931215591275>

Demir, M., McNeese, N.J., Cooke, N.J., Ball, J.T., Myers, C., and Frieman, M. (2015). 'Synthetic teammate communication and coordination with humans'. in *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 59, No. 1, pp. 951-955). held 26-30 October 2015 at JW Marriott Los Angeles Hotel. Los Angeles: SAGE Publications.

Demir, M., McNeese, N. J., and Cooke, N. J. (2016) 'Team Communication Behaviors of the Human-Automation Teaming'. in *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://doi.org/10.1109/COGSIMA.2016.7497782>>

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://doi.org/10.1016/j.cogsys.2016.11.003>>

Demir, M., McNeese, N.J., Cooke, N.J. (2018a) 'Team Synchrony in Human-Autonomy Teaming'. in Chen, J. (ed). *International Conference on Applied Human Factors and Ergonomics*, 'Advances in Human Factors in Robots and Unmanned Systems'. held 17-21 July 2017 at The Westin Bonaventure Hotel Los Angeles. Cham: Springer, 303-319. Available from <https://doi.org/10.1007/978-3-319-60384-1_29>

Demir, M., McNeese, N.J., and Cooke, N.J. (2018b) 'The Impact of Perceived Autonomous Agents on Dynamic Team Behaviors' *IEEE Transactions on Emerging Topics in Computational Intelligence* 2(4), 258-267. available from <https://doi.org/10.1109/TETCI.2018.2829985>

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 <doi.org/10.1016/j.cogsys.2016.11.003>

Demir, M., Likens, A.D., Cooke, N.J., Amazeen, P.G., and McNeese, N.J. (2019a) 'Team Coordination and Effectiveness in Human-Autonomy Teaming' *IEEE Transactions on Human-Machine Systems* 49(2), 150-159. available from <https://doi.org/10.1109/THMS.2018.2877482>

Demir, M., McNeese, N.J., Johnson, C., Gorman, J.C., Grimm, D., and Cooke, N.J. (2019b) 'Effective Team Interaction for Adaptive Training and Situation Awareness in Human-Autonomy Teaming' in Rogova, G.L., McGeorge, N., Gundersen, O.E., Rein, K., and Freiman, M. (eds.) *IEEE International Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. held 8-11 April 2019 in Las Vegas. [online] available from <https://doi.org/10.1109/COGSIMA.2019.8724202> [1 July 2021]

de Visser, E.J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S. and Parasuraman, R. (2012) 'The world is not enough: Trust in cognitive agents'. in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 263-267). held 22-16 October 2012 in Westin Boston Waterfront Hotel. Los Angeles: Sage Publications.

de Visser, E., Pak, R., and 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://doi.org/10.1080/00140139.2018.1457725>

Durso, F.T., and Dattel, A.R. (2004) 'SPAM: The Real-Time Assessment of SA'. in *A Cognitive Approach to Situation Awareness: Theory and Application*. ed. by Tremblay, S., and Banbury, S. London: Routledge, 137-154

Edgar, G.K., Catherwood, D., Baker, S., Sallis, G., Bertels, M., Edgar, H., Nikolla, D., Buckle, S., Goodwin, C., and Whelan, A. (2018) 'Quantitative Analysis of Situation Awareness (QASA): modelling and measuring situation awareness using signal detection theory', *Ergonomics* 61(6), 762-777. available from <https://doi.org/10.1080/00140139.2017.1420238>

Endsley, M. R. (1988) 'Situation awareness global assessment technique (SAGAT)'. In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, held 23-27 May 1988 at Dayton. New York: IEEE, 789-795. available from <10.1109/NAECON.1988.195097>

Endsley, M. R. (1995a) '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 <https://doi.org/10.1518/001872095779049543>

Endsley, M.R. (1995b) 'Measurement of Situation Awareness in Dynamic Systems'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37(1), 65-84. available from <https://doi.org/10.1518/001872095779049499>

Endsley, M.R. (2000) 'Theoretical Underpinnings of Situation Awareness: A Critical Review' in *Situation Awareness Analysis and Measurement.* ed. by Endsley, M.R., and Garland, D.J. London: Lawrence Erlbaum Associates, 3-23

Endsley, M.R. (2004) 'Situation Awareness: Progress and Directions'. in *A Cognitive Approach to Situation Awareness: Theory and Application*. ed. by Banbury, S., and Tremblay, S. London: Routledge, 317-341

Endsley, M.R. (2015) 'Situation Awareness Misconceptions and Misunderstandings'. *Journal of Cognitive Engineering and Decision Making*. 9 (1), 4-32. available from <10.1177/1555343415572631>

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. (2018) 'Situation awareness in future autonomous vehicles: Beware of the unexpected'. in Bagnara, S., Tartaglia, R., Albolino, T., and Fujita, Y. (eds.) *Proceedings of the 20th Congress of the International Ergonomics Associatio*n. held 26-30 August 2018 in Florence. Cham: Springer. 303-309 available from <doi.org/10.1007/978-3-319-96071-5_32>

Endsley, M.R., Bolté, B., and Jones, D.G. (2003) *Designing for Situation Awareness.* London: Taylor & Francis

Endsley, M.R., and Jones, W.M. (2001) 'A Model of Inter and Intra-Team Situation Awareness: Implications for Design, Training and Measurement' in *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments*. ed. by McNeese, M., Salas, E., and Endsley, M.R. Santa Monica: Human Factors and Ergonomics Society, 1-24. available from <Endsley: research gate>

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>

Endsley, M.R., and Robertson, M.M. (2000) 'Situation awareness is aircraft maintenance teams'. International Journal of Industrial Ergonomics 26 (2), 301-325. available from <https://doi.org/10.1016/S0169-8141(99)00073-6>

Eriksson, A., and Stanton, N.A. (2017) 'The chatty co-driver: A linguistics approach applying lessons learnt from aviation incidents'. *Safety Science* 99 (A), 94-101. available from <https://www.sciencedirect.com/science/article/abs/pii/S0925753517308731>

Evans, J.S.B.T. (2006) 'The heuristic-analytic theory of reasoning: Extension and evaluation'. *Psychonomic Bulletin & Review* 13 (3), 378–395. available from <https://doi.org/10.3758/BF03193858>

Evans, J.S.B. (2019) 'Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning'. *Thinking & Reasoning*, 25 (4), 383-415. available from <doi.org/10.1080/13546783.2019.1623071>

Eysenck, M.W., and Keane, M.T. (2015) *Cognitive Psychology*. 7th edn. London: Psychology Press. available from <https://doi.org/10.4324/9781315778006>

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012) 'Effect Size Estimates: Current use, Calculations, and Interpretation'. *Journal of Experimental Psychology. General* 141 (1), 2-18. available from <https://doi.org/10.1037/a0024338>

Furukawa, H., Nakatani, H., and Inagaki, T. (2004) 'Intention-Represented Ecological Interface Design For Supporting Collaboration With Automation: Situation Awareness And Control In Inexperienced Scenarios'. in *Human Performance, Situation Awareness and Automation*. ed. by Vincenzi, D.A., Mouloua, M., and Hancock, P.A. New York: Psychology Press, 49-55

Garbis, C., and Artman, H. (2004) 'Team Situation Awareness as Communicative Practices'. in *A Cognitive Approach to Situation Awareness: Theory and Application*. ed. by Banbury, S., and Tremblay, S., London: Routledge, 275-296. available from <https://doi.org/10.4324/9781315263977>

González-Martínez, E., Bangerter, A., and Lê Van, K. (2017) 'Building Situation Awareness on the Move: Staff Monitoring Behavior in Clinic Corridors'. *Qualitative Health Research* 27 (14), 2244-2257. available from <https://doi.org/10.1177%2F1049732317728485>

Goodman, T.J., Miller, M.E., Rusnock, C.F., and Bindewald, J.M. (2017) 'Effects of agent timing on the human-agent team' *Cognitive Systems Research* 46, 40-51. available from <https://doi.org/10.1016/j.cogsys.2017.02.007>

Grimm, D., Demir, M., Gorman J. C., and Cooke N. J. (2018) 'The Complex Dynamics of Team Situation Awareness in Human-Autonomy Teaming' in Rogova, G.L., Lebiere, C., Gundersen, O.E, Salfinger, A., and Baclawski, K. (ed.) *2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA).* held 11-14 June 2018 in Boston. New York: IEEE, 103-109. available from <https://doi.org/10.1109/COGSIMA.2018.8423990>

Groom, V., and Nass, C. (2007) 'Can robots be teammates?: Benchmarks in human-robot teams' *International Studies* 8(3), 483-500. available from <https://doi.org/10.1075/is.8.3.10gro>

Guznov, S., Lyons, J., Pfahler, M., Heironimus, A., Woolley, M., Friedman, J., and Neimeier, A. (2020) 'Robot Transparency and Team Orientation Effects on Human-Robot Teaming'. *International Journal of Human-Computer Interaction* 36 (7), 650-660. available from <https://doi.org/10.1080/10447318.2019.1676519>

Hanoch, Y., and Vitouch, O. (2004) 'When less is more: Information, emotional arousal and the ecological reframing of the Yerkes-Dodson law'. *Theory & Psychology*, 14 (4), 427-452. available from <doi.org/10.1177%2F0959354304044918>

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://doi.org/10.1006/ijhc.2000.0454>>

Hoff, K.A., and Bashir, M. (2015) 'Trust in automation: Integrating empirical evidence on factors that influence trust'. *Human factors*, 57 (3), 407-434. available from <doi.org/10.1177%2F0018720814547570>

Hoffman, R.R., Johnson, M., Bradshaw, J.M., and Underbrink, A. (2013) 'Trust in automation'. *IEEE Intelligent Systems*, 28(1), 84-88. available from <https://doi.org/10.1109/MIS.2013.24>

Hollnagel, E., and Woods, D.D. (2005) *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. London: CRC Press

Hutchins, E. (1995) 'How a Cockpit Remembers Its Speeds'. *Cognitive Science*, 19 (3), 265-288. available from <https://doi.org/10.1207/s15516709cog1903_1>

Iwasaki, M. and Fujinami, K. (2012) 'Recognition of pointing and calling for industrial safety management'. in *Proc. 2012 First ICT International Senior Project Conference and IEEE Thailand Senior Project Contest (ICT-ISPC2012)* held 20 March to 20 April 2012 at Mahidol University.

Jipp, M., and Ackerman, P.L. (2016) 'The Impact of Higher Levels of Automation on Performance and Situation Awareness: A Function of Information-Processing Ability and Working-Memory Capacity'. *Journal of Cognitive Engineering and Decision Making*. 10 (2), 138-166. available from <doi:10.1177/1555343416637517>

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Hoffman, R. R., Jonker, C., van Riemsdijk, B., and Sierhuls, M. (2011) 'Beyond Cooperative Robotics: The Central Role of Interdependence in Coactive Design'. *IEEE Intelligent Systems.* 26 (3), 81-88. available from <<https://doi.org/10.1109/MIS.2011.47>>

Jones, D.G., (2000) 'Subjective Measures of Situation Awareness', in *Situation Awareness Analysis and Measurement*. ed. by Endsley, M.R., and Garland, D.J. Mahwah New Jersey USA: Lawrence Erlbaum Associates

Jones, D.G. (2015) 'A Practical Perspective on the Utility of Situation Awareness'. *Journal of Cognitive Engineering and Decision Making*. 9 (1), 98-100. available from <doi:10.1177/1555343414554804>

Jones, D.G., and Kaber, D.B. (2005) 'Situation Awareness Measurement and the Situation Awareness Global Assessment Technique'. in Handbook of Human Factors and Ergonomics Methods. Ed. Stanton, N., Hedge, A., Brookhuis, K., Salas, E., and Henrick, H. London: CRC Press, 42-1 – 42-8

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>

Khastgir, S., Birrell, S., Dhadyalla, G. and Jennings, P. (2018) 'Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles'. *Transportation research part C: emerging technologies*, 96, 290-303. available from <https://doi.org/10.1016/j.trc.2018.07.001>

Kitchin, J., and Baber, C. (2016) 'A comparison of shared and distributed situation awareness in teams through the use of agent-based modelling' *Theoretical Issues in Ergonomics Science* 17 (1), 8-41. available from <https://doi.org/10.1080/1463922X.2015.1106616>

Klein, G. (2000) 'Analysis of Situation Awareness from Critical Incident Reports'. in *Situation Awareness Analysis and Measurement.* ed. by Endsley, M.R., and Garland, D.J. London: Lawrence Erlbaum Associates, 45-62

Klein, G. (2015) 'Whose Fallacies?' *Journal of Cognitive Engineering and Decision Making*. 9(1), 55-58. available from <10.1177/1555343414551827>

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 <https://doi.org/10.1109/MIS.2004.74>

Lago A.S., Dias J.P., and Ferreira H.S. (2020) 'Conversational Interface for Managing Non-trivial Internet-of-Things Systems'. in Krzhizhanovskaya, V., Zavodszky, G., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Brissos, and B., Teixeira, J., (ed) *Proceedings of the 20th International Conference on Computational Science – ICCS 2020*, 'Lecture Notes on Computer Science'. held 3-5 June in Amsterdam. Cham: Springer, 384-397. available from <https://doi.org/10.1007/978-3-030-50426-7_29>

Large, D.R., Burnett, G., Anyasodo, B., and Skrypchuk, L. (2016) 'Assessing cognitive demand during natural language interactions with a digital driving assistant'. In *Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*. held 24-26 October 2016 at Ann Arbor, USA. New York: Association for Computing Machinery, 67-74. available from <doi.org/10.1145/3003715.3005408>

Large, D.R., Clark, L., Burnett, G., Harrington, K., Luton, J., Thomas, P., and Bennett, P. (2019) '"It's Small Talk, Jim, But Not As We Know It." Engendering Trust Through Human-Agent Conversation In An Autonomous, Self-Driving Car'. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. held 22-23 August 2019 at Dublin Ireland. New York: Association for Computing Machinery 1-7. available from <doi.org/10.1145/3342775.3342789>

Lashley, H., Thorpe, A., Tylor, R., and Grabham, A. (2019) Measuring effectiveness of human autonomy teaming. in *Proceedings of NATO STO Meeting on Human Autonomy Teaming (NATO STO*

HFM-300-03) held 15-17 October 2018 at Seasouth, UK. Salibsury: Dstl Knowledge and information Services. available from <u><STO-MP-HFM-300-03></u>

Lavrynenko, O., Konakhovych G., and Bakhtiiarov, D. (2016) 'Method of voice control functions of the UAV'. in *2016 4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC)*, held 18-20 October 2016 in Kiev. New York, IEEE. 47-50. available from <u><doi.org/10.1109/MSNMC.2016.7783103></u>

Lehtonen, E., Airaksinen, J., Kanerva, K., Rissanen, A., Ränninranta, R., and Åberg, V. (2017) 'Game-based situation awareness training for child and adult cyclists' *Royal Society Open Science* 4 (3) 1-14. available from <u><doi.org/10.1098/rsos.160823></u>

Lee, J.D. and See, K.A. (2004) 'Trust in automation: Designing for appropriate reliance'. *Human factors*, 46(1), 50-80. available from <u><https://doi.org/10.1518%2Fhfes.46.1.50_30392></u>

Li, H., Wickens, C.D., Sarter, N., and Sebok, A. (2014) 'Stages and Levels of Automation in Support of Space Teleoperations'. *Human Factors* 56 (6), 1050-1061. available from <<u>10.1177/0018720814522830</u>>

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. available from <u><doi.org/10.1016/j.trf.2016.09.006></u>

Lyons, J.B., and Stokes, C.K. (2012) 'Human–human reliance in the context of automation'. *Human factors*, 54 (1), 112-121. available from <u><doi.org/10.1177%2F0018720811427034></u>

Maciej, J. and Vollrath, M. (2009) 'Comparison of manual vs. speech-based interaction with in-vehicle information systems' *Accident Analysis & Prevention* 41(5), 924-930. available from <u><https://doi.org/10.1016/j.aap.2009.05.007></u>

Madhavan, P., and Wiegmann, D.A. (2007) 'Similarities and differences between human–human and human–automation trust: an integrative review'. *Theoretical Issues in Ergonomics Science* 8 (4), 277-301. available from <doi.org/10.1080/14639220500337708>

Mahajan, K., Large, D.R., Burnett, G., and Velaga, N.R. (2021) 'Exploring the effectiveness of a digital voice assistant to maintain driver alertness in partially automated vehicles. *Traffic injury prevention*, 22 (5), 378-383. available from <doi.org/10.1080/15389588.2021.1904138>

Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., and 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://doi.org/10.1016/j.obhdp.2017.08.001>

McAree, O., Aitken, J.M, and Veres, S.M. (2017) 'Towards artificial situation awareness by autonomous vehicles'. *IFAC-PapersOnLine* [online] 50 (1), 7038-7043. available from <doi.org/10.1016/j.ifacol.2017.08.1349> [1 July 2021]

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 <<doi.org/10.1177%2F0018720817743223>>

McTear, M., Callejas, Z., and Griol, D. (2016) The Conversational Interface: Talking to Smart Devices. Cham: Springer

Miller, C. A., and Parasuraman, R. (2003) 'Who's in charge?; intermediate levels of control for robots we can live with'. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. 'System Security and Assurance'.* held 8-8 October 2003 in Washington. New York, IEEE, 462-467. available from <doi.org/10.1109/ICSMC.2003.1243858>

Miller, C.A., and Parasuraman, R. (2007) 'Designing for Flexible Interaction Between Humans and Automation: Delegation Interfaces for Supervisory Control'. *Human Factors*. 49 (1), 57-75. available from <10.1518/001872007779598037>

Millot, P. (2015) 'Situation Awareness: Is the glass half empty or half full?'. *Cognition, Technology & Work* 17, 169-177. available from <https://doi.org/10.1007/s10111-015-0322-6

Minitab (2016) *Data Not Normal? Try Letting It Be, with a Nonparametric Hypothesis Test* [online] available from <https://blog.minitab.com/en/understanding-statistics/data-not-normal-try-letting-it-be-with-a-nonparametric-hypothesis-test> [19 August 2021]

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>

Murray, S., James, N., Perš, J., Mandeljc, R., and Vučković, G. (2018) 'Using a situation awareness approach to determine decision-making behaviour in squash', *Journal of Sports Sciences*, 36 (12), 1415-1422, <doi.org/10.1080/02640414.2017.1389485 >

Naderpour, M., Lu, J., and Zhang, G. (2016) 'A safety-critical decision support system evaluation using situation awareness and workload measures' Reliability Engineering & System Safety 150, 147-159. available from <https://doi.org/10.1016/j.ress.2016.01.024>

Neisser, U. (1976) *Cognition and reality: principle and implications of cognitive psychology*. San Francisco: W.H.Freeman

NHS Institute for Innovation and Improvement (2010) *Safer Care SBAR Situation Background Assessment Recommendation Implementation and Training Guide.* Coventry: NHS Institute for Innovation and Improvement

Nishimori, M., Saitoh, T., and Konishi, R. (2007) 'Voice Controlled Intelligent Wheelchair,' in *SICE Annual Conference 2007*. held 17-20 September 2007 at Kagawa University, New York: IEEE, 336-340. available from <doi.org/10.1109/SICE.2007.4421003>

O'Neil, T.A., McNeese, N.J., Barron, A., and Schelble, B. (2020) 'Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature'. Human Factors [online], available from <https://doi.org/10.1177%2F0018720820960865> [1 July 2021]

Onnasch, L., Wickens, C. D., Li, H., and Manzey, D. (2014) 'Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis'. *Human Factors* 56 (3), 476-488. available from <https://doi.org/10.1177/0018720813501549>

Operational Solutions (2021) *Introducing F.A.C.E our C-UAS & UTM C2 System* [online] available from <https://operationalsolutionsltd.co.uk/c-uas-%26-utm>

Pallant, J. (2007) *SPSS Survival Manual: A Step by Step Guide to Data Analysis using SPSS for Windows third edition.* Maidenhead: Open University Press

Parasuraman, R. (2000) 'Designing Automation for Human use: Empirical Studies and Quantitative Models'. *Ergonomics* 43 (7), 931-951. available from <<https://doi.org/10.1080/001401300409125>>

Parasuraman, R., and Riley, V. (1997) 'Humans and Automation: Use, Misuse, Disuse, Abuse'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39 (2), 230-253. available from <https://doi.org/10.1518%2F001872097778543886>

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000), "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30 (3), 286-297. available from <doi.org/10.1109/3468.844354>

Patrick, J, and James, N. (2004) 'A Task-Oriented Perspective of Situation Awareness'. in *A Cognitive Approach to Situation Awareness: Theory and Application*. ed. by Tremblay, S., and Banbury, S. London: Routledge, 61-81

Patrick, J., and Morgan, P.L. (2010) 'Approaches to understanding, analysing and developing situation awareness'. *Theoretical Issues in Ergonomics Science* 11 (1-2), 41-57. available from <https://doi.org/10.1080/14639220903009946>

Poirier, S., Routhier, F., and Campeau-Lecours, A. (2019) 'Voice Control Interface Prototype for Assistive Robots for People Living with Upper Limb Disabilities' *Proceedings of 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*. held 24-28 June 2019 at Toronto. New York: IEEE, 46-52. available from <doi: 10.1109/ICORR.2019.8779524>

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., and Mazurek, G. (2019) 'In bot we trust: A new methodology of chatbot performance measures'. *Business Horizons* 62, 785-797. available from < https://doi.org/10.1016/j.bushor.2019.08.005>

Rajaonah, B., Tricot, N., Anceaux, F., and Millot, P. (2008). 'The role of intervening variables in driver–ACC cooperation'. *International journal of human-computer studies* 66 (3), 185-197. available from <doi.org/10.1016/j.ijhcs.2007.09.002>

Richards, D. (2020) 'Measure for Measure: How Do We Assess Human Autonomy Teaming?'. in Stephanidis, C., Harris, D., Li, W-C., Schmorrow, D.D., Fidopiastis, C.M., Zaphiris, P., Ioannou, A., Fang, X., Sottilare., R.A., and Schwarz, J. (ed.) *Proceedings of the International Conference on Human-Computer Interaction,* 'Late Breaking Papers: Cognition, Learning and Games'. held 19-24 July 2020 at Copenhagen, Denmark. Cham:Springer, 227-239. available from <doi.org/10.1007/978-3-030-60128-7_18>

Rodemer, K., and Wessels, G. (2011) 'Belt-Mic for Phone and In-Vehicle Communication - Pushing Handsfree Audio Performance to the Next Level'. In Meyer, G., and Valldorf, J. (ed.) *Advanced*

Microsystems for Automotive Applications 2011. 'Smart Systems for Electric, Safe and Networked Mobility'. held 29-30 June 2011 in Berlin. Berlin: Springer, 109-117. available from <https://doi.org/10.1007/978-3-642-21381-6_11>

Rousseau, R., Tremblay, S., and Breton, R. (2004) 'Defining and Modelling Situation Awareness' in *A Cognitive Approach to Situation Awareness: Theory and Application*. ed. by Tremblay, S., and Banbury, S. London: Routledge, 3-21

Rusnock, C.F., Miller, M.E., and Bindewald, J.M. (2017). 'Observations on trust, reliance, and performance measurement in human-automation team assessment'. In: Coperich, K., Cudney, E., and Nembhard, H. (ed.) *Proceedings of the Industrial and Systems Engineering Conference*. held 20-23 May 2017, Pittsburgh, Pennsylvania. Red Hook: Curran Associates Inc, 368-373

Russell, R. (2018) 'Testing Human-Autonomy Teaming Concepts on a Global Positioning System Interface'. in: Yamamoto S., and Mori H. (ed.) *International Conference on Human Interface and the Management of Information*. 'Human Interface and the Management of Information, Information in Applications and Services'. held 15-20 July 2018 in Las Vegas. Cham: Springer, 450-464. available from <https://doi.org/10.1007/978-3-319-92046-7_38>

Salas, E., Prince, C., Baker, D.P., and Shrestha, L. (1995) 'Situation Awareness in Team Performance: Implications for Measurement and Training'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37(1):123-136. available from <doi:10.1518/001872095779049525>

Salas, E., Cooke, N.J., and Rosen, M.A. (2008) 'On teams, teamwork, and team performance: Discoveries and developments'. *Human factors* 50 (3), 540-547. available from <doi.org/10.1518%2F001872008X288457>

Salmon, P.M., and Stanton, N.A. (2013) 'Situation awareness and safety: Contribution or confusion? Situation awareness and safety editorial' *Safety Science* 56, 1-5. available from <https://doi.org/10.1016/j.ssci.2012.10.011>

Salmon, P.M., Stanton, N.A., Walker, G.H., Baber, C., Jenkins, D.P., McMaster, R., and Young, M.S. (2008) 'What really is going on? Review of situation awareness models for individuals and teams' Theoretical Issues in Ergonomics Science 9 (4), 297-323. available from <doi.org/10.1080/14639220701561775>

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. available from <doi.org/10.1016/j.ergon.2008.10.010>

Salmon, P.M., Stanton, N.A., Walker, G.H., Jenkins, D.P., and Rafferty, L. (2010) 'Is it really better to share? Distributed situation awareness and its implications for collaborative system design' *Theoretical Issues in Ergonomics Science* 11 (1-2), 58-83. available from <https://doi.org/10.1080/14639220903009953>

Sarter, N.B., and Woods, D.D. (1991) 'Situation Awareness: A Critical But Ill-Defined Phenomenon' *The International Journal of Aviation Psychology* 1 (1), 45-57. available from <doi.org/10.1207/s15327108ijap0101_4>

Schaefer, K.E., Chen, J.Y.C., Szalma, J.L., and Hancock, P.A. (2016) 'A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems'. *Human Factors* 58 (3), 377-400. available from <doi.org/10.1177%2F0018720816634228>

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>

Selkowitz, A.R., Lakhmani, S.G., and Chen, J.Y.C. (2017) 'Using agent transparency to support situation awareness of the Autonomous Squad Member'. *Cognitive Systems Research* 46, 13-25. available from < <https://doi.org/10.1016/j.cogsys.2017.02.003> >

Sharm, A., and Nazir, S. (2017) 'Distributed Situation Awareness in pilotage operations: Implications and Challenges' *TransNav, International Journal on Marine Navigation and Safety of Sea Transportation* 11 (2), 289-293. available from <dx.doi.org/10.12716/1001.11.02.11>

Sheridan, T.B., and Verplank, W.L. (1978) *Human and Computer Control of Undersea Teleoperators*. Defence Technical Information Centre [online]. available from <https://apps.dtic.mil/sti/citations/ADA057655>

Shinohara, K., Naito, H., Matsui, Y., and Hikono, M. (2013). 'The effects of "finger pointing and calling" on cognitive control processes in the task-switching paradigm.' *International Journal of Industrial Ergonomics*, 43 (2), 129-136. available from <doi.org/10.1016/j.ergon.2012.08.004>

Shively, R.J., Lachter, J., Brandt, S.L., Matessa, M., Battiste, V., and Johnson, W.W. (2017) 'Why Human-Autonomy Teaming?'. in: Baldwin C. (ed) *International Conference on Applied Human Factors and Ergonomics*, 'Advances in Neuroergonomics and Cognitive Engineering'. held 17-21 July 2017 in Los Angeles. Cham: Springer, 3-11. available from <https://doi.org/10.1007/978-3-319-60642-2_1>

Shu, Y., and Furuta, K. (2005) 'An inference method of team situation awareness based on mutual awareness' *Cognition, Technology & Work* 7, 272-287. available from <https://doi.org/10.1007/s10111-005-0012-x>

Simpson, R. C., and Levine, S. P. (2002) 'Voice control of a powered wheelchair,' *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 10 (2) 122-125. available from <doi.org/10.1109/TNSRE.2002.1031981>

Smith, K., and Hancock, P.A. (1995) 'Situation Awareness Is Adaptive, Externally Directed Consciousness'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37 (1), 137-148. available from <doi.org/10.1518%2F001872095779049444>

Society of Automotive Engineers International (2014) *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. [online] available from <https://www.sae.org/standards/content/j3016_201401/preview/)>

Sorensen, L.J., and Stanton, N.A. (2011) 'Is SA shared or distributed in team work? An exploratory study in an intelligence analysis task' *International Journal of Industrial Ergonomics* 41 (6), 677-687. available from <doi.org/10.1016/j.ergon.2011.08.001>

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

Sorensen, L.J., Stanton, N.A., and Banks, A.P. (2010) 'Back to SA school: contrasting three approaches to situation awareness in the cockpit' *Theoretical Issues in Ergonomics Science* 12 (6), 451-471. available from <doi.org/10.1080/1463922X.2010.491874>

Stanton, N.A. (2016) 'Distributed Situation Awareness' *Theoretical Issues in Ergonomics Science* 17 (1), 1-7. available from <https://doi.org/10.1080/1463922X.2015.1106615>

Stanton, N.A., Chambers, P.R.G., and Piggott, J. (2001) 'Situational awareness and safety' *Safety Science* 39 (3), 189-204. available from <doi.org/10.1016/S0925-7535(01)00010-8>

Stanton, N.A., Stewart, R., Harris, D., Houghton, R.J., Baber, C., McMaster, R., Salmon, P., Hoyle, G., Walker, G., Young, M.S., and Linsell, M. (2006) 'Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology'. *Ergonomics*, 49(12-13), 1288-1311. available from <doi.org/10.1080/00140130600612762>

Stanton, N.A., Salmon, P.M., Rafferty, L.A., Walker, G.H., Baber, C., and Jenkins, D.P. (2013) *Human Factors Methods: A Practical Guide for Engineering and Design*. Farnham, UK: Ashgate Publishing Limited

Stanton, N.A., Salmon, P.M., and Walker, G.H. (2015) 'Let the Reader Decide: A Paradigm Shift for Situation Awareness in Sociotechnical Systems'. *Journal of Cognitive Engineering and Decision Making* 9 (1), 44-50. available from <doi.org/10.1177%2F1555343414552297>

Stanton, N.A., Salmon, P.M., Walker, G.H., and Jenkins, D. (2009) 'Genotype and phenotype schemata and their role in distributed situation awareness in collaborative systems' *Theoretical Issues in Ergonomics Science* 10 (1), 43-68. available from <doi.org/10.1080/14639220802045199>

Stanton, N.A., Salmon, P.M., Walker, G.H., and Jenkins, D.P. (2010) 'Is situation awareness all in the mind?' *Theoretical Issues in Ergonomics Science* 11 (1-2), 29-40. available from <doi.org/10.1080/14639220903009938>

Stanton, N.A. Salmon, P.M., Walker. G.H., Salas, E, and Hancock, P.A. (2017) 'State-of-science: situation awareness in individuals, teams and systems' *Ergonomics* 60 (4), 449-466. available from <doi.org/10.1080/00140139.2017.1278796>

Stanton, N. A., Stewart, R., Harris, D., Houghton, R. J., Baber, C., McMaster, R., Salmon, P., Hoyle, G., Walker, G., Young, M.S., Linsell, M., Dymott R., and Green, D. (2006) 'Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology' *Ergonomics* 49(12-13), 1288-1311. available from <doi.org/10.1080/00140130600612762>

Sternberg, R.J., and Sternberg, K. (2012) *Cognitive Psychology*. 6th edn. Belmont: Wadsworth, Cengage Learning

Strayer, D.L., Cooper, J.M., Turrill, J., Coleman, J.R., and Hopman, R.J. (2016) 'Talking to your car can drive you to distraction'. Cognitive Research: Principles and Implications [online] 1, 1-17. available from <doi.org/10.1186/s41235-016-0018-3>

Strybel, T.Z., Keeler, J., Mattoon, N., Alvarez, A., Barakezyan, V., BarrazaJames, E., Park, J., Vu, K-P.L., and Battiste, V. (2017) 'Measuring the Effectiveness of Human Autonomy Teaming'. In Baldwin, C. (ed.) *International Conference on Applied Human Factors and Ergonomics* 'Advances in Neuroergonomics and Cognitive Engineering' held 17-21 July 2017 in Los Angeles. Cham: Springer, 23-33. available from <https://doi.org/10.1007/978-3-319-60642-2_3>

Strybel, T.Z., Keeler, J., Barakezyan, V., Alvarez, A., Mattoon, N., Vu, K.P.L., and Battiste, V. (2018) 'Effectiveness of human autonomy teaming in cockpit applications'. in Yamamoto, S., and Mori, H. (eds.) *Proceedings of the International Conference on Human Interface and the Management of Information* held 15-20 July 2018 at Las Vegas. Cham:Springer. 465-476. available from <doi.org/10.1007/978-3-319-92046-7_39>

Tamura, S., Iwano, K. and Furui, S. (2004) 'Improvement of audio-visual speech recognition in cars'. In *Proceedings of the 18th International Congress on Acoustics.* held 4-9 April 2004 in Kyoto. Kyoto: ICA, 2595-2598

Taylor, R. M. (1990). 'Situational awareness rating technique (SART): The development of a tool for aircrew systems design'. in *AGARD Conference Proceedings No.478 (AGARD-CP-478)* 'Situational Awareness in Aerospace Operations' Loughton: Specialised Printing Services Limited, 3-1 – 3-17

Tokadlı, G., and Dorneich, M.C. (2019) 'Interaction Paradigms: from Human-Human Teaming to Human-Autonomy Teaming'. in *Proceeding of IEEE/AIAA 38th Conference on Digital Avionics Systems Conference (DASC).* [online] held 8-12 September 2019 in San Diego. New York: IEEE, 1-8. available from < https://doi.org/10.1109/DASC43569.2019.9081665 > [1 July 2021]

UK Civil Aviation Authority (2014a), *CAP 737: Flightcrew Human Factors Handbook*. Gatwick Airport South: UK CAA. available from <http://publicapps.caa.co.uk/docs/33/CAP%20737%20DEC16.pdf>

UK Civil Aviation Authority (2014b) *CAP 744: UK Flight Information Services.* Gatwick Airport South: UK CAA. available from <https://publicapps.caa.co.uk/docs/33/20170404-CAP774_UK%20FIS_Edition%203.pdf>

UK Civil Aviation Authority (2016), *CAP 413: Radiotelephony Manual*. Gatwick Airport South: UK CAA. available from <https://publicapps.caa.co.uk/docs/33/CAP413%20MAY16.2.pdf>

U.S. Department of Transportation Federal Aviation Authority (2020) *Unmanned Aircraft Systems (UAS) Traffic Management (UTM) Concept of Operations*. [online] available from <https://www.faa.gov/uas/research_development/traffic_management/media/UTM_ConOps_v2.pdf>

Vagia, M., Transeth, A.A., and Fjerdingen, S.A. (2016) 'A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed?' Applied Ergonomics 53 (A), 190-202. available from <https://doi.org/10.1016/j.apergo.2015.09.013>

Veena, S., Rahul, D.K., Chandan, G.M., Lokesha, H., Lakshmi, P., and Durdi, V.B. (2018) 'Designing a Speech Interface for Voice activated MAV Ground Control Station'. in *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. held 18-19 May 2018 in Sri Venkateshwara College of Engineering. New York: IEEE, 1734-1737, <doi.org/10.1109/RTEICT42901.2018.9012132>

Vidulich, M.A., and Hughes, E.R. (1991) 'Testing a Subjective Metric of Situation Awareness'. *Proceedings of the Human Factors Society Annual Meeting* 35(18), 1307-1311. available from <10.1177/154193129103501812>

Walker, G.H., Stanton, N.A., Salmon, P., and Jenkins, D. (2009) 'How can we support the commander's involvement in the planning process? An exploratory study into remote and co-located command planning'. *International Journal of Industrial Ergonomics*, 39 (2), 456-464. available from <https://doi.org/10.1016/j.ergon.2008.12.003>

Waytz, A., Heafner, J., and 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://doi.org/10.1016/j.jesp.2014.01.005>

Weaver, B., and Wuensch, K.L. (2013). 'SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients'. *Behavior research methods*, 45, 880-895. available from <doi.org/10.3758/s13428-012-0289-7>

Wiegmann, D.A., Eggman, A.A., ElBardissi, A.W., Parker, S.H., and Sundt III, T.M. (2010) 'Improving cardiac surgical care: a work systems approach'. *Applied ergonomics*, 41 (5), 701-712. available from <doi.org/10.1016/j.apergo.2009.12.008>

Wickens, C.D. (2002) 'Multiple resources and performance prediction' *Theoretical Issues in Ergonomics Science* 3 (2), 159-177. available from <doi.org/10.1080/14639220210123806>

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://doi.org/10.1518%2F001872008X288394>

Wiegmann, D.A., Rich, A., and Zhang, H. (2001) 'Automated diagnostic aids: The effects of aid reliability on users' trust and reliance'. *Theoretical Issues in Ergonomics Science*, 2 (4), 352-367. available from <doi.org/10.1080/14639220110110306>

Wright, J.L., Chen, J.Y.C., and Barnes, M.J. (2018) 'Human–automation interaction for multiple robot control: the effect of varying automation assistance and individual differences on operator

performance' Ergonomics 61 (8), 1033-1045. available from

<doi.org/10.1080/00140139.2018.1441449>

Wynne, K.T., and Lyons, J.B. (2018) 'An integrative model of autonomous agent teammate-likeness' *Theoretical Issues in Ergonomics Science* 19(3), 353-374. available from

<https://doi.org/10.1080/1463922X.2016.1260181>

Appendix A. Advice On Situation Awareness Requirements For Counter UAV Surveillance

Nexus Nine Ltd
New Hockley Hall Farm
Lower Road
Hockley
Essex
SS5 5NW   Tel: 07917 391405

Ref: Your query by email

Date: 2 Dec 20

Dear Adam

I write in response to your question about the information a controller would need to know to successfully carry out the Counter UAV / UTM task you describe. In terms of discharging the task and keeping track of their progress I'd suggest that they will need to maintain situational awareness of the following:

1. **Location.** They will need to be able to remember where each of the UAVs are and what Zone they are heading towards, especially proximity to the nearest high value asset. They will need this to inform the threat estimate.

2. **Current Threat.** They will need to know which UAVs are most likely to become a threat and which are already confirmed as a threat. That means they need to know the Registration and Operations Plan information of each UAV and remember which UAV has had its risk rating increased to high risk. Specifically assessing how deviation from that plan may indicate increased threat is key.

3. **Previous Outcomes.** To plan effective responses, they will need to understand previous actions, how long they took and the consequences. It is important to be aware of how long each action takes as this directly affects the time left available for follow on decisions / reporting. They will also need to remember the success of previous responses to help identify the courses of action to consider next.

4. **Current Risks.** They will need to know what other (third party) activities are taking place in the UAVs' area of operation, normally from sources such as NOTAMs, flight plans and sensor system outputs.

As we discussed this is directly relevant to work we are doing in the BVLOS/UTM space, particularly looking at how a single operator can supervise multiple UAV missions. Therefore, I'd be grateful for sight of your final conclusions, noting, of course, the need to respect and acknowledge intellectual property rights as necessary.

I hope this response is of use. Let me know if you need any other information.

Kind Regards

CEO – Nexus Nine Ltd / lee@nexusnine.co.uk

Page (1)

Nexus Nine Ltd
Registered in England No. 10731020
Registered Office: New Hockley Hall Farm, Lower Road, Hockley, Essex, SS5 5NW T: 01608
812191 E: info@nexusnine.co.uk W: www.nexusnine.co.uk

Can We Talk? – The Impact of Conversational Interfaces on Human Autonomy Teaming Perception, Performance and Situation Awareness

Adam Bogg[1*], Andrew Parkes[1], Mike Bromfield[1]

[1]Coventry University, Coventry, United Kingdom

*bogga@uni.coventry.ac

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

Appendix B – Published Work – Publication

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

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.

This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.