Future Human-System Interaction Techniques to Influence Perceived Trust

By

Faye McCabe

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering

Department of Electronic, Electrical and Systems Engineering

The University of Birmingham

June 2023

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

The maritime defence domain presents unique challenges for the introduction of autonomous systems. With the Ministry of Defence predicting that intelligent information systems are the future of defence, recommending their integration into future systems in order to maintain a competitive advantage, the question is not “if”, but “how”.

The thesis presents a user-centred design approach to the development of automated sonar decision-support systems. It develops this approach through an understanding of Submersible Maritime Platforms (SMPs) as socio-technical systems, evaluating their informational and user requirements using context from the Trust-in-Autonomy and Human Factors literature. Requirements to produce trustworthy autonomy recommendations for the maritime defence domain are presented.

These requirements are then developed further through interviews with Subject Matter Experts, aimed to understand how they deal with informational uncertainty through the Critical Decision Method interview technique, to create a thorough understanding of the tasks involved in broadband sonar classification. This is built on through developing understanding of how they perform their cognitive classification process, in comparison to novices and other SMP crew members, using the Repertory Grid interview technique.

These literature reviews and interviews lead to the development of the VINAS: A Visual, Intelligent Narrative of Autonomous systems. This visualisation uses the cognitive constructs derived from the repertory grid to create an explanation behind an autonomous classifier’s decision-making processes, to allow an operator to evaluate its performance in a transparent and understandable way, in order to encourage appropriate trust calibration.

The VINAS visualisation is then evaluated through two experiments, which show that VINAS increases performance and trust when performing classifications utilising an autonomous classifier.

DEDICATION

This thesis is dedicated to my long-suffering parents.

Mom, Dad, I love you, thank you for being so patient with me, and for all of your help, support, and encouragement. I couldn't have got this far without you.

ACKNOWLEDGEMENTS

I am thankful for the funding I received to complete this research through the EPSRC and BAE Systems ICASE studentship.

I am also extremely grateful for the support of my supervision team in conducting this work, their expertise, camaraderie, and patience is unequalled. Thank you for your endless help, advice and support, Professor Chris Baber, and Professor Bob Stone. I am so grateful for the experiences this PhD has provided me, things I never imagined I would see or achieve, I do not know how to thank you enough for this opportunity.

A massive thank you to my best friend Tony, for his unwavering support throughout.

I would not have been able to complete this thesis without any of the people listed here. Once again, I thank you for everything you've done for me.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronym	**Meaning**
AI	Artificial Intelligence
AIS	Automatic Identification System
ANI	Artificial Narrow Intelligence

C2	Command and Control
CDM	Critical Decision Method
CO	Commanding Officer
COA	Course of Action
CPA	Closest Point of Approach
CQ	Close Quarters
DCLT	Detection, Classification, Localisation and Tracking
DEMON	Detection Envelope Modulation On Noise
DOA	Direction of Arrival
EPSRC	Engineering and Physical Sciences Research Council
FOST	Flag Officer Sea Training
HAT	Human-Autonomy Team
HCI	Human-Computer Interaction
HF	Human Factors
HFACS	Human Factors Analysis and Classification System
HMT	Human-Machine Team
ICASE	Industrial Collaborative Awards in Science and Technology
IMO	International Maritime Organisation
ISM	International Safety Management
LIME	Local Interpretable Model-agnostic Explanations
LoA	Level of Autonomy
LOB	Line of Bearing
LOFAR	Low Frequency Analysis and Recording
LOP	Local Operations Plot
MAIB	Marine Accident Investigation Branch
MO	Mission Objective
MOD	Ministry of Defence
OOW	Officer of the Watch
OPSO	Operations Officer
RN	Royal Navy
SA	Situation Awareness
SC	Sonar Controller
SMCS	Submarine Command System
SME	Subject Matter Expert
SMP	Submersible Maritime Platform
SO	Sonar Operator
SONAR	Sound, Navigation and Ranging
SPAM	Situation Present Assessment Method
ST	Sonar Trainer
TiA	Trust in Automation
TMA	Target Motion Analysis
TPK	Turns-per-Knot
UI	User Interface
VINAS	Visual, Intelligent Narrative of Autonomous Systems
VMS	Vessel Monitoring System

LIST OF APPENDICES

"To gaze into the depths of the sea is, in the imagination, like beholding the vast unknown, and from its most terrible point of view. The submarine gulf is analogous to the realm of night and dreams. There also is sleep, unconsciousness, or at least apparent unconsciousness, of creation. There in the awful silence and darkness, the rude first forms of life, phantomlike, demoniacal, pursue their horrible instincts."

- *Victor Hugo, "The Toilers of the Sea"*

CHAPTER 1: INTRODUCTION

1.1: Overview

1.1.1: Scope

The PhD research was undertaken as part of an ICASE (Industrial Collaborative Awards in Science and Technology) studentship funded by the EPSRC (Engineering and Physical Sciences Research Council) and the industrial sponsor, BAE Systems.

The research was conducted within the Human Interface Technologies Team, based in the School of Engineering at the University of Birmingham.

The focus of the research is Human Factors (HF), Human-Computer Interaction (HCI) and Human Machine Teaming (HMT), with regards to developing trust in autonomous systems (TiA) and human-autonomy interaction within the scope of the maritime defence domain.

1.1.2: Research Questions

The research questions defined for the thesis are:

> RQ1: What Level and Type of Autonomy could be suitably applicable to tasks carried out in the process of broadband sonar classification whilst maintaining appropriate levels of trust in the automation?
>
> RQ2: How can the causes of previous SMP accidents be mitigated through the introduction of autonomy?

RQ3: How do Sonar Operators cognitively classify sounds?

RQ4: How can an autonomous decision aid visually present a credible, understandable, and trustable explanation behind its decisions?

RQ5: Can a Visual, Intelligent Narrative of Autonomous Systems (VINAS) improve performance in a classification task utilising an autonomous classifier?

RQ6: Can a VINAS improve trust in an autonomous classifier when conducting a classification task?

These research questions, and how they relate to the different parts of the thesis, are explained in Section 1.4.

1.1.3: Present Work: Aims and Objectives

The overall aim of the research described in the thesis is to determine how human-autonomy teaming could be utilised in future naval defence, specifically in the underwater maritime domain, in a trustworthy, safe, useful, and advantageous way. It seeks to determine how and where human-autonomy teaming could support crew members aboard Submersible Maritime Platforms (SMPs) to conduct their work, through examining current working practices, safety concerns, and human-machine interactions. The research identifies a specific task, broadband sonar classification, and examines this task from a user-centric, Human Factors perspective, to identify how human-autonomy teaming could support operators to carry it out.

The first objective of the research is to understand how the socio-technical system of an SMP works, how crew members work together, and work with machines, to complete tasks and achieve goals. This helped to inform an understanding of the informational needs of the crew members, which could be supported by the introduction of autonomy. This understanding is developed through the Human Factors literature concerning SMPs.

The second objective is to understand how and why SMP accidents can occur from a systemic perspective. Two accidents were studied and modelled in detail, a collision that occurred in 2015,

and a near-miss which occurred in 2018. The causes of the accidents were mapped using an AcciMapping technique, to understand how they relate to each other, and how they could be prevented from happening again. This helped to identify a specific task which could be supported by autonomous information processing: broadband sonar analysis. Other informational needs are identified where information systems could be developed further to potentially support crew members manage uncertainty.

The third objective of the work is to better understand the concepts of trust and autonomy, and how they relate to each other. Trust is a complex, dynamic, multi-dimensional construct, which is affected by many intrinsic and extrinsic factors. An understanding of the different levels, degrees and types of autonomy is developed. A model of trust in autonomy is presented. A comprehensive literature review was conducted to understand the problem-space and develop a framework from which to understand how trust in autonomy can be fostered and calibrated appropriately.

The fourth objective, once a use-case for human-autonomy teaming had been identified, was to better understand the user, the Sonar Operator (SO). The task of broadband sonar classification was explored; how SOs conduct this task, how they manage and mitigate uncertainty, what their informational needs are to carry out the task successfully, and how they could be appropriately supported to carry out their work. The appropriate level, degree, and type of autonomy which could aid them in classification was defined.

The fifth objective was to develop and evaluate a visualisation using the elicited user requirements and informational needs, which could support an SO's cognitive classification method, and help to deal with the inherent informational uncertainty in the task. This was then evaluated.

1.1.4: Problem Statement

We find ourselves in a world of Artificial Narrow Intelligence (ANI); Artificial Intelligence which can outperform humans in a single-function, structured task. Since Marvin Minsky wrote "The Society of Mind" in 1983, there has been an explosion of ANI in all kinds of areas, from the virtual agents

embedded into our phones and devices, cars which can drive themselves, to the myriad uses of machine-learning algorithms and big data processing in domains as diverse as medicine, economics, logistics, entertainment, robotics, and defence. The rapid pace of development in Artificial Intelligence (AI) since the turn of the millennium represents the maturation of a field which has existed in concept for over 50 years. **However, the convergence of three key factors: powerful hardware, advanced algorithms, and vast data sets, has now made intelligent computing a reality**.

Commercial investment in AI and robotic technologies, and the recruitment of Subject Matter Experts (SMEs), dwarfs that of any state (UK Development Concepts And Doctrine Centre, 2018). Many Silicon Valley and Chinese companies spend more annually on AI and robotics research and development than the entire United States government on research and development for all mathematics, robotics, and computer science combined (Allen and Taniel, 2017).

The combination of the three key factors (hardware, advanced algorithms, and big data sets), if developed in tandem, fused, and used correctly, is set to be as revolutionary in defence as the birth of aviation, radio, or nuclear power.

There is an identified need in the Ministry of Defence (MoD) to innovate in these domains, as laid out by the Joint Concept Note 2/18: Information Advantage, and the Joint Concept Note 1/18: Human-Machine Teaming. Both of these publications identify that the research and development, and exploitation, of intelligent information systems, AI, and machine learning, is crucial in order to maintain a military advantage into the next half a century and beyond.

However, the introduction of any “bleeding edge” technology, especially in a safety-critical domain, is fraught with danger. Although it could be argued that partially autonomous and intelligent systems have been used in military technology since the Second World War, this has not been without mistakes. A recent example would be a fratricide incident which occurred at the start of the Iraq conflict in 2003, where the mis-use of the automated Patriot Missile Defence System lead to a British plane being shot down by their American allies (Hawley, 2017a, 2017b). Because of the rapid

pace of change and development of machine learning and ANI technologies, and their departure from standardised, well-understood current and past capabilities, new ways of working, interfacing, and interacting with these technologies must be researched and developed, not only to exploit their capabilities, but to do so *safely*, and in a way which *fosters trust* in their usage and outputs.

This research examines how humans could interact with ANI systems in a specific aspect of maritime defence capability, sonar analysis. It attempts to formulate an understanding of how data collection, fusion, information generation and distribution is currently conducted for this task, in order to understand how human activity can be augmented using modern information technologies.

Firstly, the thesis develops an understanding of how autonomous systems should be used and interfaced with, in order to for a user to develop an appropriately calibrated level of trust in the system, even when there are high levels of uncertainty. This is particularly pertinent with consideration to SMP operations, which are reliant on uncertain information which cannot be validated externally.

A comprehensive literature review of trust, and Human-Autonomy Teaming, is conducted, to provide a foundation of knowledge which underpins the central argument of the research.

Rather than positing a specific algorithm or classification methodology, (which would probably be out-dated before the research is even published, due to the rapid rate of innovation in the field of classifier algorithms), the thesis offers a unique methodology for identifying areas of the task which could benefit from the inclusion of autonomous information processing. Instead of centring the technologic capabilities, it instead focuses on the user, the Sonar Operator, and how to support their highly skilled work by exploiting those capabilities, using Human Factors techniques and methodologies.

The research does this in three ways. First, it analyses two recent accidents involving SMPs from a systems perspective, and tries to understand how and why these occurred, to identify where the inclusion of intelligent information processing could prevent them from re-occurring.

Second, through researching the domain of SMP information processing and communication, to identify key actors and systems, and understand how information is used to carry out tasks currently. This informs the research, to better understand how the introduction of autonomous and intelligent information systems would transform these methods, roles, teams, and processes.

Third, through eliciting user requirements and perspectives, using them to inform how the work and task of sonar classification itself works, from a naturalistic decision-making perspective. This is done through modelling and analysing the cognitive decision-making processes, heuristics, and management of uncertainty strategies, which Sonar Operators develop to perform their work, to develop user- and information- requirements, to best understand how to support them with the introduction of autonomous and intelligent information systems.

These findings are used to develop a visual aid which could support a Human-Autonomy Team working together, called a Visually Intelligent Narrative of Autonomous Systems (VINAS), and shows examples of how this could be used for sonar classification. Some evaluations of the VINAS are performed and discussed.

1.2: Approach

The thesis takes a user-centred design approach, based on the approach outlined in ISO 9241-210. ISO 9241 outlines the international standard for the ergonomics of human-system interaction, with part 210 focusing on human-centred design principles and activities related to the use of interactive systems. The standard discusses ways of enhancing human-system interaction, looking at the usage of both hardware and software aspects of interactive systems. Figure One shows the different stages of the user-centred design process outlined in the ISO.

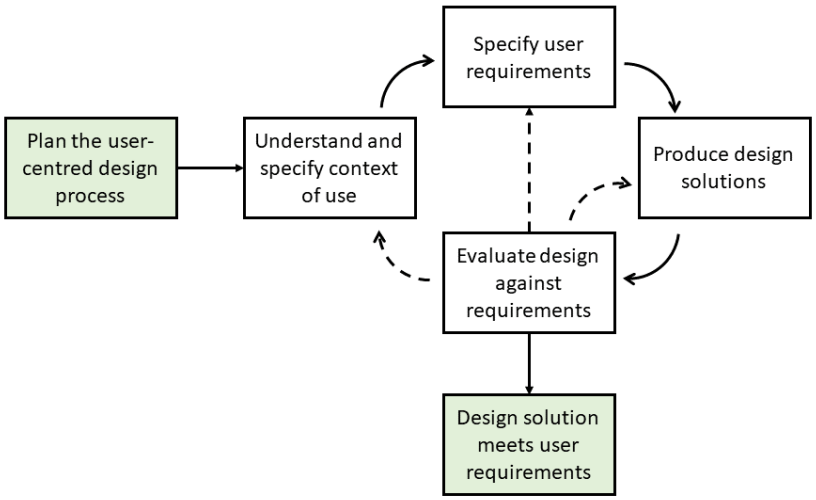


Figure 1: Diagram of the user-centred design process outlined by ISO 9241, based on ISO 9241-210:2010

As can be seen in Figure One, the user-centred design process is broken down into six stages. It is an iterative process of designing systems, with phases that re-occur and link back to each other, from understanding the context of use, to specifying user requirements, producing design solutions, and then evaluating those design solutions against the user requirements and context of use.

The ISO focuses on usability, defining this as: "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (Krippl *et al.*, 2016, p. 269).

The thesis follows this design process in a number of ways:

Firstly, by understanding the context of use through developing an understanding of the SMP as a sociotechnical system as outlined in Chapter Three. This is followed by looking at two case studies of accidents involving SMPs in Chapter Four, where the activities leading to the incidents were mapped across different layers of the sociotechnical system with a technique called AcciMapping. This helped to understand the causes of the accidents from a systemic perspective, and visualise how they related to each other. These chapters both helped to inform the specific informational requirements needed to conduct a broadband classification task, and how that information is used.

The user requirements were then explored in Chapter Five, through a Critical Decision Method (CDM) interview with an SME, helping to elicit a detailed understanding of the tasks used in broadband sonar classification, and how these are carried out by a Sonar Operator. A timeline of the tasks and events which occur in the classification process was created in conjunction with the SME, and they provided detail around each aspect.

The user requirements are then built on further in Chapter Six, through using the repertory grid interview technique with SMEs to understand their cognitive classification process. This gave insight into how the mental processes of classification could be translated into an explanation given by an autonomous classifier. Both of these chapters helped to identify key processes within the task that would benefit from autonomous support, and what kind of support would be most beneficial, centred on the user, and how they carry out these tasks. A design is presented at the end of Chapter Six.

Chapter Seven then attempts to make some initial evaluations of whether the design is useful; how it affects performance, workload, self-confidence, and a self-reported trust measure when it is used in a classification task.

1.3 Background

The Ministry of Defence posit that in the next fifty years, the integration of artificially intelligent and autonomous systems, which exploit information gathering, processing and fusion technologies and techniques, will be key to maintaining a military advantage (UK Development Concepts And Doctrine Centre, 2018; UK MOD, 2018).

The use of automation will offer opportunities to better exploit information, which will improve understanding, decision-making, and tempo. They state that efforts should focus on automation information collection, processing, and management, and that, “bespoke AI may be required for specific applications, such as automating the analysis of visual and audio data flows” (UK Development Concepts And Doctrine Centre, 2018, p. 16).

Automated systems can make decisions rapidly, faster than humans can monitor and restrain them (UK Development Concepts And Doctrine Centre, 2018). The proliferation of sensors and machine learning systems outperforming humans at recognition and pattern detection tasks is likely to increase.

For all of the expected benefits, the “chaos and friction” which they could cause must somehow be effectively managed. As these systems become integrated into more systems, the probability for unexpected interactions that rapidly spiral out of control will increase.

An example is given by (Allen and Taniel, 2017) of the stock market “Flash Crash” of May 2010, where one trillion dollars of stock market value was wiped out within minutes, because of unintended machine interactions. The U.S. Securities and Exchange Commission reported that this was “enabled and exacerbated by use of autonomous financial trading systems”. A small trader’s spoofing algorithm caused banks’ automated trading systems to enter an online loop, which crashed the stock market in under 36 minutes.

When autonomous and ANI systems get things wrong, they can have disastrous consequences. AI can be fooled, and even have biases, depending on how it is trained: systems are not infallible. Algorithmic decision-making can be flawed, and procedural consistency is not equivalent to objectivity.

Some examples of this can be seen in the recent emergence of advanced driver assistance technologies, sometimes referred to as “self driving” vehicles. The NHTSA, or National Highway Traffic Safety Administration, a federal agency in the USA, reported how Tesla’s “Autopilot” driver assistance system, known as “Full Self Driving”, has been involved in 35 crashes which lead to the deaths of 19 people since its release in 2021 (Klippenstein, 2023).

Although the mis-use of these systems can be attributed as a contributing factor for many of the crashes, with the driver not having their hands on the steering wheel, or driving whilst intoxicated,

the autopilot feature has been known to slam on the brakes of vehicles when moving at high speed (Siddiqui, F. and Merrill, 2022), known as “phantom breaking”, creating dangerous conditions. In another example of an eight-car collision where a child was injured, the car drove itself into a lane of on-coming traffic and then came to a stop (Press and Estachio, 2022).

These incidents highlight a number of points about the introduction of autonomy. Firstly, that people have biases towards autonomy where they think it is more capable and intelligent than it really is, hence the over-reliance on the self-driving features, even though the features are specifically designed to be used in conjunction with an alert driver with their hands on the steering wheel. Secondly, that autonomy is fallible; driving into on-coming traffic should never be an action carried out by a self-driving car, but when conditions are less-than-optimal, autonomy can make mistakes, which can have disastrous or dangerous consequences. Thirdly, the behaviour of such systems when used in reality can be very different from the behaviour of systems in optimal laboratory conditions, and reliance on them, especially at high speeds, in safety critical situations, needs to be carefully managed and evaluated against the potential safety risks.

Visual and audio data can be manipulated in ways too subtle for humans to be able to perceive and can be used to fool ANI. Part of ANI's vulnerability is its lack of real intelligence, making it possible to be tricked by deception.

Automated systems are also very "brittle" (Gutzwiller and Reeder, 2020); they struggle to function in situations outside of their design parameters, and can fail catastrophically when faced with unique situations outside of their trained range. Even if they can recognise when they are reaching their limit of capability, they lack the contextual thinking to intuit when it is appropriate to "hand over" to a human operator, meaning they can demand human attention at points of high workload and stress, potentially passing on a problem to a human who is insufficiently engaged, with no opportunity to understand the issue and avert disaster. Pilots are still needed despite robust modern

autopilot systems, for example, as the human-in-the-loop remains essential for the rare occasions when autopilot can no longer cope.

Humans are also capable of introducing problems to Human-Autonomy Teams (HATs). Human attention is limited, neither constant nor consistent. Passive monitoring is difficult for humans and holds their attention poorly. An unengaged human may not hold efficient Situation Awareness (SA) to suddenly orient themselves rapidly at a point of crisis (Parasuraman, Sheridan and Wickens, 2008). Highly technical skills must be used frequently to prevent them from degrading in quality; removing a human-in-the-loop prevents their ability to engage sufficiently when required. Even when including low-level autonomy, as with current self-driving vehicles, a lack of timely intervention from the driver was recorded for all fatal crashes previous to 2019 (Jenssen *et al.*, 2019).

Therefore, although there is an operational need to incorporate these new technologies into the defence domain, to maintain a tactical and defensive advantage, there is also an operational need to manage them, and interactions with them, safely, to mitigate the potential for them to create danger.

This is why the study of Human-Machine Teaming (HMT) and Human-Autonomy teaming is important. Developing HATs which augment human strengths, is crucial to their success. Not only that, but a human operator must be able to trust the information generated and displayed by autonomous systems, and be able to rapidly evaluate its accuracy, in order to use it in safety-critical situations.

Humans and machines have different strengths and weaknesses; one cannot simply replace another. They also have different ways of analysing a problem and making decisions, which may not be easily communicable or understandable. In order to be effective, they must work interdependently on problems, augmenting each other's strengths, and mitigating each other's weaknesses. Ways for

them to communicate effectively with each other, share SA, and calibrate appropriate levels of trust must be exploited.

Therefore, advantages will not automatically be provided by implementing the newest, or most expensive, algorithm; instead, it lies with the most effective HMT.

Through the exploration of these issues, the thesis hopes to be able to answer RQ1: “What level and type of autonomy could be suitably applicable to tasks carried out in the process of broadband sonar classification whilst maintaining appropriate levels of trust in the automation?” This understanding of how to create trusting interfaces is developed through a literature review which explores what level and type of automation could be appropriate, and how to interface with it, in order to develop an appropriate level of trust.

1.3.1: Submersible Maritime Platform Operations

The maritime domain presents new and unique challenges for HAT development. Below periscope depth, data transfer from the outside world is practically non-existent. External views of the environment are completely mediated by sensor systems, and so gaining SA, building a tactical picture and an understanding of an operational environment relies on complex socio-technical systems of data processing, fusion, and display, and information analysis, communication, and assimilation, between human and computer actors. This could be considered an example of distributed cognition.

Distributed cognition is "characterised by multiple individuals and teams working together in pursuit of a common goal (comprising multiple interacting sub-goals)" (Hagberg, 1997; Hutchins, 2005; Stanton, 2014a). High levels of communication and coordination are vital, with technology often facilitating this.

When operating covertly, very little information is available for the team aboard an SMP to develop understanding about their position and environment. They rely on passive SOund NAvigation and

Ranging, or sonar, sensor data to understand the environment, and what vessels, or contacts, are present in the surrounding area. Sonar "is a system for the location and ranging of objects using sound propagation and listening" (Fay, Stanton and Roberts, 2019). The four main functions of sonar are Detection, Classification, Localisation and Tracking (DCLT) (Hughes *et al.*, 2010).

A very high-level description is provided here, but a more detailed description can be found in Chapter Three. SOs use passive sonar systems to listen to noise radiated by vessels (ships or SMPs) using hydrophone arrays. To detect a vessel, an operator must distinguish the sound it makes from oceanic background noise, or sight it on their waterfall display forming a line. An estimated Direction of Arrival (DOA) is made from the acquired signals, in order to inform the presence of a target in a determined direction, known as the bearing. Analysis is then performed using DEMON (Detection Envelope Modulation On Noise) display, and/or LOFAR (Low Frequency Analysis and Recording). Both rely on spectral frequency estimation, and support the detection and classification of targets. This allows operators to derive the speed of a contact, and some of its engine characteristics, such as number of propellers.

Tracking of a vessel is done by analysing broadband trends over time, to try and determine actions a vessel is taking. For every time period that a sonar array returns data, it is plotted as a line. When a line is added, it moves all of the others down the display, giving a waterfall-like effect (Asplin and Christenson, 1988). When an SO detects a vessel, they assign it an identifier, allowing the sonar system to automatically track and update its location.

Once the SO has an estimated speed for the contact, they pass the data cuts and speed on to the Sonar Controller (SC), who manages all SOs in the sound room. The SC acts as an informational filter between the sound room and control room. The SC then passes this information over to be used for Target Motion Analysis (TMA). TMA is the process of "analysing positional data from contacts derived from passive sensors to produce a location and predicted movements" (M. Murphy, 2000). This is a "solution", comprised of speed, course, range and bearing for a contact (Genç, 2010). This is

a line of best fit, as all information is derived from estimations. The more data cuts for the vessel, the more accurate the TMA solution can be. Once a TMA solution has been derived that seems to fit the contact, it can then be uploaded to a geographical display, which can be used by the Command team.

Command teams have three main objectives: remain safe, remain undetected, and complete mission objectives (Stanton, Roberts and Fay, 2017). Generally, the Officer of the Watch (OOW) leads the command team, and is responsible for making safe navigational decisions. The command team is in the control room, separate from the sound room, where SOs work with the sonar. Control rooms are "nerve centres", where trained operators and advanced technology is utilised, to understand the environment, and develop Courses of Action (CoA), to meet operational and strategic goals. (Stanton and Bessell, 2014; Stanton *et al.*, 2017). The control room contains multiple operators working as a team to communicate different information, utilising different sensors to generate knowledge of the environment, and to complete mission objectives (Ly, Huf and Henley, no date).

The OOW does not have their own display; instead, they can view repeater screens of different displays. They have to assimilate information from the multiple screens, whilst communicating with the wider team, and make mental calculations and plan actions, to gain SA. SA is distributed across many system agents, both social and technical. Each actor, whether technology, or operator, contributes to the distributed situation awareness. The OOW must use this collective SA to create a mental, tactical picture, in order to make decisions. The command team works together, constantly updating the tactical picture under direction from the OOW (Dominguez, Long, Miller, Wiggins, *et al.*, 2006).

From this high-level description of operations, it can be seen that an SMP is an extremely complicated socio-technical system, with dynamic information sharing occurring constantly and

rapidly between different rooms, information systems, and team members, to try and build an accurate tactical and navigational picture.

There is uncertainty inherent throughout this system, and if mistakes are made, they can be propagated and negatively impact the distributed SA and tactical picture. Sonar classification and TMA are integral to environmental understanding, even though they are mostly based on estimations. If an error is made during contact classification, it can have dangerous consequences, resulting in a lack of understanding of where a contact is in relation to own-ship, which can result in collisions (Washington, no date; Marine Accident Investigation Branch, 2015).

This leads to the formation of RQ2: "How can the causes of previous SMP accidents be mitigated through the introduction of autonomy?" The thesis identifies why previous accidents and near misses occurred, and how autonomy could be used in order to prevent them from happening again.

Interfaces for broadband classification have changed very little over time (Fay, Stanton and Roberts, 2019). Despite many capability advances, User Interfaces (UIs) for sonar data still consist of green and black waterfall displays. This can be attributed to reducing risk, SME familiarity, or maintaining training readiness (Hall, 2012). However, as the technologies and capabilities advance, it is sensible to presume displays may need to be re-designed, in order to incorporate these new technologies effectively, and improve the SA of the Operators.

SMPs present an interesting paradigm for the consideration of the introduction of autonomous systems - their trust in technology must be high, as it mediates all understanding; however, new systems can introduce new risks, and new uncertainties, into an already perilous environment. Highly skilled operators could benefit from innovative new interfaces, for example, despite representing a 360° aural signal, the sonar waterfall display is not circular; this requires operators to mentally translate the plot of their surroundings, which could be increasing their cognitive work; however, in such a safety-critical, highly dynamic and uncertain environment, making even small changes to ways of working could have consequences which affect all parts of the system. If, as the

Ministry of Defence posits, the future of defence lies in the incorporation of autonomous information systems, this could drastically affect all aspects of the socio-technical system, including methods of work, roles, and crew configurations.

This leads to research question RQ3: “How do Sonar Operators cognitively classify sounds?” By understanding how subject matter experts mentally deal with this information, and come up with classifications, it may be possible to identify how best to support them with these cognitive tasks through the provision of new, autonomous tools and interfaces. This then leads to RQ4: “How can an autonomous decision aid visually present a credible, understandable, and trustable explanation behind its decisions?”, which examines how to best present this autonomously-derived information in order for trust to be calibrated correctly.

This research attempts to understand how SMP crews of the future could benefit from the inclusion of autonomous information systems in such a way that they can be trusted appropriately, and will positively contribute to safety, and distributed situation awareness. Future autonomous support and human-machine teaming for such a complex domain takes careful consideration.

1.4: Research Questions

This work hopes to answer the following research questions:

> *RQ1: What Level and type of Autonomy could be suitably applicable to tasks carried out in the process of broadband sonar classification whilst maintaining appropriate levels of trust in the automation?*

This question is answered by:

- Thoroughly examining the literature in order to understand how to select an appropriate Level of Autonomy (LoA) which could be applied to various tasks conducted in the process of broadband sonar classification

- Using the literature review to understand where different types of autonomy are appropriate, and which types could be applicable to the process of sonar classification

- Interviewing a Sonar Operator, and using this interview to identify key tasks and uncertainties on the classification process, to understand how autonomy could be applied to them

As explored in the literature view, there are many different levels (Onnasch *et al.*, 2014a) and types (Parasuraman, Thomas B Sheridan and Wickens, 2000a) of autonomy. Combined in two dimensions, these create the concept of degrees of automation. The tasks elicited from the CDM interview with a Sonar Operator are then evaluated to understand what type and degree of autonomy would be most appropriate for various tasks.

RQ2: How can the causes of previous SMP accidents be mitigated through the introduction of autonomy?

The research answers this question by analysing two recent accidents involving Royal Navy vessels, AcciMapping them, and evaluating where the addition of autonomous systems could help prevent them from happening again.

RQ3: How do Sonar Operators cognitively classify sounds?

RQ4: How can an autonomous decision aid visually present a credible, understandable, and trustable explanation behind its decisions?

The research answers these questions by conducting a study involving the repertory grid interview technique to try and elicit the cognitive classification process from a Sonar Operator. This leads to the development of the VINAS display.

RQ5: Can a VINAS improve performance in a classification task utilising an autonomous classifier?

RQ6: Can a VINAS improve trust in an autonomous classifier when conducting a classification task?

The research answers these questions by evaluating the VINAS grid in two experiments, measuring the trust and performance of participants when conducting a task with, and without, a VINAS, and when information is congruent, and incongruent.

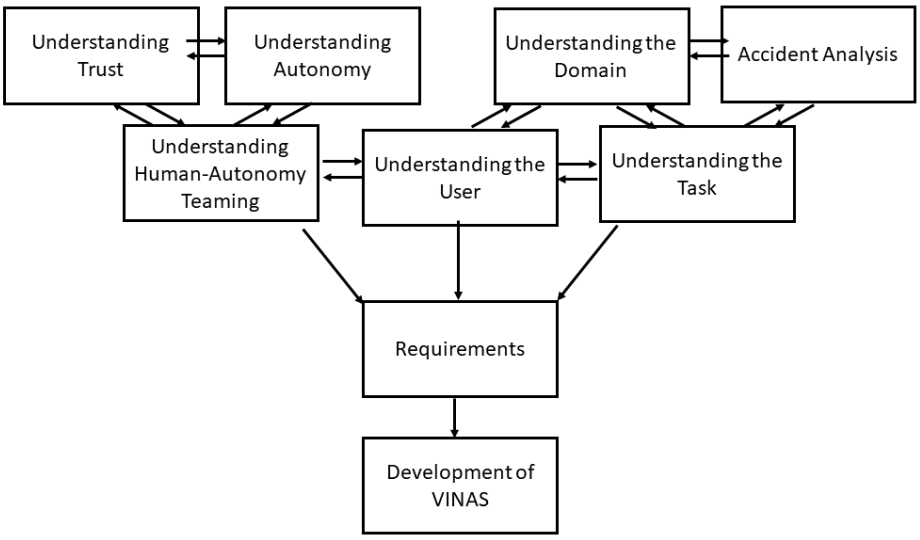


Figure 2: How the sections of the thesis relate to each other

In Figure Two, the different objectives of the thesis can be seen, and how they relate to each other. Developing an understanding of trust, and autonomy, informs an understanding of human-autonomy teaming, and how autonomy can be interfaced with effectively in order to develop trustworthy human-machine teams.

By developing an understanding of the domain, through considering a submersible platform as a socio-technical system, and examining the specific task of broadband sonar classification, an understanding of the task and its information requirements is formed. Looking at case studies of accidents allows an understanding to be developed of what happens when things go wrong, and what the causes can be.

All of these parts give context to understanding the Sonar Operator, their role, their tasks, and their requirements. This is substantiated further by interviews with an SME who has worked in this role,

to understand better how they carry out their tasks and how they deal with the information uncertainty inherent to that task.

Understanding human-autonomy teaming, understanding the user, and understanding the task, allows requirements to be developed which inform how, why, and in what form, autonomy could be applied to the task. These requirements then underpinned the VINAS display development. This is reflected in the user-centred design process outlined in Figure One, with an understanding of the user, and understanding the tasks and their context of use both contributing to user requirements, and an informed design which meets those user requirements.

1.5: Thesis Structure

An outline of the thesis structure is as follows:

Chapter Two presents a comprehensive review of the trust in autonomy literature and defines a theoretical framework from which to understand trust, and trust in autonomy. It presents a model for understanding trust in autonomy, and identifies key factors which affect it, which must be exploited within an interface design to ensure appropriate trust development, mediation, and calibration. Strategies for managing uncertainty and creating interdependent human-machine teams, with shared situation awareness, are presented and discussed. Examples of experimental interfaces for Human-Autonomy teams in the defence domain, which have been developed to foster trust and interdependence, are discussed.

Chapter Three explains key actors, technologies, and roles within an SMP, what they do, and how they share and communicate information. It seeks to describe how the overall socio-technical system operates, how humans and machines function within that system to operate safely and develop a distributed understanding of their environment. It seeks to show how the tactical picture is generated and maintained, which is used to carry out their mission objectives.

Chapter Four presents the accident analyses for two incidents which involved Royal Navy SMPs. Visualisations of the causes of the incidents are presented, created from accident reports compiled by the Marine Accident Investigation Branch (MAIB). The reasons behind the incidents are discussed, and informational requirements which could be exploited through the use of intelligent information systems are identified. The role of sonar information in the accidents is contrasted.

Chapter Five explores the broadband sonar classification task, and the role of the Sonar Operator. It seeks to understand how Sonar Operators carry out the classification process, and their strategies for mitigating uncertainty. It presents the analysis of a Critical Decision Method interview conducted with an SME, and shows a high-level timeline for the task which was produced during this. The level, degree, and type of autonomy which could be used to support the task is defined.

Chapter Six builds on Chapter Five by looking at the cognitive processes, constructs and concepts which are used in contact classification. It presents the results of experimentation which elicits these cognitive concepts and constructs through the use of a repertory grid interview technique. The methodology for the experimentation is presented. A visualisation to help explain an autonomous classifier's decision-making is presented, called the VINAS. A VINAS is produced for both a Sonar Operator and an Officer of the Watch, showing distinct cognitive constructs derived from each. This is explained through their differing roles and information requirements.

Chapter Seven presents the results of two experiments used to evaluate the efficacy of the VINAS grid. Experiment One evaluates how trust, performance, perceived workload, and confidence are affected by the inclusion of a VINAS in a simulated contact classification task, utilising a simulated autonomous classifier. Experiment Two tests how trust, performance, workload, and confidence are affected when there is incongruent information presented to make a classification decision using a VINAS and a simulated autonomous classifier. Results are presented and discussed.

Chapter Eight concludes the work, and presents a summary of what the thesis has explored. Limitations of the current research, and ideas for further work, are discussed.

1.6: Published Works

A list of works published pertaining to the research:

- Ergonomics and Human Factors 2020 - Classifying Vessels Using Broadband Sonar: Considerations for Future Autonomous Support - Full conference paper, available at: https://publications.ergonomics.org.uk/publications/classifying-vessels-using-broadband-sonar-considerations-for-future-autonomous-support.html
- International Conference on Multi-Modal Interaction 2021 - Feature Perception in Broadband Sonar Analysis – Using the Repertory Grid to Elicit Interface Designs to Support Human-Autonomy Teaming - Full conference paper, available at: https://dl.acm.org/doi/10.1145/3462244.3479918

CHAPTER 2: WHAT IS TRUST? A LITERATURE REVIEW

2.1: What is Trust?

To understand trust in automation, and how it is affected by HSI techniques, first a concept of trust must be defined.

Trust has been studied in the context of many different disciplines and fields of research including business, psychology, sociology, economics, decision-making, robotics, and neuropsychology. This multidisciplinary perspective has created confusion about what trust can be conceptualised and defined as, with many different schools of thought about what trust is. As well as this, trust is a "hypothetical construct", unable to be directly observed or measured in any physical sense. It could be considered to be an "intervening" variable, residing in the human mind, and mediating a person's observable responses to environmental stimuli in a similar way to mental workload. This inability to measure trust directly makes it difficult to describe its very nature (Muir, 1994).

Early literature seeks to define trust as a fundamental ingredient inherent to a "healthy personality", as described by Erikson in 1953. It is implied in the field of psychoanalysis in the earlier twentieth century that our capacity and ability to trust is borne out of early childhood, and the level of security one is provided in their inter-personal and familial relationships as a baby (Lunsky, 1966).

A failure to trust others was cited as an "important determinant in delinquency" in the 1950s, with little regard as to the qualities, or trustworthiness, of the trustee. Instead, trust is considered to be an individualist trait tied to the formation of personality, and beginning with the development of sufficient confidence in the goodness of the mother figure (Erikson, 1950). This is then extrapolated to affect an individual's trust within society as a whole, and their ability to form trust both interpersonally, and organisationally. This "basic trust" or "proto-trust" can impact the basic development and ability to learn of an individual, as education involves a reliance of trusting the information provided by a person without directly being able to observe the evidence for it.

Rotter developed this idea of intrinsic, personal trust further under the idea of social learning theory. Instead of seeing one's capacity to trust being fixed from inter-personal relationships in formative years, social learning theory expands on this by claiming behaviour is determined more generally by "expectancy"; the expected outcome of a behaviour, and the value a person places in the outcome of that behaviour (Rotter, 1954). This affects all social learning, and social relationships, but instead of being set in early childhood, it is fluid, and based on previous experience. Therefore, trust is seen to be reactive, in the sense that experience will affect our ability and propensity to trust, and intrinsic, in the sense that a personal capacity to trust is developed through the internalisation of one's experiences and expectations. Experiences provide positive or negative re-enforcement for expectations. This led to the development of Rotter's Interpersonal Trust Scale (Rotter, 1967).

In the paper about the development of this scale, Rotter describes interpersonal trust as "an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon". This touches upon two important concepts which are echoed through many different models of trust, even in the modern day; expectancy, and reliance.

Rotter does little to separate capacity to trust from something (or someone's) trustworthiness. There is also a conflation between a person's inherent propensity to trust, and the action of trusting itself.

(Athos, Gabarro and Holtz, 1978) explored the multi-dimensional nature of trust, through clinical interviews. They were defined to include:

1) integrity, honesty, and truthfulness
2) competence, technical and interpersonal knowledge, and skills required to do one's job
3) consistency, reliability, predictability, good judgement in handling situations
4) loyalty or benevolent motives, willingness to protect and save face for a person
5) openness or mental accessibility, willingness to share ideas and information freely.

Many of these concepts propagate through the more modern trust literature, and also through the literature concerning trust in automation, especially the ideas of competency, reliability, and openness. However, this was seen through a lens focused on power dynamics between subordinates and superiors, positing that depending on the dimensions differed in importance depending on the relative status of an individual. For example, the work suggested that "the integrity, loyalty, and openness of one's superiors are more important the superiors' competence and consistency".

Butler and Cantrell (Butler and Cantrell, 1984) tested some of these hypotheses further, with participants responding to cues describing hypothetical superiors and subordinates. They found that many of them were not supported. They did show some support for the idea that "there was no difference between the importance of one's subordinates and the importance of the integrity of one's superiors" (Butler and Cantrell, 1984).

The focus on these power dynamics in both Rotter's and Gabarro's work was shown to be less important than first expected; trust is less about submission or dominance and can be seen in relationships which are non-hierarchical in nature. Again, aspects of honesty, integrity and dependability are seen to be more crucial to trusting relationships than implied subordination or superiority. Therefore, trust was shown to be multi-faceted, but also based on more than a person's position within society. A dyadic relationship, involving some give and take between trustor and trustee, had been established.

Rousseau (Rousseau *et al.*, 1998) performed a seminal review, looking at how trust was researched across a wide variety of disciplines in an attempt to compare and quantify what the fundamental properties of trust were in an organisational context. This work brought together a plethora of definitions and studies, and concluded on a number of definitive characteristics of organisational trust.

Rousseau defined trust in terms of *vulnerability*, identifying this as an intrinsic characteristic across many different fields of research. Trust fundamentally includes a willingness to be vulnerable, and

also, a willingness to take a risk; without these aspects of vulnerability and risk, there is certainty, or belief, but no real trust. However, trust is more than simply risk-taking or vulnerability, it is a psychological state which mediates those actions, rather than an action in itself. People can be pressured to take risks, or take risks under duress, which is distinctly different from an action which comes from trust.

The second characteristic Rousseau finds across the literature is *interdependence*. This is defined as "where the interests of one party cannot be achieved without reliance upon another" (Rousseau *et al.*, 1998). The degree of interdependence modifies the levels of risk and trust. Therefore the overall definition of trust was determined to be, "the willingness to be vulnerable under conditions of risk and interdependence."

2.1.2: Mayer's Organisational Model of Trust

Mayer's integrative model of trust (Mayer *et al.*, 1995) is one of the most referenced and widely used models to understand trust, and like Rousseau's review, Mayer sets out to create clarity from the many different definitions of trust found in the various schools of research. Mayer shows a distinction between trust's antecedents, processes, context, and the products of trust; it describes six primary components of trust, but only one component is trust itself. In this way, it manages to incorporate many different facets of trust that were identified in the earlier literature, and explain how they co-exist and relate to each other.

This model has become foundational to understanding trust in automation, and is widely used as the basis of well-accepted models of this (see (Lee and See, 2004; Hoff and Bashir, 2015)). It also manages to make the distinction between pre-dispositional trust, something individual, stable over time, and static; with the fluid, dynamic and reactive trust that changes depending on the level of risk-taking and interdependence, as both identified in previous works.

This section will attempt to explain the different components of Mayer's model, so that the different components of trust can be discussed in a cohesive manner throughout the rest of the literature

review, and also so that the chosen models of focus for trust in autonomy can be shown to be logically and tangibly related to an established model of trust as a whole. It also has influenced the chosen measures used to understand trust in later experiments.

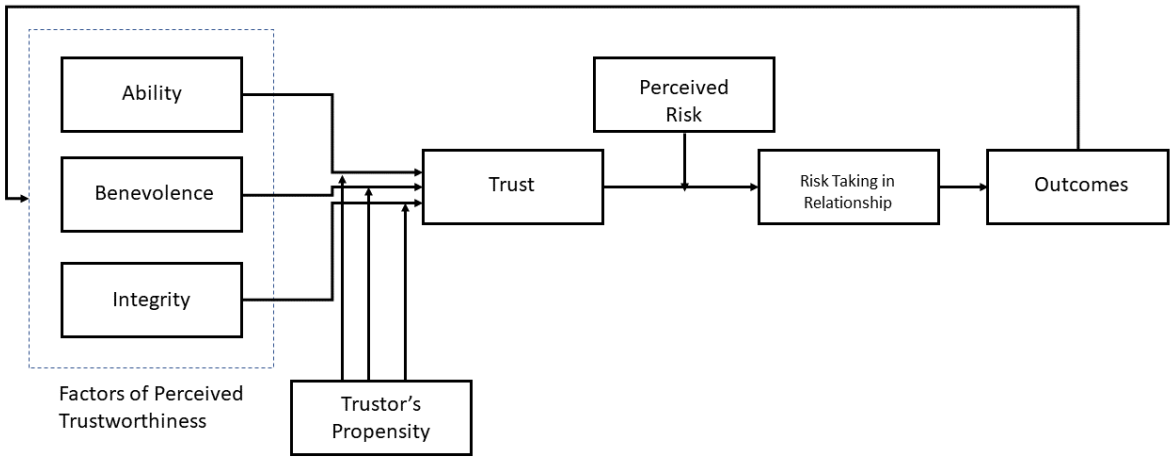


Figure 3: Mayer's Proposed Model of Trust. Adapted from Mayer et al., 1995

Figure Three shows a diagram of Mayer's model. Each component has been separated, and their relationship to each other established. Trust can be seen to be influenced by two distinct concepts: Factors of Perceived Trustworthiness, and a Trustor's Propensity.

The three Factors of Perceived Trustworthiness are ability, benevolence, and integrity. Integrity is defined as "the degree to which the trustee adheres to a set of principles which are acceptable to the trustor".

Ability is defined as "a group of skills, competencies and characteristics that enable a party to have influence within some specific domain". The domain is specified because skills and competencies in one area, for example, a technical area, may not necessarily be present in another area, for example, interpersonal relationships. Therefore affordance of trust in one area may not carry over to another, and so this shows how trust is domain-specific.

Benevolence is defined as "the extent to which a trustee is believed to want to do good to the trustor". Benevolence is chosen distinctly from intentions, or motives, because they can represent

something selfish or competitive, such as profit-seeking, or something which can be motivational to be dishonest or deceptive for one's own gain. Benevolence implies a relationship for its own sake, with no extrinsic reward for the trustee. Therefore a trustee has a positive perception towards the trustor. Because of the inherent risk-taking present in trust, these Perceived Factors of Trustworthiness provide encouragement and promote willingness to take risks.

The second key factor influencing trust is what is called Trustor's Propensity. Propensity relates back to the earlier concepts of dispositional trust identified in the works of Erikson and Rotter. Propensity to trust is considered to be a stable trait which influences a person's perceptions of ability, integrity, and benevolence, or trustworthiness, as a whole. Propensity is considered to be innate, and is influenced by an individual's culture, personality, and experiences. It impacts whether a trustor will trust before they have shared any experience with a trustee. It refers to long-term tendencies which can arise from both biological and environmental influences (Hoff and Bashir, 2015). Hoff and Bashir identify four primary sources in this foundational trust layer; culture, age, gender, and personality (Hoff and Bashir, 2015).

Trust has been shown to vary across countries, races, religions, and generational cohorts (Garske, 1976; Doney, Cannon and Mullen, 1998; Yoo, Donthu and Lenartowicz, 2011; Chien *et al.*, 2017, 2018, 2020; Hillesheim *et al.*, 2017). These influences establish a pre-conception of a trustee's trustworthiness (Kohn *et al.*, 2021). This predisposition to trust explains how some people can engage in "blind trust", whilst others may choose not to trust even when there is ample evidence to do so.

As a trustor accumulates experience with a trustee, propensity begins to play less of a role in the trusting relationship (Mayer *et al.*, 1995), (Merritt and Ilgen, 2008). The inter-relationship between propensity to trust, and perceptions of trustworthiness, lead to the development of trust in the diagram above.

This trust then mediates the perception of risk. Mayer makes the distinction between a willingness to trust, and the "behavioural manifestation" of the willingness to be vulnerable, and to take a risk, i.e.: behavioural trust. "Trust is the willingness to assume risk; behavioural trust is the *assuming* of risk" (Mayer *et al.*, 1995). Risk can be seen as a situational modifier (Kohn *et al.*, 2021), in the sense that individuals are less likely to engage in trusting behaviour when they have a high perception of risk (Lyons and Guznov, 2019). Trustors therefore weigh up their trust attitude with their perception of a situational risk. If trust outweighs the risk, they will engage in an expression of trust and increasingly take a risk (Mayer *et al.*, 1995; Colquitt, Scott and LePine, 2007; Solberg *et al.*, 2022). The trustor will rely on the trustee, making themselves vulnerable in order to support and meet their goals.

The outcome of this behaviour, whether positive or negative, becomes a feedback loop which influences future trust attitude (Mayer *et al.*, 1995; Kohn *et al.*, 2021). This influences the perceptions and evaluation of integrity, ability, and benevolence, as can be seen by the arrow feeding back to the start of the diagram.

2.2: Trust in Automation

Now that an understanding of the components and processes affecting trust and trust behaviours has been established, it can now be applied to the central focus of the thesis, trust in automation.

Automation can be defined as "technology that actively selects data, transforms information, makes decisions, or controls processes" (Lee and See, 2004, p. 50). Bradshaw defined autonomy as, "an idealised characterisation of observed or anticipated interactions between the machine, the work to be accomplished, and the situation" (Bradshaw *et al.*, 2013, pp. 4, 5). Both of these definitions encompass different aspects of autonomy; it is technology that deliberately acts, and an interaction between technology, situation, and work. This complex conceptualisation is similar to the conceptualisation of trust as an all-encompassing feeling, belief, mediator, and trigger for action.

The study of Trust in Automation is borne from the literature on supervisory control. As automation began to proliferate and become increasingly pervasive, researchers tried to understand and analyse the changing relationship between humans and machines in complex technical systems. Sheridan modelled supervisor behaviour throughout the seventies and eighties, advancing the hypothesis that supervisors' intervention behaviour could be based upon their trust in automation.

Sheridan and Hennessey observed in 1984 that "supervisory control demands that the system be trustworthy" (Sheridan and Hennessy, 1984). This implies that by their very existence, automated systems must be trustworthy, otherwise they would not be implemented in control systems, and therefore trust in those systems is implicit in the very act of supervisory control.

Sheridan and Hennessey see trust as the deciding factor in overriding autonomous systems; when trust in the system falls, and a supervisor no longer believes automation can control a process safely and effectively, they will choose to override the automation and take manual control of a process or system (Sheridan and Hennessy, 1984).

They also posit that the perception of trustworthiness of a system is what governs operator behaviour, rather than actual trustworthiness, which is a common theme to this day in TiA literature. They determine two dimensions of trust which affect supervisory behaviour: the predictability of the consequences of a system's actions, and their desirability (Sheridan and Hennessy, 1984).

Muir attempted to model this relationship between supervisory control and trust in 1985, but this work remained unpublished until 1994. She posits that trust develops in machines in a "developmental sequence", with stages of predictability, dependability, followed by faith, once the relationship and experience with the system matures (Muir, 1994).

She identified that in order to understand the predictability of a machine, its behaviour must be observable, and so a system must be transparent. Once predictability can be established, through extensive experience with a system which is observable, to the extent where uncertain and risky

scenarios can be observed, beyond the normal operational parameters of the automation, an operator is able to evaluate its dependability. Therefore, it can be seen that the basic components of trust, vulnerability, and risk, mediate the ability to trust an autonomous system to maintain control over its functions without manual intervention.

Muir then posits that faith in automation must be based upon the development of an understanding of its predictability and dependability because of an acknowledgement of its complexity; processes under supervisory control being so complex, that they "defy complete understanding" (Sheridan and Hennessy, 1984; Muir, 1994). Faith is especially pertinent when autonomy is exposed to novel situations, as belief in the expected behaviour must stretch beyond any observed or available evidence.

Lee and Moray build on Muir's model of trust in machines by identifying how her "developmental sequence", proposed as orthogonal to key components of trust, are actually complementary to them. Instead, Lee and Moray posit four dimensions to trust. The first being the "persistence of natural laws". The second is identified as performance, depending on "the expectation of consistent, stable and desirable performance or behaviour." The third is process, built on an understanding of the characteristics which govern behaviour. Their final dimension is labelled purpose, reflecting the intentions of the designer in creating a system (Lee and Moray, 1992).

As Mayer established the three factors of perceived trustworthiness as ability, benevolence, and integrity, TiA literature has established factors of perceived trustworthiness in automation which directly relate back to Mayer's model.

Lee and See relate the "three Ps"; process, purpose, and performance, back to Mayer's Factors of Perceived Trustworthiness. As performance refers to the operation of the automation, and its characteristics such as reliability, and predictability, it directly relates to Mayer's factor of ability; defining not only what the automation does, but its competency and expertise "as demonstrated by its ability to achieve the operator's goals" (Lee and See, 2004).

Process is defined as the appropriateness of the automation's algorithms for a given situation. In other words, process describes how the automation operates. This corresponds to Mayer's idea of integrity, or how consistently actions adhere to a set of acceptable principles (Lee and See, 2004). Considering process, trust is held in the agent, and not in its specific actions. Therefore the process basis of trust is reliant on inferences which can be drawn from the performance of the agent. It focuses on how the algorithms and operations by which it achieves its goals can be easily understood by the operator and can be assessed as capable at achieving them, implying a dispositional understanding of the automation.

Purpose is shown to refer to "the degree to which the automation is being used within the realm of the designer's intent". Therefore, it describes why the automation was developed. This is posited to correspond to benevolence, as it "reflects the perception that the trustee has a positive intention towards the trustor" (Lee and See, 2004). Purpose looks at why the automation is developed and whether the desired outcomes of the human and the automation are aligned between the two. This relates to the idea of value congruence in human-human trust, where an assessment is made on the intentions and motivations of the trustee. In human-automation trust, this depends on whether an operator is able to perceive the designer's intent, allowing the operator to trust the automation to achieve the goals which it was designed to achieve.

The three Ps of a system being presented in a transparent and accessible way are crucial for an Operator to develop appropriate levels of trust in a system. Lee and See (2004) define bases of trust that are strengthened or degraded through observation and reflection on how an autonomous system realistically performs, evaluating whether the process by which it achieves the operator's goals is appropriate, and whether the autonomy truly reflects the designer's intent, or purpose (E. Chancey *et al.*, 2017, p. 335). These levels of trust are dynamic, transient, reactive, and are heavily influenced not only by knowledge, observation and familiarity with the system, but also through social and cultural identity, personality, personal and cognitive bias, psychology, mood, cognitive

load, environmental factors and past experiences (Lee and See, 2004) . The three dimensions identify three separate types of goal-oriented information which contribute to the development of an appropriate level of trust.

Observation of the automation's performance will support inferences the user makes regarding the internal mechanisms associated with the automation's process, as observations of the processes can support understanding of the designer's intent, and therefore, the autonomy's purpose. Having a clear understanding of the autonomy's purpose will allow a user to evaluate its performance, and so the three dimensions also depend on each other.

Systems that focus on making the purpose, process, and performance of the autonomy as transparent as possible to the operator will encourage an appropriate calibration of trust in the automation, as their observations and inferences regarding the automation are aligned. However, inconsistency in what can be inferred will lead to poor coherence (Lee and See, 2004).This is similar to human-human trust, as observed by Gabarro, who saw that when there was incongruency between the perceived intentions conveyed by a manager (their purpose) and their actions (their performance), this would heavily negatively impact trust in them (Athos, Gabarro and Holtz, 1978). Therefore, to generate stable and robust trust, it must be based on all three factors equally, as discrepancies between the coherence or the observation of the three elements undermine trust.

This also affects the design of training and interfaces for automation. Making the three Ps clear to the operator can enhance the appropriateness of trust. However, the mere availability of information does not guarantee it will be interpreted by the operator, and so automation must be designed and presented in a way which is consistent with the cognitive processes which are foundational for trust development.

Lee and See split these cognitive processes into three categories: analytic, analogical and affective, with all three being assimilated in different ways. These depend on the evolution of the relationship between the trustee and trustor, what information is available to the trustor, and how it is

displayed. Just like with the process, purpose, and performance, the three influence each other. Affective processes can influence both analytic and analogical processes, more-so than analytic processes can influence affect.

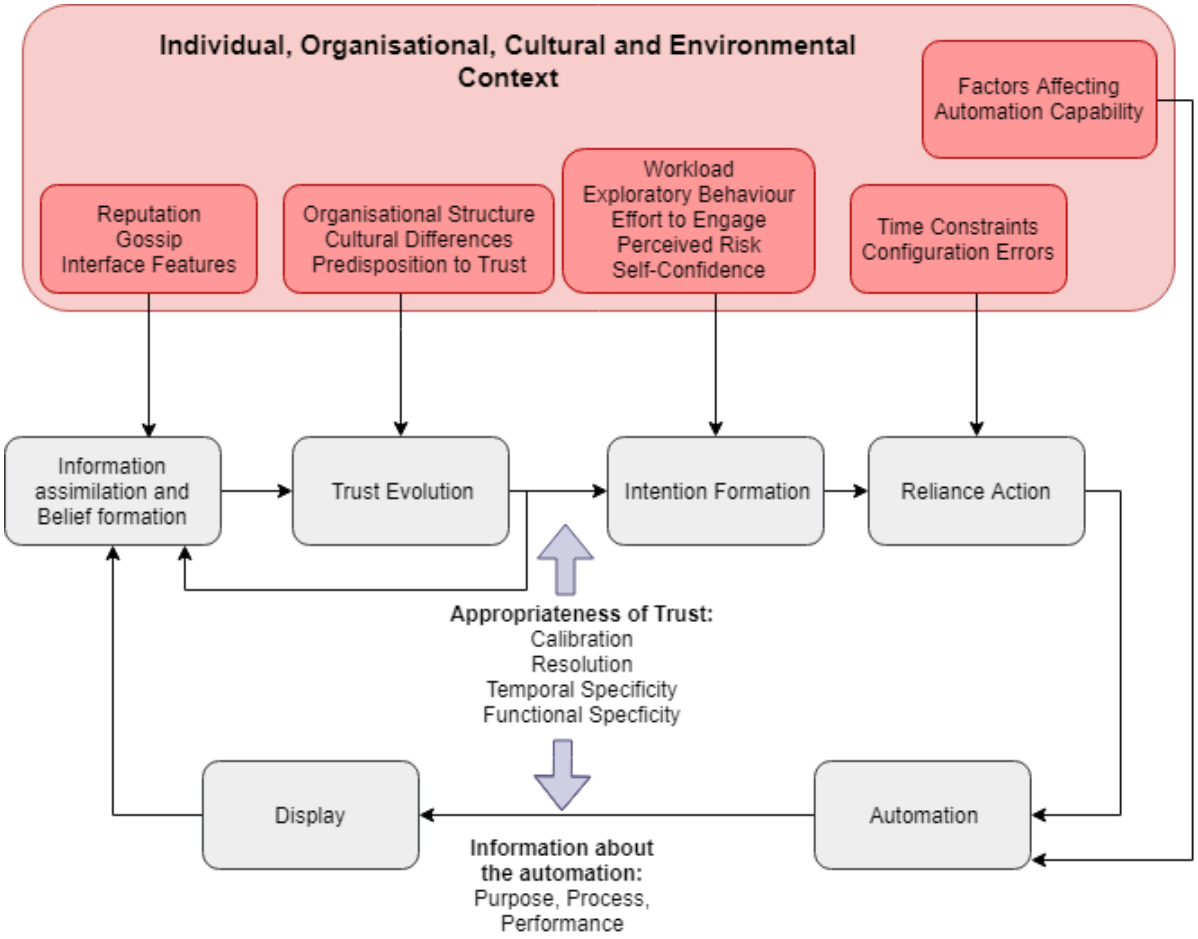


Figure 4: Lee and See's conceptual model of the dynamic process which governs trust and its effect on reliance. Adapted from Lee and See (2004)

In Figure Four, a conceptual model of trust in autonomy and the processes which affect it are shown. It can be seen that the process of developing trust in autonomy is a loop, with many influences. How trustworthy a system is does not necessarily affect how much the user trusts it; instead, their perception of a system, and perceptions of automation and technology in general, mediate how much trust is generated.

The diagram shows the process of moving from forming a belief, to developing trust, which leads to forming an intention, and then a **trust-based action** of compliance or reliance. Three critical

elements are the closed-loop dynamics of trust and reliance, the mediating effect context has on trust and reliance, and the role of the information display on developing appropriate levels of trust.

As information is gathered, trust evolves, but it is affected by many different organisational, cultural, and environmental contexts, such as the user's predisposition for trust, the features of the interface, how confident the user feels, and constraints. Many of these contextual factors also influence the automation's performance as well; for example, environmental variability, such as weather conditions, or a history of inadequate maintenance, could also degrade the performance of automation, rendering reliance inappropriate (Lee and See, 2004).

Trust combines with other attitudes, such as perception of workload, level of engagement, perceived risk, and self-confidence, to form the intention of whether to rely on the automation. Once the intention is formed, factors such as time constraints and situational familiarity then affect whether the actor relies on the automation.

Trust and its effect on behaviour and action can be seen as a dynamic interaction between the operator, context, automation, and interface (Jian, Bisantz and Drury, 2000). This feedback loop influences any interaction with the automation. If the system is under-trusted, it is unlikely to be used, and if it is not used, then an operator will have limited information and experience by which to understand its capabilities, making it difficult for trust to increase. As an integral part of trust in autonomy is through the observation of its behaviour, automation must be relied upon in order for trust to grow (Muir and Moray, 1996).

Human-automation trust and interpersonal trust are dependent on different attributes (Hoff and Bashir, 2015). Whereas interpersonal trust begins with little more than predictions of a trustee's actions, until knowledge of their dependability or predictability is known, evolving eventually into faith, in the trustee's benevolence (Lee and See, 2004), human-automation trust develops in the reverse order.

As shown, people often exhibit a positive bias towards novel automated systems (Dzindolet *et al.*, 2003), assuming that they operate perfectly, which is essentially faith-based. This trust rapidly degrades when exposed to system error. As the relationship with the autonomy progresses, dependability and predictability replace faith as the primary basis for trust in the system (Madhavan, Wiegmann and Lacson, 2006). Although there are some related concepts, as expressed above in the synthesis of Lee and See's model of trust in autonomy, interpersonal trust directly differs from trust in technology in two distinct ways.

Firstly, technology lacks intentionality, unlike humans. Whereas interpersonal trust has a basis in the altruism of the trustee, or their intentions for completing the same task as the trustor, automation is based on system capabilities, scripts, and algorithms for a specified use case (Hopko and Mehta, 2021). Autonomy does not truly have individual intent, unlike humans (Madhavan and Wiegmann, 2007). It can be argued that trust in autonomy is actually trust in the designer's intent once-removed, and its design may be reflective of the designer's intents and biases. As automation in safety-critical systems is assumed to be designed with the improvement of the system in mind, operators can assume it is intended to work in support of them. The reciprocal, dyadic nature of interpersonal trust does not apply to trust in automation. Instead, the perceived capability, or purpose, of the autonomous system, is considered the primary basis for trust in automation (Chen *et al.*, 2018).

The other major difference between interpersonal trust and trust in automation is the lack of anthropomorphisation of many autonomous systems, and therefore, the lack of the accompanying social expectations (Hopko and Mehta, 2021). Users like to personify technology (Nass and Moon, 2000a). Anthropomorphised qualities can increase trust in autonomous systems (Nass and Moon, 2000b; Oleson *et al.*, 2011; de Visser, Pak and Shaw, 2018a; Zhang and Yang, 2022), however, there is a point where this can result in an extreme degradation of trust levels, known as the uncanny valley (Mori, MacDorman and Kageki, 2012; Lay *et al.*, 2016; Latoschik *et al.*, 2017). Ascribing human

qualities to autonomous systems can have a harsher negative impact on trust when they fail to behave in the same way as people do, and can be interpreted as deceptive, especially implemented in safety-critical systems.

2.2.1: Table of User Requirements

Table One summarises the user requirements identified from the research thus far.

Table 1: A summary of user requirements elicited from the trust in autonomy literature

User Requirement Identified	**Reasoning**
A system must strive to be transparent (Muir, 1994)	Through transparency, a system's behaviours become observable. Familiarity can be developed. Allows for the user to evaluate dependability and predictability, both crucial for trust development
A system must be predictable (Muir, 1994)	By creating predictable systems, a user is able to anticipate the system's actions, and so can become familiar with a system's behaviour. This also enables them to know when a system is not behaving correctly, so they can spot when something goes wrong
The process behind a system must be transparent (Lee and See, 2004)	By understanding the processes behind a system, a user is able to identify how it comes to conclusions, and so can assess their efficacy and understand when they are correct or incorrect
The purpose of a system must be transparent (Lee and See, 2004)	By understanding the purpose behind a system, a user is able to identify whether it is working for their benefit, a key factor in developing trust in the system's intentions
The performance of a system must be	By understanding the performance of a system, a user can then judge if it is working effectively, and develop familiarity with the way it works, which is crucial for correct trust calibration

transparent (Lee and See, 2004)	
The three Ps must be presented in a way the user can interpret and understand (Lee and See, 2004)	Information about the three Ps must be easily understandable for a user to be able to analyse them effectively. Observability mediates the level of reliance a user will display

2.2.2: The Compliance-Reliance Paradigm

Reliance on automation must be correctly calibrated. Over- and under- reliance on automation can result in poor performance, or even disaster. Reliance and compliance can be seen to represent two different types of trusting actions, or responses, towards automation, where the reliability of each can affect trust differently, with trust being the attitude which influences an operator's actions through reliance, or compliance.

Compliance can be defined as a salient response to a signal issued by a system. An operator refraining from a response when there is no signal, implying normal operation, is reliance (Rice and Geels, 2010; E. T. Chancey *et al.*, 2017). Together, reliance and compliance create dependence on a system.

The reliability of the automaton impacts the responses of an operator in different ways. If a system is prone to giving false alarms, the rate of compliance will be degraded, as an operator may choose to ignore the alerts of a system. If a system is prone to misses, this can degrade reliance (Dixon and Wickens, 2006), as an operator cannot trust that a system has not failed to alert them to something important.

The effects these types of errors have on the trust-based actions of reliance and compliance differ in the literature. Meyer proposes that false alarms and misses affect compliance and reliance

individually (Meyer, 2001). Dixon and Wickens used a UAV paradigm to investigate this and found that automation false alarms negatively affected both compliance and reliance (Dixon and Wickens, 2006). This was corroborated in (Dixon, Wickens and McCarley, 2007), whereby false alarm prone automation was shown to hurt overall performance more than miss-prone automation.

This effect can be explained by a number of reasons. System reliability has a mediating effect on these actions. If a system is quite unreliable at alerting to a problem, this may draw an operator's attention to the information separately to the automation, and so they may be able to spot a miss without reliance on the automation. This relies on an awareness of the system's unreliability, which can be framed as an understanding of the system's performance.

However, if a system has high reliability, this can lead to complacency in an operator, who will assume good performance, even when there is a system failure, and so the chances of spotting a miss are decreased (Rovira, McGarry and Parasuraman, 2007; E. T. Chancey *et al.*, 2017).

A system which gives an alarm presents a salient, explicit choice for an operator to comply. However, intervention without a signal is more difficult, as an operator may lack awareness of the unreliability of the system. Therefore, as trust develops from the observation of a system (Lee and See, 2004), there is a large penalty to trust when the system alerts the operator to its unreliability, which can affect both compliance and reliance (E. T. Chancey *et al.*, 2017).

In Rice (2009), participants performed a simulated combat task by examining aerial photographs for the presence of enemy targets. A diagnostic aid provided recommendations during each trial. By manipulating the reliability and response bias of the aid, Rice showed that both false alarms and near misses can affect both reliance and compliance actions, therefore demonstrating a multiple-process theory of operator trust. False alarm rates were shown to have strong selective effects on operator compliance, and weaker nonselective effects on operator reliance. Miss-prone automation was shown to have a strong selective effect on operator reliance, and weaker nonselective effects on operator compliance (Rice, 2009).

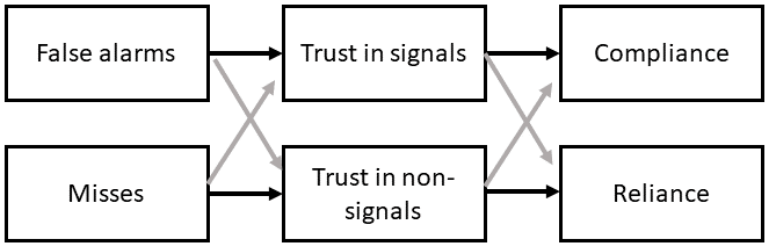


Figure 5: Rice's multiple-process theory of operator trust. Weaker relationships are shown in grey. Adapted from Rice (2009)

This shows a need for system designers to consider carefully which type of error is more critical. If a miss cannot be afforded, as in a safety-critical system, the automation bias must be adjusted accordingly. However, if a system which is sensitive and produces too many false alarms could lead to the "cry-wolf" effect (Breznitz, 1983) whereby the automation could be turned off or ignored when critical, the bias must be adjusted in the opposite direction (Rice, 2009).

2.2.3: Automation Bias and Complacency

Complacency and automation bias represent different manifestations of automation mis-use, with considerable overlap (Parasuraman and Manzey, 2010a).

People tend to have mis-conceived, or biased, expectations of autonomous performance. Some of the "myths of automation" (Bradshaw *et al.*, 2013) are that autonomy unilaterally reduces human workload, is self-sufficient, performs better than a human, and will eventually replace humans, obliviating the need for human-machine collaboration.

The perceived benefits in performance and reduced workload are true; to an extent. Onnasch et al. show that when operating at perfect levels, automation does have these benefits to performance and workload (Onnasch *et al.*, 2014b). However, when it fails, the consequences can be severe, with humans not able to quickly identify the automation failure, and so being unable to prevent or mitigate it suitably.

Even when automation has the ability to be self-sufficient, a machine is incapable of taking responsibility for its own actions, and so to limit risk, humans will still need to take some kind of responsibility, or manual control, over the autonomy. Therefore there is a significant interaction between the responsibility for outcomes, and the delegation of authority; ethical and moral considerations which govern how, and to what extent, automation can be used. This is especially pertinent when considering the domain of defence; autonomous team-mates cannot take accountability or be punished for their actions – if automation fails, who will take responsibility for the consequences?

This can also lead to situations of under-reliance, where the capabilities of automation are purposefully limited because of a fear of the consequences if they do not achieve the expected result. The other extreme of this is over-trusting automation. Over-trust of automation can lead to complacency, and accidents, such as the grounding of the cruise ship *Royal Majesty*, caused by a failure of an automatic radar plotting aid due to a broken cable, and the loss of GPS signal, which was not noticed by the crew, who did not monitor the other sources of navigational information sufficiently, showing an over-reliance on the ARPA system (NTSB, 1997; Parasuraman and Manzey, 2010b).

Humans have a propensity to believe that automation performs more rationally and objectively than a human (Dijkstra, Liebrand and Timminga, 1998; de Visser, Pak and Shaw, 2018b). Trust can be rapidly degraded when a system performs erroneously, more-so with human-machine interaction than with human-human interaction (de Visser *et al.*, 2016). This could be due to the expectancy of better performance from automation.

There are a number of "ironies of automation" (Bainbridge, 1983); the more powerful and complex automation becomes, and the more it is integrated into important or safety-critical systems, the more important a human-in-the-loop becomes, to ensure that the automation is operating correctly. However, if automation takes over a complex task which requires highly skilled operation, and little

input from a human, human skills and performance can degrade over time and with dis-use, resulting in a decreased ability to intervene if a system fails. This can be the case for manual control, or for the application of expert knowledge, both of which can be lost if not practiced frequently. Humans are also ill-suited to monitoring systems for long periods of time, especially when they require little interaction from the human, leading to inattention, complacency, and reduced situational awareness.

One of the ideal purposes of automation is to reduce human workload to increase performance of a task, by freeing up human attention and delegating work to an autonomous system. The irony here is that because of a human having responsibility for both themselves and the autonomy, this supervisory role can actually increase workload, as the human now has to monitor the automation's activities as well as attend to their own work. The introduction of autonomy to a task not only introduces new supervisory tasks, but also requires an operator to know how to use the autonomy efficiently. This introduces additional costs on the operator to understand how the autonomy works and what it is doing, in order to be able to evaluate if it is behaving correctly.

The brittle nature of automation is another problem with its integration into safety-critical systems. Automation has no capacity to deal with unique situations which it has not been trained on how to perform, which can lead to unexpected and poor performance when faced with novel situations.

Adding or expanding the role of automation profoundly changes the human's role in the system, and their interactions with it (Woods and Dekker, 2000). The addition of an automotive aide to a task can be considered equivalent to the introduction of a new team-mate, with new co-ordination costs introduced (Christoffersen and Woods, 2000), requiring them to ensure their actions are synchronised and consistent.

Complacency has been implicated as a contributing factor in aviation accidents. Wiener reported more than half of 100 experienced airline captains stated complacency was a leading factor in accidents. Parasuraman (Parasuraman, Molloy and Singh, 1993) showed that complacency was

inversely related to automation consistency, with mean detection rate of automation failure being markedly higher for the variable-reliability condition compared to the constant reliability condition. They also found that detection of automation failure was significantly higher when that was their sole task. This showed complacency as an active reallocation of attention in cases of high workload, rather than a passive state.

Singh, Molloy and Parasuraman studied the spatial positioning of automation and its effect on complacency. When the automation was centrally located, there was similar performance to (Parasuraman, Molloy and Singh, 1993), therefore it was not affected by a more centralised positioning of the automation.

Molloy and Parasuraman (Molloy and Parasuraman, 1996) performed a similar study to (Parasuraman, Molloy and Singh, 1993), but this time made the automation more subtly unreliable, with only a single occasion of failure, and tested whether early or late failure would make a difference to detection rate. They found that in the singular task condition, where participants only monitored the automation, they mostly did spot the automation error, whether it occurred early or late. However, for the multi-task condition, they observed that only half of the participants spotted the singular failure, and even less if it occurred later.

These studies show that automation complacency occurs even for highly reliable systems. As signal detection decreases with reductions in signal probability, this indicates that monitoring for automation failure will often result in poor performance. This complacency cost can off-set benefits automation can provide, especially for safety-critical systems, where misses could have grave consequences.

Wickens and Dixon (Wickens and Dixon, 2007) found that once automation reliability falls below 70%, the benefits of automation are the same as if there was no automation. They found that people would still use this automation if it was presented to them, however. Yet De Visser and Parasuraman (De Visser and Parasuraman, 2007) found that even below the 70% reliability

condition, automation could still support human operators if they have the raw information sources, which operators can combine with automation output to improve overall performance.

These ideas of over-trust and complacency are significant and problematic when considering the introduction of autonomy to the maritime defence domain. They show that for tasks where misses could have dangerous consequences, even for very high reliability automation, incorporating such autonomous systems could have dangerous consequences, as they would require a high level of supervision and vigilance from human operators, who could easily miss a failure in performance.

Instead, humans and automation should work together interdependently on a task, with a shared information space and common ground, meaning opportunities for complacency are minimised, and even low-reliability automation can still help with the achievement of common goals.

This can also help to form more resilient trust with automation, as it can also help to mitigate the effects of automation bias, as described above. By sharing common ground and working together, the purpose and processes of a system are clearer to an operator, helping them to understand the actual capabilities of the automation, and calibrating trust more accordingly. Keeping humans in-the-loop and sharing work interdependently prevents inattention from long periods of supervision, and maintains engagement, meaning reduced complacency. Thus, interdependent Human-Machine Teams (HMT), rather than rigid, allocated roles for humans and automation, create a more productive and trusting working relationship.

2.3: Human-Autonomy Teaming

2.3.1: Types of Autonomy

Appropriate allocation of system functions within a Human-Autonomy team is a vitally important question, as introducing autonomy transforms human activity, and potentially imposes new coordination demands on the human operator (Parasuraman, Sheridan and Wickens, 2000; Onnasch *et al.*, 2014b). Introducing autonomy into a system needs careful consideration; high levels of

autonomy can lead to problems with complacency, loss of situational awareness, and manual skill decay (Endsley and Kaber, 1999; Onnasch *et al.*, 2014c) as the operator may over-rely on, or over-trust the system, and so does not actively monitor the system as they expect it to perform well.

It can also reduce the human operator to a supervisory, passive role, and even if the autonomy is highly accurate, this is ill-suited to human attention, and can have heavy cost-benefit penalties if, or rather when, it fails. Most automation is not perfectly reliable. Automation can fail because of software or hardware failures, or because it is used for something outside of its functionality (Onnasch *et al.*, 2014a). This, therefore, creates a trade-off between how useful automation is when performing at optimum capacity, and the negative consequences that can occur if the automation fails.

This shows a need to maintain "directability": allowing the prevention or modification of any action in a timely manner, which in turn requires a system to provide anticipatory indicators and explanation which can be used to quickly predict and validate its decisions (Johnson *et al.*, 2012; Defense Science Board, 2016). The benefits of this are two-fold, as offering transparency, where the purpose, processes and performance strategies of automation is clear, also provides a strong foundation for developing a resilient and trusting interdependence with the autonomy (Lee and See, 2004; Chancey *et al.*, 2017b).

Parasuraman et al. (2000) identified four primary types of automated systems, including information acquisition, information analysis, decision selection, and action implementation (Parasuraman, Sheridan and Wickens, 2000). These are not mutually exclusive, and vary depending on the level of control the human operator has over their function. This adds another dimension to the "level" of autonomy, as the autonomy could operate to different degrees depending on the type of action it is performing.

Onnasch (Onnasch *et al.*, 2014a) visualises this model, by plotting three levels of automation, manual, low, and high, against the automation actions as defined by (Parasuraman, Sheridan and

Wickens, 2000), and termed this “degrees of automation” (DoA). This model fits nicely with the idea of directability; the correct level of automation could be selected depending on what kind of automation type would be appropriate.

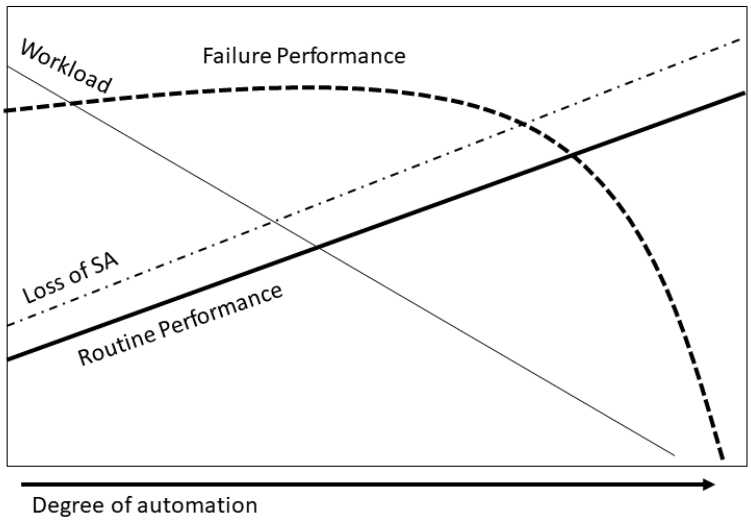


Figure 6: The trade-offs between loss of SA, workload, and failure performance depending on the degree of automation and its reliability. Adapted from Onnasch (2014)

Figure Six illustrates the trade-offs between high levels of automation in terms of workload, SA, and performance, known as the “lumberjack analogy”: “the taller the tree, the harder it falls” (Wickens and Dixon, 2007). Routine performance increases when automation is high, but results in extreme failure when automation fails to work correctly. Loss of situation awareness is directly correlated with performance. Workload is negatively correlated with performance; it decreases as the automation becomes more independent.

Table 2: Parasuraman's Automation Actions, with examples for high and low levels

Automation Type	**Type Description**
Information Acquisition	Low: Mechanically move sensors, find and grasp an object, Medium: Organise incoming information, highlight information High: Filter out information
Information Analysis	Low: Extrapolate or project, show projection

	High: Integrate inputs into a single value, augment perception and cognition, provide context dependent summaries
Decision Selection	Low: Recommend action High: Execute course of action
Action Implementation	Replace a manual function of a human

Table Two shows a summary of examples of different types of automation, as described by (Parasuraman, Sheridan and Wickens, 2000).

They are modelled loosely on the four-stage model of human-information processing, which describes how humans first acquire and register multiple sources of information, which refers to pre-perception processing of sensory data we receive. The second stage would be "conscious perception", and the manipulation of retrieved information from working memory. This can include cognitive operations such as rehearsal, integration, and inference, occurring prior to the point of decision. The next stage of human-information processing is where decisions are made based on such cognitive processing, the fourth stage would be implementation. This is a simplistic interpretation of human cognition, but it can be seen how the different stages of autonomous action are closely coupled to it.

2.3.1.1: Acquisition Automation

The first level of automation action, acquisition, is related to supporting human sensory processes. At a low level, it could consist of mechanically moving sensors to scan and observe. Parasuraman uses the example of radars in air traffic control scanning the sky in a fixed pattern, or a robot using sensors to allow it to find and grasp an object.

A moderate level acquisition system will organise incoming information for an operator, maybe prioritising information in a list, or highlighting parts of the information. Highlighting, rather than

selecting, is an important distinction; highlighting does not remove any raw data, and effectively still lets an operator see the full reality of the picture, as does ranking information. A higher level of automation acquisition would filter data, therefore removing the ability of the operator to perceive it.

2.3.1.2: Analysis Automation

Analysis automation involves cognitive functions such as inferential processes and working memory (Parasuraman, Sheridan and Wickens, 2000). At a low level, automation could extrapolate or predict what happens to data over time. An example of this is a projected display used in a cockpit, showing the predicted path of another plane in the airspace (Chandra *et al.*, 2009). Another example from aviation would be a Converging Runway Display Aid, where the approach path of an aircraft onto a converging runway could be visualised, reducing cognitive load by visualising what the operator would have to otherwise mentally compute and project. A more complex form would be an "information manager", which summarises data to a user in a context-dependent way. In this way, information integration tries to augment current operator perceptions.

Rovira found that information automation had less of a cost on performance than decision automation in a simulated C2 (Command-and-Control) task when the reliability of the automation was 80%. Decision automation had the biggest performance benefit, however, when reliability was higher. They posit that when automation is reliable yet imperfect, performance is better with an information support tool, as the user will still generate their own courses of action, and so is not so detrimentally influenced by inaccurate information. However, when the automation was only 60% reliable for the information automation, performance was worse, showing sensitivity to the overall automation imperfection (Rovira, McGarry and Parasuraman, 2007).

2.3.1.3: Decision Automation

Here, automation can select from decision alternatives. This could augment or replace human decision selections with machine decision selections. They make implicit or explicit assumptions

about costs and values of possible outcomes. Examples include route planning for pilots, diagnostic tools in medicine, and C2, such as the algorithm FOX-GA described in (Schlabach, Hayes and Goldberg, 1999), which generates and evaluated plans for manoeuvres. The LoAs for this action encompass recommendation up to execution of an action.

2.3.1.4: Action Automation

Parasuraman describes this stage as "involving different levels of machine execution of the choice of action, typically replacing the hand or voice of the human". For example, a photocopier sorting, collating or stapling (Parasuraman, Sheridan and Wickens, 2000). A more complex version would be virtual agents, tracking user interactions and executing subtasks automatically, such as described in (Lewis, 1998, pp. 67–78) with regard to Norman's Human-Computer Interaction model (Norman, 1984).

Rieger and Manzey conducted an experiment manipulating time pressure and reliability of an autonomous decision support system which was used to aid a luggage screening task. They also manipulated whether the participant made a decision first manually, and then was shown an autonomous suggestion, or first saw the autonomous suggestion, and then made a decision. They found that under high time pressure, reliance and performance actually decreased, but only when shown the autonomous suggestion first. When they made a decision manually first, their performance and reliance increased comparatively (Rieger and Manzey, 2022). Performance was worse with human intervention than it would have been with no human intervention, with participants making increased incorrect rejections of the decision support system's suggestions. This was especially true for the high-performance condition.

Bartlett and McCarley also showed this effect in a different experiment, where participants used automation to identify if a pattern on the screen had more blue or orange dots (Bartlett and Mccarley, 2021). Again, performance was worse in the high reliability automation condition when

operators were given the option to over-ride, than it would have been if the automation had done the task alone.

This is problematic concerning the use of high-reliability automation. It implies that it should be used with a high level of autonomy in order to increase performance, however, if the automation fails, this can lead to the negative human-performance effects as shown above. This appears to be another trade-off with high-level automation, as the rapid decline in performance when it fails could make it unsuitable for safety critical tasks, even if performance is highly enhanced under normal behaviour.

2.3.2: Transparency and Situation Awareness

As argued above, in order to better calibrate trust, and provide improved resilience to trust, a transition from human-machine interaction where a human exacts control over the machine, to human-agent teaming (HAT), which encourages interdependence and collaboration with autonomy in order to accomplish a task, is needed.

Human agent teaming posits agent transparency is essential, as it promotes three key factors for human-agent interaction: mutual predictability of teammates, shared understanding, and the ability to redirect and adapt to one another (Lyons and Havig, 2014; Chen *et al.*, 2018).

A challenge to HAT is enabling an autonomous agent to be able to clearly communicate its intentions to human team members. Capability to infer intent of a team member requires a way to facilitate explicit, directed communication Schaefer *et al.*, 2016.

A team can be defined as, "two or more people who interact dynamically, interdependently, and adaptively towards a common and valued goal/objective/mission, who have each been assigned specific roles or functions to perform, and who have a limited life-span membership" (Salas *et al.*, 1992).

By establishing common ground with team members, actions can be mutually understood and expected (Schaefer, Evans and Hill, 2015). This facilitates a shared understanding of mental models and goals. Even though autonomous agents have different decision-making processes and contexts compared to human team members, it is still possible to establish the development of team situation awareness.

Team SA can encompass joint decisions and actions. Each individual agent may have their own SA to carry out goals, but a shared SA will enable coordination, so that individual sub-goals can support the accomplishment of overall goals (Mica R. Endsley, 1995).

Transparent user displays enable an autonomous agent to communicate its reasoning process to a human agent, thus increasing the shared SA of the team. This can help to reduce ambiguity, misunderstandings, errors and unnecessary interactions Schaefer *et al.*, 2016. This supports the idea that information transparency is a critical part of building trust in teams, and developing shared SA (Chen and Barnes, 2014; Ososky *et al.*, 2014).

The Situation Awareness Transparency (SAT) model (Barnes *et al.*, 2014; Ososky *et al.*, 2014) can help to organise information requirements for better HAT performance. It has three levels. The first level provides a user with basic information about the autonomous agent's current state and goals, intentions, and proposed actions. This is to assist the human's perception of the agent's current actions and plans.

The second level provides contextual reasoning, including constraints and affordances, to assist a human's comprehension of the agent's behaviours. The third level includes information regarding the agent's projection of future states, consequences, and likelihood of success or failure, as well as uncertainty associated with the projections. This helps to assist the human's projection of future outcomes.

Mercado et al (Mercado *et al.*, 2016) used the SAT model to design a display for simulated operation of an unmanned vehicle. They tested three levels of SAT, with the first level showing which vehicles were used and the path they utilised, the second level with an additional text box with the agent's rationale explained, and the third level with additional information about uncertainty, by modifying opacity and colour of the icons, and additional bullet points in the textbox explaining the reasons behind the uncertainty.

Operator performance was significantly better when using the SAT level 3 display. Correct rejection rates were also significantly higher for the SAT level 3 display. They also measured no significant workload differences between the three conditions. Trust was also measured using the Trust Between People and Automation checklist (Jian, Bisantz and Drury, 2000). Trust was found to increase as transparency level increased.

Transparency improves operators' trust in less reliable autonomy by revealing situations where an agent has high levels of uncertainty. This helps calibrate trust in an autonomous agent by making the agent's limitations clear (Schaefer *et al.*, 2016; Chen *et al.*, 2018).

Lyons (Lyons and Havig, 2014) focuses on human-robot teaming, but makes valid points; they posit that "moving from tool to teammates requires the systems be designed with more naturalistic interaction styles, which may attempt to leverage the nuances of human-human interactions".

Developing shared awareness of environmental constraints within a task domain, conveying intent, explaining decisions in a timely and accurate manner, and communicating limitations, and progress towards goals, are all positive strategies for both increasing transparency and improving team situation awareness.

Chen does suggest that if level three SAT information is ambiguous, it may lead to increased complacency. In the RoboLeader experimentation discussed in (Chen *et al.*, 2018), where participants controlled robotic vehicles in a simulated environment, when participants had limited

information about the environment and medium transparency reports on RoboLeader reasoning reduced complacent behaviour and improved performance on the operator's route selection task, without increasing workload. However, when timing information was added, which was potentially ambiguous, participant behaviour and complacency behaviour was negatively affected. For the high and the low transparency conditions, participants showed more complacency behaviour, and lower task performance.

This shows that information requirements could be linked to both personal differences in complacency potential, or to knowledge of the environment, which can change dynamically throughout a task, and so bi-directional communication between the human and the agent could be beneficial, to establish better shared situation awareness.

Lyons (Lyons *et al.*, 2017) tested three conditions of transparency in a simulated study using commercial pilots to land multiple planes using an automated aid providing decision support. The automated aid provided no text-based reasoning in the first condition, values for likely success in the second condition, and logic, or rationale for its recommendations in the third condition. Trust was measured to be highest during the third condition, supporting the idea that sharing reasoning through increased transparency promotes trust in HATs.

Table 3: Further user requirements identified through trust in automation literature

User Requirement Identified	Reasoning
A system should be calibrated to be sensitive in order to minimise the chance of near misses (Rice, 2009)	In a classification task within a safety-critical environment, missing a contact would have worse consequences than receiving too many false alarms for a contact, therefore systems should be sensitive. However, careful calibration is required, as false alarm and near-miss rates affect both compliance with and reliance on a system

Human-in-the-loop activities must be maintained in order to ensure adequate SA (Onnasch *et al.*, 2014b)	Complacency is high when automation is accurate, and the further removed an operator is from a system's work, the harder it will be for them to spot when things go wrong
Automation should be at least 70% accurate to be implemented (Wickens and Dixon, 2007)	If accuracy is below 70%, operators have been shown to still depend on automation, and performance has been shown to be worse when compared to no automation
Automation should not replace a human when doing a task; human and automation should work together as a team to accomplish a task	Teaming, with both contributing to goals, prevents the negative implications of low SA, high complacency, and can help trust to be calibrated appropriately
Information autonomy would be more prudent to implement rather than action autonomy in a safety-critical domain (Onnasch *et al.*, 2014b)	There is a heavier cost to performance when action autonomy is implemented when compared to information autonomy when considering autonomy failure. Mitigating risk is important in a safety-critical domain, and therefore erring on the side of caution when considering what type and level of autonomy to implement is sensible
Information autonomy should seek to highlight areas of a display to an operator, rather than removing any raw information (Parasuraman,	Removing data from an operator's view makes it more difficult for them to validate autonomy's performance and can lead to reduced SA and over-reliance. Therefore, low levels of information automation are best in a safety-critical domain

Thomas B. Sheridan and Wickens, 2000)	
High levels of "team SA" should be encouraged through as much context and transparency behind automation's decisions as possible (Endsley and Kaber, 1999; Schaefer, Evans and Hill, 2015; Mercado *et al.*, 2016; Chen *et al.*, 2018)	Explanation behind automation's decisions can help evaluate their usefulness. Completing a goal as a team helps to reduce complacency and loss of SA. Higher performance in C2 tasks has been recorded when the automation offers increased levels of explanation for its decisions

2.4: Explainable Artificial Intelligence

As established in the trust literature, for an artificially intelligent or autonomous system to be trusted appropriately, its processes, performance and purpose must be transparent to the operator. This causes problems when considering many machine learning algorithms use classifiers which are opaque to the operator.

Explanation behind decision-making is crucial to be able to evaluate the efficacy of a classifier. In (Freitas, 2014), they present an example of an artificial neural network which classifies tanks as friendly or an enemy. It had very high accuracy when classifying the test images, but in the field, the accuracy was very low. This was because photos of friendly tanks were taken with a sunny background, and photographs of enemy tanks were presented with a darker, more overcast background in the test data, and so the sky was the feature being used for classification.

In (Ribeiro, Singh and Guestrin, 2016), a model is trained to distinguish between pictures of huskies and wolves. However, all photographs of wolves contained a snowy background, and the model was distinguishing between the two by looking at whether the picture contained snow.

In both of these cases, some kind of explanation for what was being analysed would have been beneficial in determining whether to trust the classifier's output.

A growing body of literature explores how to create more explainable artificial intelligence, known as eXplainable Artificial Intelligence (XAI).

Schaekermann et al. (Schaekermann *et al.*, 2020) compare two AI assistants which provide classification labels for medical time-series data. Both assistants integrate uncertainty estimates of their own performance. However, the "ambiguity aware" AI additionally provides a human-interpretable argument for any ambiguous or conflicting labels. This argument was either selected randomly, or by experts. Schaekermann's research shows that the provision of this human-interpretable explanation increases performance, especially when the arguments are highly relevant. Conversely, when arguments are random, accuracy is heavily negatively impacted, with less than 50% accuracy, lower than random guessing (Schaekermann *et al.*, 2020).

This work makes a very important point; that self-confidence or uncertainty percentages with context are much more useful in decision-making than those with no context, or random or bad explanations, which can have a negative impact on performance, trust and perceived workload (Schaekermann *et al.*, 2020).

The explanation provided by the ambiguity aware AI was text-based, and takes up a sizeable proportion of the experimental interface. This type of explanation may be more suitable in a medical field, where screen space is less of a valuable commodity. However, in the context of SMP operations, screen-space is limited and precious; text-based explanations may not be the most appropriate. Also, when experiencing high time constraints and a heavy volume of classification

decisions to make, text-based explanation may focus attention on reading explanations instead of detecting emerging contacts.

Ehsan et al. study automated rationale generation as a real-time explanation approach. This is done by a computer model learning to translate an autonomous agent's internal state, turning data representations into natural language (Ehsan *et al.*, 2019). They did this by creating a training data set by getting humans to play the arcade game Frogger, and explain the rationale behind their decisions using a think-aloud protocol. The definition used for automated rationale generation is "a process of producing a natural language explanation for agent behaviour as if a human had performed that behaviour". (Ehsan *et al.*, 2018).

Ehsan shows that users prefer detailed rationale, allowing them to form a stable mental model of the agent's behaviour. However, any rationale, no matter how broad, increased confidence and understandability in the agent. This goes further to support a need for an explanation of autonomous decision-making in order to maintain trust and confidence.

However, although this method works well for a game, with very linear constraints and predictable, limited behaviour, it may not be so readily applicable to a problem with more dynamic variables which are open to a wider range of interpretation, or that does not behave in a linear, sequential way.

Again, the explanations offered were text-based, which also may not be appropriate in a domain with limited screen-space and a need for focused visual attention.

Another point of contention would be the desire to make explanations "as if a human had carried out the behaviour". Although the initial results seem to indicate participants felt confident in the rationales provided, this feels like a mis-representation of the capabilities of the system; humans and AI do not possess the same thought processes or ability to think contextually – providing rationale which could have come from a human may encourage an operator to infer other human-

like characteristics about the system, which could lead to rapid trust degradation when the system behaves unexpectedly, or makes a mistake, and acts in an un-humanlike manner.

In their paper discussing theoretical explainable AI techniques for an autonomous ferry, Glomsrud et al. raise some interesting points about explainability (Glomsrud *et al.*, 2020). They posit that currently the methods for explaining AI or autonomy are insufficient, as they are mainly framed towards AI developers, and do not help actors with less familiarity or knowledge of AI systems to understand the decision-making processes of a system. They therefore posit that explanations depend on context of use, which corroborates (Ehsan *et al.*, 2019), and that the explanation required changes depending on the use-case and user. However, they do not provide any additional information about how these explanations could be generated, or what they look like.

They highlight the key irony that as a vehicle becomes more autonomous, it becomes more difficult to switch control effectively between the human and the vehicle as the demands on situational information and alertness of the driver are higher. Therefore it requires a closer coupling between the human and vehicle, despite the increase in autonomous behaviour. This corroborates the work of (Parasuraman, Thomas B Sheridan and Wickens, 2000b).

Local Interpretable Model-agnostic Explanations (LIME) is a popular method for explaining the predictions of complex and black-box machine learning models (Dieber and Kirrane, 2020). LIME can be used to provide an explanation for a specific prediction. LIME then creates slightly different versions of the selected data points by making small changes to its features. These changes are random, but follow specific rules to maintain realism. These are known as “perturbed samples”. It then uses the perturbed samples and corresponding predictions to create a simple model, which approximates how the black box model behaves. LIME can then generate a list of feature importance scores and coefficients to show how different data was used and weighted by the black box model. Positive scores mean that a feature had a positive influence on prediction. The larger the absolute value of the score, the more influential the feature was in making the prediction (Ribeiro,

Singh and Guestrin, 2016). LIME has been shown to increase trust in output with its explanations, and outperforms other explainable models (Ribeiro, Singh and Guestrin, 2016).

However, LIME does have some limitations. For example, the choice of perturbations applied to the input data can lead to differing explanations, even when the variations are very small (Dieber and Kirrane, 2020). This makes the results less robust. LIME can provide explanations in terms of feature importance, but these explanations may still not be intuitive or easy to interpret by non-expert operators. It also is time consuming to generate explanations, especially when a large number of explanations are needed, or for a real-time application. This would make LIME unsuitable for our research, which requires explanations on-the-fly, and which involves many nuanced feature sets which are extremely sensitive to change.

Another method of providing an explanation behind AI is to use a decision tree. A decision tree is a machine learning algorithm used for classification and regression tasks (Blockeel *et al.*, 2023). It makes decisions by recursively partitioning data based on features, resulting in a tree structure. Decision trees provide an easily understandable and traceable way to see how an algorithm reached a certain outcome. They represent a particular approach to modelling data. It also gives a visual representation of a problem, which could be useful in domains where text is not appropriate.

Decision trees are often lauded as an easily assessable way of visualising a decision making process (Blockeel *et al.*, 2023). However, they have limitations. Decision trees are prone to overfitting data, and can create very complex, overly detailed trees which do not generalise well to unseen data. This can make them more difficult to interpret (Dietterich, 1995). Even small changes in the input data to a tree can cause significant changes to the structure. This can make decision trees relatively unstable (Li and Belford, 2002). They also struggle to capture complex relationships between features as effectively as neural networks, and so are better used for problems with simplistic decision boundaries.

As well as this, bias in a data set can be greatly amplified by a decision tree, as it can learn and perpetuate the biases through its rules (Dietterich and Kong, 1995). They also struggle to handle continuous data, as they convert continuous variables to discrete ones, which can lead to information loss and less accurate modelling (Blockeel *et al.*, 2023). Similarly to LIME, decision trees lack global understanding, and may not offer a global understanding of a model's behaviour. They can struggle with high-dimensional data, which can lead to complex trees which are difficult to interpret.

In conclusion, current methodologies for providing on-the-fly explanations for decisions made by AI are not capable of providing easily interpretable information in real-time. Explanations are very sensitive to biases within the training data, and can lack robustness, as very small changes can produce variances in explanation. They are vulnerable to adversarial examples. This is especially true for decision trees. Adversarial examples are data points which have been intentionally manipulated to cause a misclassification. Adversarial examples can be used to poison a data tree, where attackers can identify weaknesses in a model's decision boundaries and manipulate the input data to cause erroneous classifications. They can be very small and difficult to spot. This particular weakness would make decision trees unsuitable for use in a defence environment, where erroneous data is used systematically to obfuscate or confuse.

2.5: Discussions and Conclusion

The literature review summarises current research into human-autonomy teaming, and highlighted a number of pertinent paradigms which must be taken into account when introducing new forms of intelligent information systems into established socio-technical systems.

The models of trust discussed within the review both highlight that individual differences in propensity to trust automation, personality and culture, and pre-conceptions of a system's capabilities, can all influence an individual's initial interactions and trust towards autonomy.

These pre-conceptions can cause over-reliance, through over-trust, or under-reliance, through distrust, depending on a person's perceptions of their own abilities and the system's abilities. In order to reduce mistrust, which can lead to inappropriate interaction with a system, operators must be appropriately trained and primed in order to have realistic expectations of a system's capabilities and limitations.

Mitigating these pre-conceptions through appropriate training and understanding is imperative to prevent mistrust, which can lead to sharp reductions in performance and trust which can be difficult to recover from.

It may seem counter-intuitive, but the literature shows that perfect performance of autonomy can lead to worse performance overall, as it can introduce factors associated with complacency and low situation awareness. As well as this, when an autonomous system can work at perfect levels of accuracy, some studies have found operators' expectations will lead them to reject the outputs of the system, as distrust and mis-calibrated expectations still affect interactions.

Therefore, a system must operate with as much transparency as possible, allowing for operators to interpret and evaluate the system's suggestions and actions accurately, which allows for more appropriate calibration of trust in autonomy.

Transparency on its own may not be enough when it is difficult for users to interpret the reasoning behind a system's actions, which shows a clear need for better explainability of autonomous systems, and so ways to facilitate understandability and inter-communication between the operator and autonomy are key for reconciling user preconceptions and systems' performance in reality.

This also must be coupled with careful design not only of the system, but of the form of interaction for an operator with that system. The more removed an operator is from the system's functionality, the worse performance will occur when something goes wrong. Methods for ensuring the human is

kept in-the-loop, and not on-the-loop or even outside-the-loop is vitally important, especially with regards to safety-critical tasks.

Cognitive dissonance between actual and perceived reality of performance, and of awareness of the reality of a situation, both incur heavy penalties on performance. Therefore robust design will always seek to introduce as much transparency as possible, as well as limiting erosion of an operator's understanding of the true state of reality. Therefore it is recommended that any introduction of autonomy into the domain must attempt to keep raw information available and viewable to an operator, with any manipulation (highlighting, for example, or prioritising information in a display) not obscuring the real state-of-events, whilst still providing recommendation. This allows humans to form their own assessment of events when required, and facilitates better situation awareness, which can decrease the gap in understanding when things do inevitably go wrong.

The results of the literature review directly help to answer RQ1. Information automation at low levels is preferrable in safety-critical tasks; this is because it does not hide raw data from an operator, and so helps to reduce the penalty to SA of high-reliability automation. Action automation incurs a more severe penalty when automation does not behave properly, and so should be avoided for safety-critical tasks. Humans and automation should work together to achieve a goal, rather than one carrying out the job of another; this also helps to reduce skill-loss, complacency, and over-reliance, as well as loss of SA. Automation should be at least 70% reliable; otherwise this can lead to a penalty for overall performance. Automation should also try and offer high-level explanations for its decision-making which a human operator can easily understand; this helps to create shared, or team SA, and also keeps trust calibrated to an appropriate level, by helping make the performance and processes of automation clear to the operator, and offering them a way to assess the automation's decision-making, which can help to prevent the effects of automation bias.

It is also concluded that it would be more dangerous for a system to fail to alert an operator to a problem, than a system which gives too many false alarms; this is also because of safety concerns:

missing an important contact could have worse consequences than treating a contact with greater concern than warranted; it would always be better to be hyper-vigilant in this scenario, as underestimating a contact could have severe consequences.

CHAPTER 3: A SUBMERSIBLE MARITIME PLATFORM AS A SOCIO-TECHNICAL SYSTEM

3.1: Chapter Aims

This chapter aims to provide context and understanding to help answer RQ1: “What Level and Type of Autonomy could be suitably applicable to tasks carried out in the process of broadband sonar classification whilst maintaining appropriate levels of trust in the automation?” This is to better understand and specify the context of use, as shown in the diagram in Figure 1, representing the user-centred design process as defined by ISO 9241. By understanding what information is used, by whom, and how it is generated and communicated, this can help identify areas where autonomous systems could help optimise the process and reduce uncertainty or cognitive load. An overview of the tasks and information used in an SMP to understand their environment is presented.

3.2: Introduction to the Task

This chapter seeks to summarise the key actors, technical systems, and information exchanges and interactions which occur on board an SMP to better facilitate understanding of common tasks, and how these could be affected by the introduction of autonomous systems.

SMP Command teams rely on the ongoing interpretation of sensor data to navigate safely when operating from below periscope depth. When “going deep”, limited sensor information is available. SOs use signal processing techniques and frequency analysis to identify and classify contacts by listening to noise picked up from hydrophone arrays, contributing to a distributed, team model of the operating environment. Not all agents who contribute to the model have the ability to see the complete operational picture (Stanton, 2014a). The complete situational model is held mentally by the OOW.

SMP control rooms have been described as sociotechnical systems; a system which "involves the interaction of human operators and technology, with interdependence to pursue broader goal-

directed behaviours creating the conditions for successful overall performance (Salmon *et al.*, 2006; Stanton *et al.*, 2009; Walker *et al.*, 2009). They rely heavily on effective communication and teamwork (Stanton, 2008). Communication and teamwork can be impactful on team workload, sometimes even more-so than the work itself, by either facilitating rapid and effective information communication, or by preventing it (Driskell, Salas and Hughes, 2010).

Below periscope depth, SMPs are able to operate covertly, undetected from the surface. The trade-off for remaining undetected is a lack of external information to make sense of the environment; all information pertaining to the operational area must be derived from integrating sensor system data (Dominguez, Long, Miller and Wiggins, 2006; Roberts, Stanton and Fay, 2018) or collected pre-mission, and is mediated by complex computer systems and chain of command. Returning to the surface to gain information could compromise the SMP's operations.

3.3: Task Description

In an SMP control room, and socio-technical systems in general, cognitive processes and situation awareness are distributed across actors, and often are not fully comprehended by individual actors (Stanton, 2014a; Roberts, Stanton and Fay, 2015). Team configurations and how the technology facilitating their communications is used can both influence their effectiveness.

Onboard an SMP, verbal communication is highly structured, and follows rank hierarchy (Fay, Stanton and Roberts, 2019). Verbal messages are received, acknowledged, and repeated back to ensure that correct information has been received (R. R. Murphy, 2000). Verbal information is in this way filtered and aggregated, until it reaches the Officer of the Watch (Carrigan, 2009), who is usually in charge of the control room, and is tasked with making navigational decisions. For example, SOs report to the Sonar Controller, who then can communicate information to the command room via the Operations Officer (OPSO), therefore acting as an informational filter.

The OOW directs the command team from within the control room, that works together to generate a tactical picture (Dominguez, Long, Miller, Wiggins, *et al.*, 2006). A tactical picture is a dynamic

model of the SMP's environment, including the perceived positions of contacts, which are vessels or objects which have been detected by periscope, sonar, or radar. Contacts are analysed within the control room, to try and understand their behaviour in relation to own-ship. The tactical picture directly informs strategic and operational decisions, and so accuracy of the tactical picture is of strategic importance.

A high volume of sensor data must be interpreted and communicated in order to develop a shared situation awareness of the developing tactical picture. As more advanced sensors and methods of data collection are employed in the control room, the potential for this information to exceed the capacity of an operator to interpret it effectively is increased (Woods, Patterson and Roth, 2002). This can lead to incorrect processing or interpretation of data, causing a degradation in the quality of the tactical picture. Highly demanding situations, such as high numbers of contacts, has been shown to reduce the cognitive capacities of operators, and reduce the volume of information they can handle (Roberts and Cole, 2018).

Most operators use a specific display to analyse a specific kind of sensor system or information, using their expertise to analyse the data and communicate their extrapolations up the chain of command. However, the OOW, who must maintain the overall tactical picture, looks at multiple repeater displays for many different information types to try and assimilate information across different technical systems. This includes geographic and sonar information. They must try and quickly interpret information across these different systems, and perform mental calculations, to constantly update the tactical picture. The Commanding Officer (CO) may be the only person holding a complete tactical picture. However, the tactical picture is dynamic, and even the CO is not explicitly aware of all information transactions (Dominguez, Long, Miller, Wiggins, *et al.*, 2006). COs are concerned with what is going to happen, and must have a big-picture understanding which is future-oriented in order to plan effective COAs. However, they have to do this by looking at several displays which present current, and near-past information, each for specific systems (Dominguez,

Long, Miller, Wiggins, *et al.*, 2006). Vital signs can be dispersed across different displays, requiring the SO to move between them, constantly integrating different pieces of information into a collective picture (Dominguez, Long, Miller, Wiggins, *et al.*, 2006)

Sonar relies on interpretation of passive hydrophone recordings. Generally, hydrophone arrays are positioned at the bow and flank of SMPs, with an additional towed array able to be deployed, combined with an antenna system and signal processing method for ranging targets passively (Brinkmann and Hurka, 2009). An incident acoustic wave front curvature is measured using the sonar system. This allows a target range to be calculated. This is very dependent on the acoustic transmission conditions of the area. Background noise - both ambient noise from the ocean itself, and also from the SMP itself, can make it difficult to identify potential contacts.

Detecting a vessel requires an operator to identify discreet noise against this background noise aurally, or visually on their waterfall display, where it forms a line, at a specific Direction of Arrival (DoA). Modern boats, with quieter engines or advanced hull designs, may operate quietly, and so it can be harder to detect their presence. They may not produce a readily discernible, clear trace (Matthews *et al.*, 2005). This is exacerbated by systems not highlighting emerging traces to the operator. Other factors affecting the clarity of contacts are the conditions of the ocean itself, including weather effects and salinity, temperature and pressure (Gimse, 2017), shallowness, or the complexity of the sea current (Shar, Li and Shar, 2000). These can affect the accuracy of measuring the bearing (Shar, Li and Shar, 2000). Time-of-arrival measurements between the different hydrophone arrays allow range to be inferred, but this is quite inaccurate, especially at the beginning of detection, and is adversely affected by target distance.

Bearing angle is displayed over time to produce a waterfall display. For every time period that a sonar array returns data, it is plotted as a line on the display. When a new line is added, all other lines are moved down, which creates the "waterfall" effect (Asplin and Christensson, 1988).

Once the DoA estimation has been performed, two types of analysis can be implemented to extract relevant signal features: DEMON (Kemper *et al.*, 2019) and LOFAR. DEMON is a narrowband analysis which visualises propeller characteristics: the number of shafts, shaft rotation frequency, and blade set (De Moura, De Seixas and Ramos, 2011). DEMON works over the cavitation noise of the target propeller (De Moura, De Seixas and Ramos, 2011). It shows a demodulated broadband signal and can help to estimate speed by combining the frequency of the shaft with a Turns per Knot value (TPK), which is how many times a propeller turns per one “knot” of speed. TPK is obtained from a classification database, or can be estimated.

LOFAR is a broadband analysis, estimating the noise vibration of target machinery (De Moura, De Seixas and Ramos, 2011). Independent Component Analysis algorithms are employed to minimise signal interference from neighbouring directions to try and aid with target detection.

Both are based on spectral estimation. Classification is typically performed using narrowband (DEMON). The classification is not validated by the system, and so operators are not aware if a classification is potentially incorrect. As the classification is not validated, this can mean the TPK value is incorrect for a vessel, and so speed can also be calculated incorrectly. This is dangerous, as it can invalidate a contact's known position. The only thing which is known to be accurate is bearing.

Tracking is performed by analysing the broadband trends over time to try and determine what actions a vessel is taking (Fay, Stanton and Roberts, 2019). When a vessel is detected, operators can assign it an identifier, or tote, which allows the system to automatically track and update its location, and communicate these details to other consoles. The identifier is a tracker (Fillinger *et al.*, 2010).

Once a target has been initially classified, the "cuts" of the data, estimated speed, and tracker, are passed to the control room for TMA. A cut is a straight line, representing the Line of Bearing (LOB) for a signal.

TMA is a solution to try and determine the location and predicted movements of a contact by analysing positional data derived from passive sensors (Cunningham and Thomas, 2005). This is called a "solution", and consists of speed, range, course and bearing for a vessel (Genç, 2010). The bearing is the relative direction to a contact from own-ship.

In order to generate a solution, a Local Operations Plot (LOP) can be used. A LOP is a chart plotted with the previous detections of a contact, which can allow a solution to be calculated (Clarke, 1999). The LOB of a signal is plotted on the LOP between own-ship's position and the maximum detection range of a sensor, which provides an equivalent angle to the detection bearing (Fay, Stanton and Roberts, 2019).

Operators can merge cuts from different sensors and treat them as a singular contact (Huf and Brolese, 2006) to provide more information on its behaviour. However, sometimes operators may merge multiple discrete contacts' cuts, which then can cause incongruent information to be displayed.

Once multiple cuts have been collected, the operator can then analyse a vessel's path using a “speed strip”. This is a visualisation of a “vessel's historic path in the water”. There are marks along a line to represent where a vessel would be if the path is correct. It can then be positioned, with marks added to the strip for each cut. An operator aligns the strip over the cuts. If the marks intersect the cuts, it provides a solution as to how the vessel may be travelling. As a solution builds and becomes more accurate, dots, representing the strip's intervals, will form a “stack” (Huf and Brolese, 2006). Multiple speed strip configurations could align, which means there is ambiguity, as only one alignment will be correct (Cunningham and Thomas, 2005). There are infinite solutions for this, which must be narrowed down using all of the available information. More cuts can rule out certain solutions.

The process of TMA is manual in this sense, rather than cognitive, where operators are manipulating a waveform and trying to get dots to stack, searching a problem space mechanically rather than recognising a solution based on information presented to them (Cunningham and Thomas, 2005).

Once a perceived to be accurate solution, which matches all of the cuts, is defined, this can then be shared (Huf and Brolese, 2006). This allows a contact to be plotted on a geographical view. Dead reckoning is used to plot positional data of a contact, extrapolating previous trends (Murphy, 2000). This geographical view can be seen in the command room.

Design issues, such as transient signals not being highlighted on the waterfall, or TMA solutions not being constrained, can add further complexity and cognitive workload to the tasks (Fay, Stanton and Roberts, 2019).

Once a contact has been classified, the OOW uses all available information to assess what the Closest Point of Approach (CPA) for the contact should be. The CPA defines a safe distance that must be maintained between own-ship and a contact. If a contact enters the CPA, a SO must call a Close Quarters (CQ) procedure, requiring Command to assess the situation and advise of what action should be taken. This is to avoid collision with the contact.

3.4: Discussion

Through the description and analysis of an SMP as a socio-technical system, it becomes clear that there are inherent uncertainties throughout the contact classification process. Although previous literature studies information assimilation and tactical picture generation for SMP command teams in detail, and posits the importance of track management and contact classification, the specific task itself from a stand-alone perspective, remains under-studied.

Coming into port, getting underway, coming to periscope depth, and transiting congested and constrained waterways can all mean a high volume of dynamically moving contacts must be classified and tracked (Dominguez, Long, Miller, Wiggins, *et al.*, 2006). This can put cognitive strain

on both operators and command teams, as knowing where contacts are in position to own-ship at all times is paramount to safety. This picture is never static, but constantly changing and moving.

There is inherent uncertainty in track management, partly due to a combination of the limitations of passive sonar capabilities, and also deception techniques used by hostile forces (Kirschenbaum *et al.*, 2014; Stanton, 2014b; Stanton *et al.*, 2020b). However, even though the uncertainty was acknowledged, it is still unclear how it is mitigated.

Much previous work focuses on how this passive sonar data is integrated into a larger tactical picture held by the command team (Stanton, 2014b; Loft *et al.*, 2015, 2016; Roberts and Stanton, 2018a).

Inherently, the building of the tactical picture, and the successful monitoring, classifying, and tracking of targets, is intertwined. Loft (Loft *et al.*, 2016) speaks of the track manager, who communicates closely with the watch leader, equivalent to the OOW, to develop an understanding of the bearing, range and speed of contacts held by the SMP sensors, and their relation to own-ship and strategic landmarks. In order to safely manoeuvre, constant communication is required to maintain situational awareness of contacts, with range and bearing rate being constantly re-assimilated and communicated between the two parties.

When beneath periscope depth, track management underpins the core situational awareness of the SMP (Loft *et al.*, 2016).

Stanton, and Stanton and Bessell, (Stanton, 2014b; Stanton and Bessell, 2014) posit that SMP crews do not create stable mental representations of their task environment; instead, the representations are dynamic, and informed by frequent interactions with information displays and team members, with different actors monitoring different facets of the tactical picture.

Much of the literature focuses on building models of distributed situational awareness, communication and information networks, and studies how these impact the development and maintenance of this dynamic tactical picture (Stanton, 2014b).

Some of this work relies on simulations of scenarios, and therefore the contacts used in the scenarios represent an actual, situational truth. However, this is not altogether realistic; uncertainty always underlies the represented position of contacts on tactical displays (Kirschenbaum *et al.*, 2014), affected by the unpredictability of the ocean itself, as well as hostile contact deception and stealth (Loft *et al.*, 2016).

Contact localisation is inherently ambiguous, with multiple possible contact solutions available, as can be seen from this task description. In order to reduce uncertainty, Target Motion Analysis and SMP manoeuvres iteratively try to create convergence in the potential solutions. This process is repetitive, takes time, and can fail, due to ambiguity and obfuscation of contacts and tracks. This means there is variability in how much the tactical picture matches the "ground truth".

When studying the relationship between uncertainty and SA using simulated track management tasks, Loft found significant differences depending on the expertise of participants. Expert participants were less willing to answer Situation Present Assessment Method (SPAM) – a type of SA measurement - queries in a timely manner when compared with students, implying they were cognitively too busy to self-report, making it difficult to accurately infer precise measurement of SA and workload for expert participants.

What can be inferred is that much of the work of SOs is heavily cognitive, and incorporates tacit knowledge, meaning it is difficult to explain, evaluate and understand. Experts explained to Loft that "answering SA queries was not as high a priority" to them as other aspects of their simulated task.

When simulated tasks rely on self-report techniques during simulation, especially with experts, results can be limited due to the task and the self-reporting both competing for the attentional resources of the participant, making self-report less accurate.

Methods for understanding workload and situational awareness often employ self-report during scenarios, where the method itself of measurement competes for cognitive attention with the task, resulting in only partial understanding of the cognitive processes and resources dedicated to the task.

Previous work focuses on the development and communication of the overall tactical picture, with distributed focus across a larger informational network, instead of micro-focus on the task of classification.

Although students, or participants without expertise, may be more willing to provide information via self-report, this does not accurately reflect the experiences of experts in simulation, as they lack the cognitive strategies that differ with expertise, and so cannot always be used to infer a realistic representation of the demands of the true task for an SME.

Therefore, this thesis chooses to focus solely on the task of contact classification itself, positing that the actual cognitive work of contact classification is under-studied, and even sometimes mis-represented, in previous works.

This can be because contact positions are directly related to ground truth during simulation in previous studies, meaning the task is inherently different to the reality of track management, where ground truth is rarely, or even never, known. As discussed in the CDM interview in Chapter Five, sometimes Operators are informed many weeks after an event that they actually missed a contact of interest, only once data has been analysed back onshore. This shows how far removed from ground-truth “in the moment” classification really is.

It can also be because the participants used to study track management tasks do not possess the expertise needed to accurately convey the mental work involved with the real-world task.

This work hopes that by focusing and developing a broader understanding of the cognitive processes employed in classification, a task which underpins the formation of an accurate tactical picture, recommendations for autonomous support specifically for this task can be developed which will have an impactful effect on improving that tactical picture generation.

Facilitating information flow between the sound room and TMA Operators would be beneficial. In much of the discussed research, they are situated in different rooms, causing extra steps for information to be shared among the two teams. Having SOs in a separate room is a legacy feature of an SMP (Stanton and Roberts, 2020); with waterfall displays and better noise-cancelling head-sets, it may be sensical to position the two tasks closer together physically, or develop better information exchange protocols, potentially through the use of autonomy. The OPSO is often under very high cognitive load, both informing the tactical picture of the OOW, and communicating with other operators. By providing a way for TMA and SO operators to communicate directly (with permission), this could potentially lower the chance of the bottleneck in the form of the SC, and give more cognitive bandwidth to the OPSO.

Similar recommendations have been made in (Stanton and Roberts, 2020), who also identify that having SOs in the same room would have the additional benefit of helping the command team develop improved awareness of what is happening currently in the sound room.

This Chapter has highlighted a number of areas within the socio-technical system which could benefit from some level of autonomous support. These include:

- Focusing on broadband sonar classification, as this underpins much of the tactical picture used by Command to understand their operational environment, and yet it contains many inherent uncertainties

- Reducing the cognitive load of an SO when there are a high number of contacts to detect, classify and monitor
- Supporting information transfer between the sonar and TMA Operators, as this currently requires messages to be passed back and forth between teams, and heavily relies on the SC and OPSO to accurately assimilate, filter and communicate information, which can lead to a high cognitive load for these crew members in particular

In this way, Chapter Three helps to answer RQ1 by identifying a specific task, broadband sonar classification, which could benefit from the addition of some level of autonomous support.

CHAPTER 4: CASE STUDIES OF ACCIDENTS TO INFORM ADDITIONAL REQUIREMENTS

4.1: Chapter Aims

This chapter aims to answer RQ2: "How can the causes of previous SMP accidents be mitigated through the introduction of autonomy?" This is done by looking at two case studies of incidents, with particular regard to how sonar information was used (or mis-used) in those incidents, as the process of classifying contacts was highlighted in Chapter Three as a particular area where the introduction of autonomous support systems could be beneficial to the specific task, and to the information requirements of the socio-technical system in general. These incidents were chosen as there were issues with mis-classification of contacts in both. This chapter helps to identify further context of use and user requirements as outlined in Figure 1, showing the user-centred design process outlined in ISO 9241, by identifying examples of why this task is crucial to overall operations, and providing real-world examples of why the task could benefit from autonomous support.

The chapter begins by looking at the human factors literature surrounding surface ship collisions, to provide a starting point to understand human factors in maritime accidents, and to help inform the accident analysis methodology.

4.2: Human Factors in Surface Ship Collisions

A large body of work exists pertaining to the human factors behind shipping accidents, specifically collisions. These collisions have been analysed from a human factors perspective, to highlight areas where risks could be reduced or mitigated, and to show how the interactions between the human actors and the technical systems used in surface shipping can influence or cause collisions.

Understanding the human factors behind surface ship collisions can act as a starting point to better understand the human factors behind collisions involving SMPs.

This work is reviewed to highlight human causes behind maritime collisions. It also discusses the similarities and differences between surface ship collisions and collisions involving SMPs. Lastly, it helps to highlight that more human factors research is needed pertaining to the causes of collisions involving SMPs, showing a large gap in the literature. Critically reviewing the accident analyses of shipping collisions also helps to inform the accident analysis of the thesis and justifies the chosen methodologies.

Surface ship collisions are one of the primary causes of serious casualties at sea (Graham, 2012), accounting for around 50% of risk when considering high-traffic shipping routes (Mou, Tak and Ligteringen, 2010). Although collisions are becoming less common, they represent 71% of accidents in European waters when including grounding (European Maritime Safety Agency, 2010). This shows a vital need for continuous analysis of the causes of collisions, in order to minimise their occurrence.

Several studies highlight the role of organisational and human factors in maritime safety (Hetherington, Flin and Mearns, 2006; Chauvin, 2011; Chauvin *et al.*, 2013). These factors are often central to the causes of surface ship collisions.

The shipping industry is heavily regulated, with international regulations which seek to harmonise equipment requirements, ship design, and best operational practice within international waters. These regulations are set by the International Maritime Organisation (IMO). With the implementation of the International Safety Management (ISM) code in 1994, produced after the capsizing of the *Herald of Free Enterprise* in 1987, people in charge of a ship are required to establish a Safety Management System. This is less prescriptive than previous stipulations for design and manning. Instead, it seeks to move away from specific safety training and checklists assigned to specific personnel, and instead create a more inclusive culture of organisational safety in the shipping industry, through development of an industrial safety culture which involves all human actors, both on and off the ship itself. It has been mandatory for all vessels since 2002 (Chauvin *et al.*, 2013).

It is important to note, however, that not all countries accept all of the proposed legislation. Other factors, such as the size of a vessel, and type of vessel, affect which legislation comes into play. This can create issues around communication between ships, as the expected behaviour of other vessels does not align with their actual behaviour. This breakdown of communications has been shown to be a causal factor in many collisions in multiple works (Chauvin, 2011; Chauvin *et al.*, 2013; Sotiralis *et al.*, 2016).

However, despite an increase in regulation which highlights human and organisational factors since 1994, the link between inappropriate operations and collisions has been clearly shown in (Chauvin *et al.*, 2013) as a significant contributor to collisions. Chauvin categorises this as "unsafe leadership", which can either be related to inappropriate operations (insufficient manning, too high speeds considering environmental conditions), or a disregard for existing rules and Safety Management Systems, which occurred in 33.33% of all accidents reviewed. The idea that non-compliance with rules and the Safety Management System increases risk of collision is supported by the Marine Accident Investigation Branch, who state that, "collisions should theoretically be avoided if every vessel abided by the International Rules for the Prevention of Collisions at Sea 1972" (MAIB, 2004).

Chauvin et al (Chauvin *et al.*, 2013) present a modified Human Factors Analysis and Classification System (HFACS) analysis, which they use to identify patterns and trends in the causes of 27 collisions reported by the Marine Accident Investigation Branch. They report that unsafe acts are mainly related to decision-making (85%), and/or the non-perception of vessels concerned in 15% of the accidents analysed (Chauvin *et al.*, 2013). They also highlight poor visibility, and the mis-use of instrumentation, as the main environmental causes behind the accidents. This is backed up by the MAIB's finding that improper use of radar appeared in 73% of the cases they investigated (MAIB, 2004).

Unsafe acts, non-perception of the concerned vessel, poor visibility, and the mis-use of instrumentation, have all been identified as contributory causes in SMP collisions, as reviewed in

Chapter Four, showing a strong link between the human factors which apply to surface collisions and the human factors which apply to SMP collisions.

In terms of the condition of operators, Chauvin identifies deficit of attention, and poor situation awareness as contributory factors to collisions. They highlight the high workload on the bridge as a contributory factor, with navigational tasks, collision avoidance, and administrative tasks often being carried out concurrently, therefore detracting from attention and situation awareness.

A lack of situation awareness may be a problem which is compounded when considering SMP operations, as less external information is available when developing an understanding of the operational environment. As well as this, the OOW is lacking any bespoke display, and so must be amalgamating information from a variety of different sources, whilst also performing calculations to ensure navigational safety. This can lead to a reduction in attention. However, crew configurations and duties are regimented, with a clear chain-of-command mediating communication and duties on-board military vessels. This means two things; situation awareness is more highly distributed across the socio-technical system, and operators should be able to focus their attention on particular tasks rather than it being spread over multiple concurrent duties.

There are a number of limitations to Chauvin's work. The work uses HFACS analysis, based largely on Reason's Swiss Cheese Model (Reason, 1997). The Swiss Cheese Model posits that each level of the socio-technical system has weaknesses in it, known as voids, analogous to the holes in a slice of Swiss cheese. A weakness, or hole, in one layer, could allow a problem to pass through. However, the next level may not have a hole in the same place, stopping the problem from permeating. If the holes align in all of the layers of the socio-technical system, an accident occurs.

Although the model is widely used and accepted, this is not without criticism. The model has been criticised for being an oversimplification of how accidents occur, with no real understanding developed of links between different individual, causal or organisational factors (Larouzee and Le Coze, 2020). Reason actually criticised the breadth of his own model, quoted as saying, "the

pendulum may have swung too far in our present attempts to track down possible errors and accident contributions that are widely separated in both time and place from the events themselves" (Reason, 1997, p. 234).

Another criticism of the Swiss cheese model would be that it can lead to a false sense of security, as it can make the causes of the accident seem very difficult to align. The model also does not show detailed links between the identified causalities, making it hard to infer the precise relationship between them (Perneger, 2005; Larouzee and Le Coze, 2020).

HFACS is based on similar principles to the Swiss cheese model, and so inherits some of the problems associated with it. Some which are highlighted in the literature include wildly different reports of reliability, and little task standardisation, with the HFACS often being modified to suit a particular domain or organisation, leading to high levels of variability in how the method is applied (Cohen, Wiegmann and Shappell, 2015). Specifically, Cohen highlights how deriving the casual factors from accident or incident reports can cause higher variability and reduces reliability amongst raters (Cohen, Wiegmann and Shappell, 2015). This is the methodology used by Chauvin, who also modifies the HFACS to be better aligned with maritime collision analysis.

Chauvin suggests the small number of accidents used in the analysis as a limitation. This is one reason the HFACS may not be an appropriate accident analysis methodology to apply to this research, as the number of accident reports for SMP accidents is drastically smaller, often with much less, or even incomplete, details of the accident provided, as not to reveal any security sensitive information.

The accidents which were analysed only have two commonalities; they were reported on by the MAIB, and included a collision. This must have introduced many factors which lacked strong relationships, considering the different technologies, crewing configurations, shipping types, operating conditions, which would have made it difficult to identify clear patterns or relationships.

As well as this, although some specific problems were identified, such as poor situation awareness, or a failure to identify the other vessel involved in the collision, many are vague, such as an incomplete or ignored Safety Management System. The ways in which the Safety Management System can be incomplete, or why or how it is ignored, appear to be outside of the scope of the work, even though it is identified as a key reason for accidents occurring.

Overall, the meta-analysis provides a good introduction to the human factors which can contribute to shipping collisions, but is very broad. Some of the points discussed, such as the poor communications between colliding ships, are specific to surface shipping, and are less relevant to the operations of SMPs operating covertly below periscope depth. These collision incidents have not been analysed from a human factors perspective, and this is a gap in the research which this thesis begins to address.

Sotiralis et al. (Sotiralis *et al.*, 2016) present a Bayesian Network model of the human factors which influence the risk of surface ship collisions. The model uses an event tree which represents the consequences of a collision event, coupled with a Bayesian network which was developed to calculate the probability of collision, and to model the factors which contribute to the competence of the OOW, their detections, assessments, and actions (Sotiralis *et al.*, 2016). This can be used to quantify the probabilities of collision based on ship type, environmental conditions, and the mental state of the OOW.

However, Sotiralis groups together many different kinds of collisions, occurring between different kinds of ships, from different nationalities, in different waters (Sotiralis *et al.*, 2016). By grouping so many disparate incidents, which may have confounding factors, such as some legislation applying to some ships or countries and not others, different operational practices for the OOW depending on industry or type of ship, this may weaken the implied relationships they can infer about the causes of collisions.

Wang et al. study the causes of 200 ship collisions, using logistic analysis to work out the maximum impact factors behind the collisions (Wang *et al.*, 2019).

Like Sotiralis, they group together a high volume of different collisions. The model is linear, and so only deals with the "main" cause of each accident, which could be considered reductive, and does not help to understand the relationships between the different factors that align to cause unsafe conditions. They also do not specify the date range of the collisions they are analysing, or give details about the sources of all of their information. Wang and Sotiralis both potentially make the mistake of amalgamating many incidents which occurred before certain standards or regulations, such as the ISM code mentioned above, with recent shipping collisions. This could introduce information on causes of collisions which have been rectified through better equipment requirements or standards.

Antão et al. create a model for maritime accidents by applying a Bayesian Belief Network which maps the main causes and types of accident and their consequences (Antão *et al.*, 2009). The paper is very clear with their data, analysing reports for 857 accidents which occurred in the last ten years, using the database entries of the Portuguese Maritime Authority. This reduces the problems found in the previous works.

Their data draws interesting conclusions about the effects weather, time of year, and type of ship can have on the likelihood of collisions. However, they do not consider the human factors behind the accidents, instead, focusing more on environmental conditions and ship type.

Robust models can be computed for collision causation when there is reliable and similar data recorded for all of the collisions analysed. From the literature, it seems that the recording of human factors historically has been vague when considering surface ship collisions, which has a knock-on effect when trying to perform quantitative analysis. This is worsened by the inclusion of disparate data, either because it covers a large time-frame where legislation and equipment requirements have radically changed, or because it is fusing data from many different kinds of environment, vessel types and locations. For example, the USA does not accept some fishing vessel regulations which

other countries abide by. Therefore collisions involving these ships could flag as collisions involving a disregard of rules, where the rules simply did not apply.

The other issue is that a lot of the reports seem to have a larger focus on the "here and now" when a collision is occurring, which is understandable, but does mean that often the higher levels of the socio-technical system receive less attention. This means that when these causes are modelled using HFACS or the Swiss cheese model, some of the later slices or higher segments are less filled than the lower ones. This is an issue with what could be considered a bias within the larger data-set; individuals or equipment which is directly involved in an incident is often the focus, leaving larger issues to do with organisational culture or governing bodies unexplored to a larger degree.

This short review has highlighted a number of factors which must be considered with regard to this research. Firstly, there appears to be some overlap between the human factors identified in surface ship collisions, and those identified for collisions involving SMPs. These include a lack of knowledge about the other vessel involved in the collision, the OOW's decision-making and mental state, and poor situation awareness, as identified in (MAIB, 2004; Chauvin *et al.*, 2013), just as they are identified in (Marine Accident Investigation Branch, 2015, 2020) with regard to SMP collisions.

There is potentially a vast amount of data available for analysing surface ship collisions, through accident reports. Many incidents that are suspected of involving SMPs are never officially confirmed. Reports for accidents involving SMPs can take a long time to come out, sometimes many years, and on release are often heavily redacted for security reasons. This makes it difficult to create a large-scale model to perform some kind of qualitative analysis; the data-set is simply unavailable. There are also issues with creating such a large data set even when the information is readily available, as confounding or conflicting information can easily be introduced to the model.

As there is limited information available about SMP collisions, instead of trying to create a larger model to predict risks and trends, this thesis considers two recent SMP incidents, both involving RN SMPs, both with mis-classification listed as a cause, and examines and compares them in detail. Both

reports are compiled by the same source, and the methodology chosen to analyse the accidents allows them to be wholly visually mapped, with all causes connected. This is known as AcciMapping, and is discussed in more detail in Chapter Four.

This allows for a fuller picture of the data across the layers of the socio-technical system; it is less abstracted than the Swiss cheese model, making causality clearer, and tries to see the relationships between all of the different causes, rather than splitting them into different workflows as in HFACS. This will allow for a more thorough comparison and detailed analysis of the two chosen incidents.

4.3: Accident Analysis

An overview of the information requirements and transfer techniques has been established in the previous chapter. This information is used to build a tactical picture, the understanding of which has been shown to be distributed across the socio-technical system, with only the CO holding an overall picture, which is dynamically changing over time, and must be assimilated through constant reiteration, through interaction with a number of sub-systems and actors. The tactical picture is based on a number of information sources which contain inherent uncertainties. This is especially true for understanding the external environment; with this process relying on broadband sonar classification when operating below periscope depth. This process cannot be externally validated, and relies on expert analysis of hydrophone recordings, with only the bearing of contacts known. All other information about a contact is inferred from this bearing information. This can lead to errors, which then propagate through the system, and negatively impact the reliability of the tactical picture.

The research will now attempt to identify how these uncertainties can have a negative impact on the overall safety and operation of submersible maritime platforms.

This serves two purposes; to firstly understand where and how the uncertainties of the task can negatively affect the safety of the SMP; and to therefore highlight potential areas where theoretically autonomous support could mitigate these effects in the future.

To do this, two case studies of incidents reported on by the Marine Accident Investigation Branch (MAIB) were chosen for further analysis.

The incidents were analysed using a technique called AcciMapping. AcciMaps can be used to understand the causes of accidents and how they relate to each other, combining to create dangerous conditions. This is done by following Rasmussen's "systems approach" of accident analysis, where individuals are not blamed for error when accidents are provoked by systemic deficiencies and failures (Reason, 1995; Branford, Naikar and Hopkins, 2009)(Rasmussen, 1997).

Rasmussen developed this technique in 1997. The purpose of AcciMapping is to develop proactive risk-management strategies for complex socio-technical systems (Branford, Naikar and Hopkins, 2009). The technique can be used to better understand the causes of accidents and how they relate to each other. These identified conditions, when combined, could lead to danger.

Complex socio-technical systems involve a high degree of integration and coupling of information derived from multiple systems. Effects of a singular decision can have effects which propagate rapidly and widely throughout the entire system (Svedung and Rasmussen, 2002). This makes it important to understand the impact of decisions across the whole of the socio-technical system.

Accidents are not caused by the "coincidental alignment" of independent failures and human errors, but instead, can be seen to be allowed by a systemic migration of organisational behaviour. This can be especially true when there are added cost or competitive benefits from operating at the edge of the usual, accepted practice (Svedung and Rasmussen, 2002). In other words, the structure of the system, and the risks that are sometimes taken to maximise the advantageousness of performance, are the reasons accidents occur, not at an individual level of actors within the system, who are simply working in a way in which the system as a whole allows them to.

The AcciMap is split into six levels representing increasing levels of abstraction of the socio-technical system, going from physical processes at the bottom, to laws and regulations at the top. Nodes are then connected that show how causes across the levels are related – either sequentially, or by task.

This is particularly useful in understanding how causes can occur as side-effects of decisions made at different points in time by different actors, distributed at different levels of the socio-technical system. The activities can be functionally disconnected, with only the accidents revealing their relational structure (Svedung and Rasmussen, 2002).

The AcciMap is aimed at improving the design of systems, rather than as an allocation of responsibility. It can allow for representative identification of factors which are sensitive to improvement within the system as a whole.

4.4: Incident Selection

The two incidents discussed were selected from accident analysis reports produced by the MAIB. The MAIB is an official government body in the U.K. which investigates marine accidents that involve U.K. vessels worldwide, or which occur in U.K. territorial waters.

The specific incidents discussed in this chapter were chosen because they were officially documented by the MAIB with a full, publicly available report available. Both involved RN SMPs. Inaccurate contact solutions were direct causes of both incidents.

4.5: *Karen* Incident Overview

The first incident involves a collision which occurred in the Irish Sea in 2015 between an RN SMP and a fishing vessel named *Karen*. The RN SMP snagged the fishing gear of *Karen*, submerging the vessel underwater. The crew of the *Karen* was unharmed. The primary cause of the collision was the mis-classification of *Karen* as a merchant vessel instead of a fishing vessel, which meant the TMA solution derived for *Karen* was incorrect. An incorrect TMA solution meant that the crew's prediction of *Karen*'s position in the water was incorrect (Marine Accident Investigation Branch, 2015).

This mis-classification occurred because of a lack of trawl noise – the noise trawling fishing nets make in the water, used in the hydrophone analysis conducted by SOs. This is an example of negative confirmation bias; the absence of a classifier being used to confirm a different classification. Many vessels were mis-classified as merchant vessels in the operational area, meaning the RN's guidance on fishing vessel avoidance was not followed, as there was a lack of awareness of how many fishing vessels were operating in the area.

The RN SMP was dived, which meant that no RADAR, Automatic Identification System (AIS) information, or external views were available to verify the positioning of contacts in the area.

There was also a heavy concentration of contacts in the area, and the SMP was operating at a high speed, meaning contacts had to be classified relatively quickly. This may have meant that attention was spread over a large number of tasks, meaning there was little opportunity to investigate the incorrect classifications further, by producing more data tracks, and refining contact solutions.

Another cause for the collision was that the Command team had suspended the requirement for Close Quarter procedures, in an attempt to normalise passing close to merchant vessel contacts. However, there were at least two fishing vessels within 4000 yards of the SMP at the time of the collision. Because of the mis-classification, the crew of the SMP were not aware of the risk the close proximity contacts were creating, as they were perceived to be merchant vessels, and so snagging could not occur.

4.5.1: AcciMapping Methodology

The MAIB report was used as the primary source for information from which the AcciMap was visualised. This contained all of the factual information available about the accident, and had been compiled from evidence obtained directly from the RN and from the crew of *Karen*, as well as admiralty charts, plots and charts obtained directly from technical systems onboard *Karen*, and raw AIS and Vessel Monitoring System (VMS) information. Therefore it was considered to be a thorough and all-encompassing source to use as the basis for the diagram.

The AcciMap was first physically conducted on a wall using post-it notes, allowing for the actions, goals, beliefs and regulations to be moved around across the different abstract layers of the map.

Firstly, key actors were identified through a "first pass" of the report, identifying actors who performed actions before, during, and after the incident. These actions were all recorded on separate post-it notes.

Secondly, the report provided a sequence of events which occurred before, during, and after the collision. These were also broken down into individual events and actions, as well as contextual goals, and recorded on separate post-it notes.

Actions (e.g.: "abandoned CQ procedures", "course changed 15 degrees") and events ("no trawl noise was heard", "operating 'fast and deep'", "no fishing vessels identified in pre-mission planning") that occurred directly before, during and after the incident were then arranged across the bottom four layers of the AcciMap, as they described the physical processes and activities, equipment, and surroundings, and their operational management. These were linked together to understand a rough path of causation.

The causes of the accident were discussed with reference to this sequence of events, and also with regard to legislation ("COLREGs"), best practices ("fishing vessel avoidance guidelines"), and rules of conduct for both shipping vessels and SMPs. These were all recorded separately on post-it notes. The best practices, rules of conduct, and legislation governing the situation were plotted over the higher levels of the AcciMap, representing government policies, regulating bodies and associations, and local planning regulations. These were then linked to the causes as defined in the report.

More abstract causes for the accident, such as beliefs ("belief that FV always made trawl noise"), goals and needs ("need to remain covert") and expectations ("expected to maintain radio silence") were then identified through a third pass, written out on post-it notes and added across the layers of the AcciMap depending on where and how they applied to the incident.

Sections of the report referring to previous accidents, and recommendations, were not included in the passes made in the report, as they contained information outside of the scope of the incident.

The diagram was built on a wall, and then a digital copy of the diagram was recreated. Analysis started from the bottom layers of the AcciMap, linking immediate causes together, and then moved upwards, tracing the effect of the legislations, beliefs, and planning activities. The completed diagram can be seen in Section 4.4.2.

The same procedure was followed for creating the diagram of the second incident, seen in Section 4.6.1. This methodology followed methodologies for other AcciMap analyses (Tabibzadeh and Meshkati, 2015; Hamim *et al.*, 2019; Banks, Plant and Stanton, 2020) and was also based on guidance found in (Rasmussen and Svedung, 2000).

4.5.2: AcciMap for the *Karen* Incident

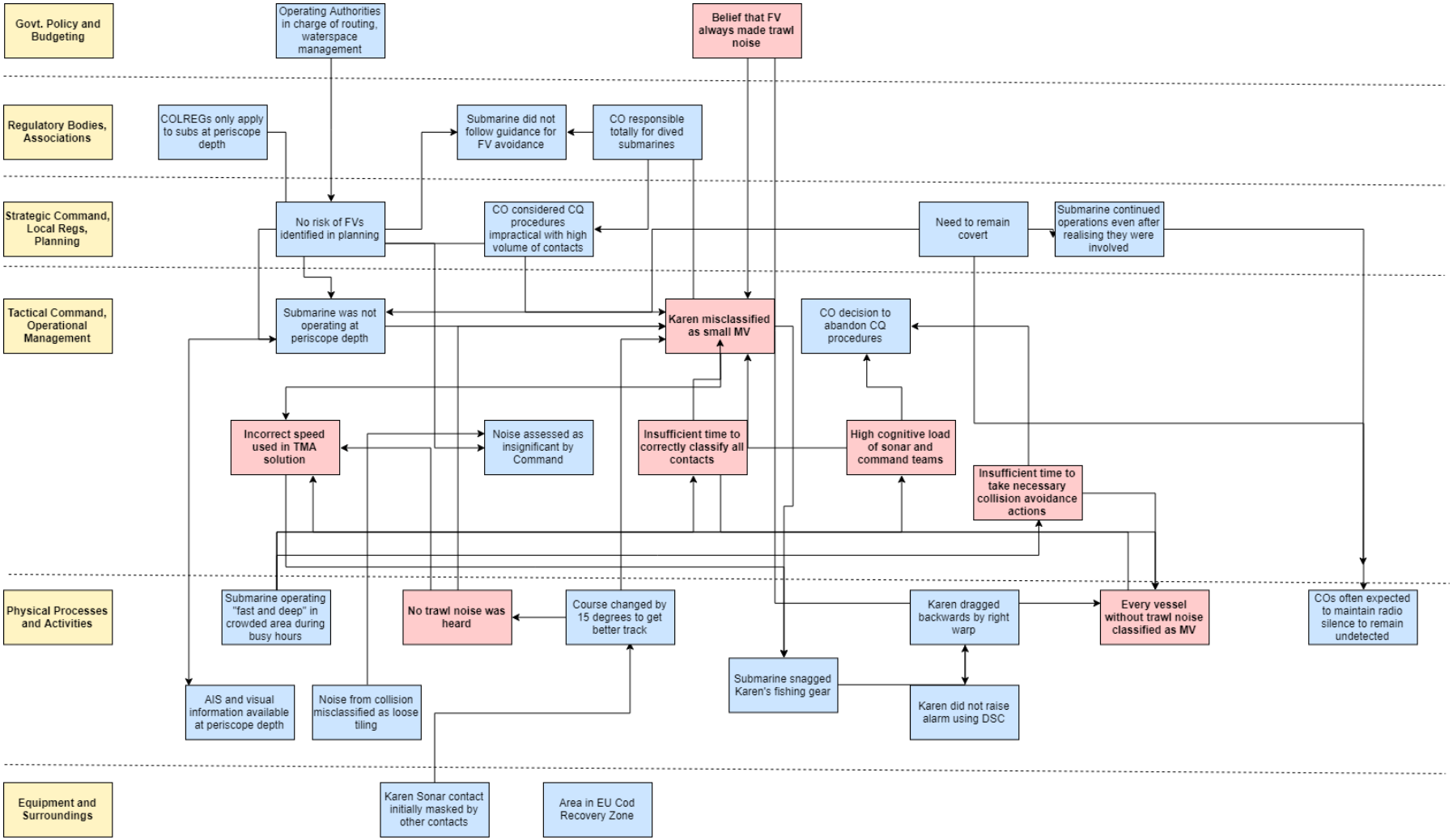


Figure 7: The AcciMap produced for the collision between an RN SMP and fishing vessel Karen in 2015

4.6: *Karen* Incident Discussion

Looking at Figure 7, reasons for the accident can be seen spanning across multiple layers of the socio-technical system. Key reasons for the accident have been coloured in red.

An integral contributor to the accident was the mis-classification of *Karen*, as can be seen on the central node, which impacts many other nodes of the AcciMap. Insufficient time, and heavy cognitive load, both contribute to this mis-classification. These causes are impactful because, coupled with the lack of trawl noise identified for *Karen*, they reduce the possibility of collecting further data tracks.

It is only through the iterative process of collecting more information on a contact that the ambiguous solutions are able to converge to reduce uncertainty around the solution, as stressed in the CDM interviews, and also in the literature (Dominguez, Long, Miller, Wiggins, *et al.*, 2006; Schunn, Kirschenbaum and Trafton, 2013; Loft *et al.*, 2016).

The lack of trawl noise being used to discount the possibility of *Karen* being a fishing vessel is particularly pertinent, for a number of reasons. This confirmation bias is a breakdown of logic. By their very nature, fishing vessels do not always produce trawl noise; in fact, this noise is linked to one very specific circumstance, trawling nets. Fishing vessels are known to be particularly dangerous because of their behaviour; often engaging and disengaging their engines, and moving and stopping to trawl for fish.

The expectation, or lack thereof, of finding fishing vessels in the operational area highlights a key area of informational need influencing safe passage. All vessels at sea have a responsibility to be transmitting information about their vessel using either AIS or Vessel Monitoring System (VMS) via radio transceiver and satellite respectively.

AIS is an automatic tracking system communication technology, with AIS information supplementing marine radar to provide collision avoidance for water transport. It provides information such as

unique identifiers, position, course and speed, which can be displayed using electronic chart displays or on a screen (Roberts *et al.*, 2004). This information allows maritime authorities to track and monitor vessel movements, and is required by maritime law to be fitted aboard voyaging ships with 300 or more gross tonnage, and all passenger ships regardless of size (*AIS Transponders*, no date).

VMS is a more general term which describes systems used in commercial fishing which allow regulatory organisations to track and monitor the activity of fishing vessels. It describes the specific application of monitoring commercial fishing boats. Different VMS systems employ various communication technologies, depending on VMS initiatives at national and regional levels.

The Global Fishing Watch initiative combines publicly available AIS information and integrates it with VMS information made available through governmental partnerships globally (Global Fishing Watch and Dicaprio, 2018). They use machine learning to combine AIS tracking and radar, optical and night-time imagery, and vessel registries, to produce publicly available visualisations of shipping activity around the world, both current and historic (Global Fishing Watch, 2019). Resources such as Global Fishing Watch could be useful for during pre-mission planning stages. Although this information cannot be accessed by the SMP crew when operating below periscope depth, it could certainly be used to enhance predictions around what traffic to expect, at what times, and where.

4.7: *Stena Superfast* Incident Overview

The second incident chosen for analysis was a near-miss between an RN SMP and a ro-ro ferry, *Stena Superfast VII,* in the North Channel in 2018.

The crew of the SMP were conducting pre-deployment safety training, with experienced members of the Flag Officer Sea Training (FOST) organisation on board. The SMP was operating at periscope depth to facilitate training exercises. This meant that they had more tools at their disposal to assess the operational environment, including the periscope, RADAR system, and AIS information broadcasts, although these were not clear, due to the short length of the periscope meaning there was interference, and so AIS information was intermittent.

A trainee periscope Watchkeeper reported sighting a new surface contact with an estimated range of 9000-10000 yards, which was identified as a ferry via the periscope. This information was input into the Submarine Command System (SMCS), and therefore was the information used by the Command team when making navigational decisions. None of the more experienced Watchmen checked the range estimate.

The OOW estimated the speed of the ferry to be 15kts. It was travelling at an actual speed closer to 21kts, causing a large disparity between the predicted time to CPA and the actual time to CPA. The periscope watchkeeper advised the OOW that the ferry was closing and heading towards the SMP. The OOW gave an order to turn the SMP to port, thinking this would move them out of the ferry's predicted path.

The Sonar team were tracking the ferry, and recognised the bearing was steady, despite the evasive manoeuvre. The sound room tried to initiate a Close Quarters (CQ) procedure. At the same time, the periscope was spotted by the OOW aboard the *Stena Superfast*, who realised there was imminent risk of collision, and applied the port rudder to increase CPA from the periscope.

The CO of the SMP, upon awareness of the ferry's turn to port, cancelled the CQ procedure and directed the OOW to remain at periscope depth, rather than going deep to avoid collision. They did not acknowledge that the ferry's change of course was evasive action, assuming instead it was course correction. Therefore the near miss was not acknowledged on board the SMP.

Once *Stena Superfast* had passed clear of the SMP, the master notified Belfast coastguard that the SMP's periscope had passed down the starboard side of the ferry at a range of 50-100m.

4.7.1: *Stena Superfast VII* Incident AcciMap

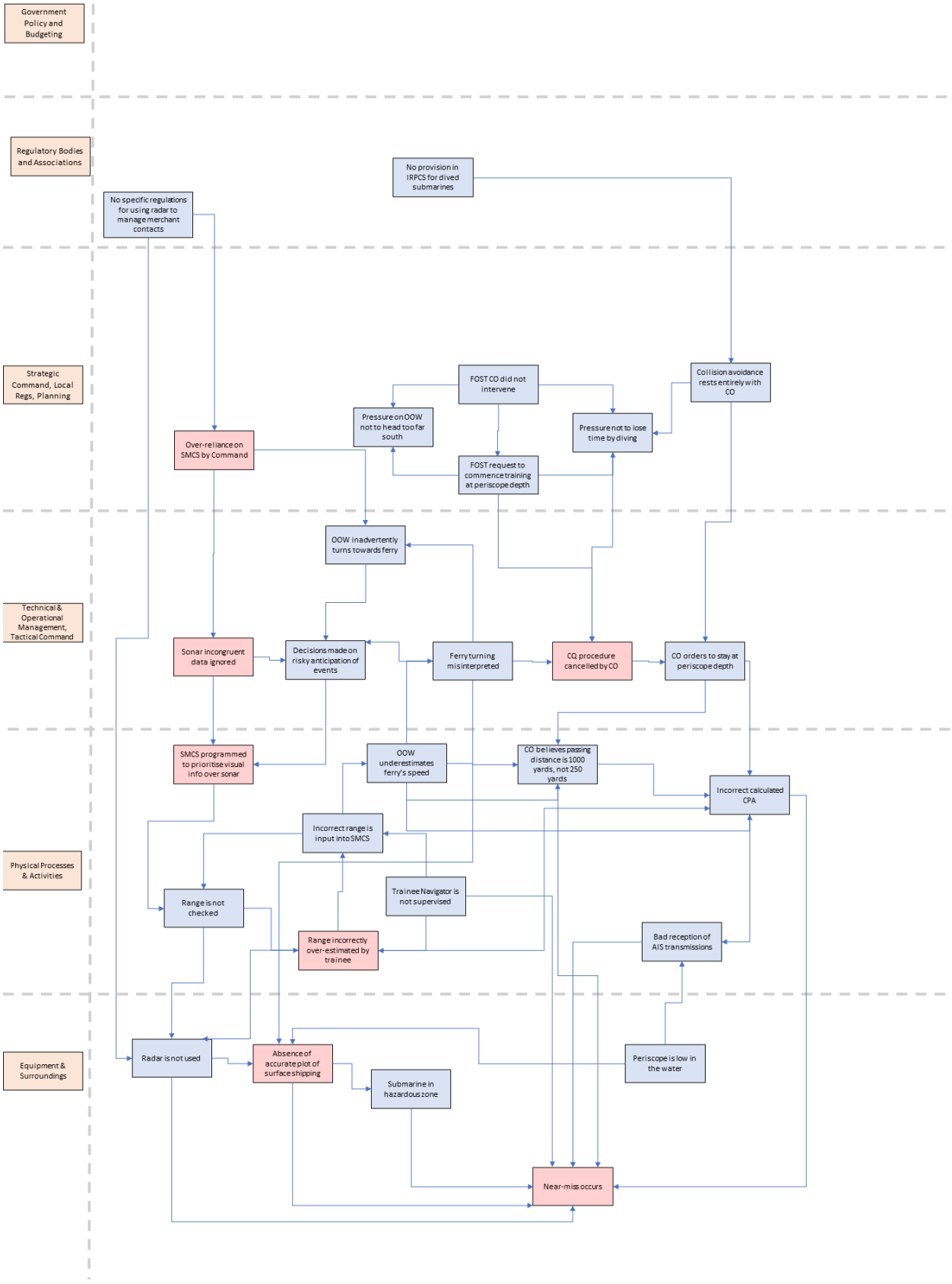


Figure 8: AcciMap for Near-Miss with Stenna Superfast VII

4.7.2: *Stena Superfast VII Incident* Discussion

The near-miss between the *Stena Superfast VII* and the RN SMP had multiple causes, which can be seen in Figure 8. The initial range estimation of the ferry was inaccurate, and this, combined with an underestimation of the ferry's speed by the OOW, meant there was significantly less time to take avoiding action than was calculated by the OOW. This, coupled with the turn towards the ferry, created a very real risk of collision.

This example shows how inaccurate information can propagate. The incorrect range estimation led to inaccurate speed and CPA calculations, all of which combined to create a very likely possibility of collision.

More accurate bearing information was available to the SMP's command team from passive sonar and periscope camera. Command relied instead on the SMCS data, which was exclusively based on the inaccurate initial estimated range data, to make their navigational decisions.

The sonar characteristics for the ferry would have presented contradictory evidence of the ferry's range, however, this was ignored. Sonar Operators had detected a potentially dangerous situation and had called a CQ procedure. However, command decisions were made instead with the SMCS information, which presented a safer picture than in actuality, and so the decision was made to remain at periscope depth, even though the situation warranted a dive for collision avoidance.

When the ferry was observed to alter course to port, this information was used as a "reinforcing bias", fuelling the assumption that the ferry was altering course to regain its original planned track.

When faced with incongruent information which did not align with the environmental picture presented by SMCS, a safer decision would have been to reassess options with consideration to the contradictory sources of information.

There was also pressure to commence training at periscope depth that afternoon, which may have impacted the decision to cancel the CQ procedure, as well as the inaccurate representation of

distance to CPA and under-estimation of speed, all based on the estimated, and incorrect, range used in SMCS.

Radar not being used also prevented a more accurate assessment of surface vessels, and a correction of the estimated range.

The SMP was positioned in a hazardous zone, with very lively shipping traffic, and the MAIB assert that this was clear when viewing historical AIS data for the location. Although the OOW may have thought they were clear of the ferry lanes, this was not actually the case: what they saw as a thin line of traffic to be avoided was actually a much wider corridor. This incident again highlights how more visualisation and study of historical shipping information could help to obtain a more accurate picture of an operational environment. Both this incident, and the *Karen* incident show an absence of a realistic plot of surface shipping which contributed to poor navigational decision making.

4.7.3: Findings

The accident analyses present a number of observations which could influence what autonomous support could be introduced to aid with tasks, where, and why. This helps to answer both RQ1 and RQ2, identifying what kind of autonomy would be useful, and how its introduction could prevent the same kind of incidents from re-occurring.

Firstly, an incomplete understanding of the surface picture, with reduced awareness of what surface vessels were operating in the area, contributed to both the collision and the near miss. This happened when operating at periscope depth, even with multiple sensor systems available to assess the environmental picture, and also when operating below periscope depth.

The *Karen* accident occurred in a busy fishing area, and yet the crew seemed unaware of how likely they were to encounter fishing vessels, and the *Stena Superfast VII* incident occurred in a hazardous area, but command did not appreciate this, believing they were not navigating in the shipping lane, underestimating its size.

This indicates that there is scope to develop a more thorough understanding of expected surface contacts both in the pre-mission planning stages, and through improved accuracy of plotting shipping routes and fishing areas. Historical AIS information could have been utilised to better understand hazards in the operational environment for both of the incidents described. This highlights a key area where increased autonomy could be beneficial for safe navigation, through the form of analysis of historical AIS and VMS information to better understand patterns of life and to analyse safer routes. Big data and machine learning can be used to accurately visualise this information, as shown by projects such as Global Fishing Watch, and could help to optimise routes by analysing historic traffic information.

The way in which sonar information was treated was different across both incidents. Whereas for the *Karen* collision, there were high levels of ambiguity for contact classifications, with many fishing vessels being mis-classified as small merchant vessels, during the near miss incident, the sound room were aware that the ferry was much closer to the SMP than Command believed, and even went as far as to initiate a CQ procedure, which was overruled.

This highlights that the inherent uncertainty around contact classification can be damaging in different ways. When not enough information is present to confirm a classification, this should be re-evaluated, as in the *Karen* accident, where mis-classification occurred. When information sources conflict, or are incongruent, as in the near miss, with SMCS and sonar data representing two distinct and conflicting solutions for a contact's behaviour, this information should be re-integrated, so that understanding is recalibrated. Both incidents show a need for better management of information uncertainty.

Although there is always going to be uncertainty when relying on a tactical picture generated through sensor systems and not through direct observation, better ways to evaluate and mitigate this uncertainty are needed.

Other key points to note are that better facilitation of sharing information from the sound room and the control room may have helped to make the disparity between the interpretations of *Stenna*'s movements clearer to command, in the near miss situation. This was highlighted in the last chapter also.

Reducing the heavy cognitive demand in a high-contact situation could also be of benefit, as highlighted by the *Karen* incident. This provides examples of additional area where highlighting information to an SO would be useful, such as to make particularly ambiguous contacts clearer to operators; for example, vessels that could potentially be fishing vessels, even if trawl noise is absent.

It is clear that sonar classification is integral to developing understanding of the tactical picture, and the mis-use, or dis-use of this information can create unsafe situations, as illustrated by the two case studies explored above.

This provides more evidence that the specific task of broadband sonar classification could benefit from autonomous support, and gives examples of what kind and level of autonomy would be useful, helping to answer RQ1 and RQ2.

The following chapter now builds on this, by developing an understanding of the task from a user-centric perspective.

CHAPTER 5: UNDERSTANDING BROADBAND SONAR ANALYSIS DECISION-MAKING

Parts of this chapter were prepared using the published work "Classifying Vessels Using Broadband Sonar: Considerations for Future Autonomous Support". Section 5.4 contains text from this publication.

5.1: Aims of Chapter

This chapter aims to build on Chapter Three and Chapter Four to develop understanding of broadband sonar analysis. Chapters Three and Four gave a context-driven overview of the task, and how it is performed and used within the wider socio-technical system. They identified how crucial the task of classification is to building an accurate and useful tactical picture, and what can happen when this goes wrong.

This chapter analyses the tasks from the perspective of an SO, to derive specific user requirements to answer RQ1:

RQ1: What Level and Type of Autonomy could be suitably applicable to tasks carried out in the process of broadband sonar classification whilst maintaining appropriate levels of trust in the automation?

This chapter begins to answer RQ3 and RQ4:

RQ3: How do Sonar Operators cognitively classify sounds?

RQ4: How can an autonomous decision aid visually present a credible, understandable, and trustable explanation behind its decisions?

This is done by developing understanding of the decision-making tactics employed by an SO, and how the decision-making process can be best supported by an autonomous decision-aid. The

cognitive process of classification is explored through an interview where the SO presents a scenario in detail, which is then further analysed to understand the cognitive processes at play.

5.2: Introduction to the Task

If all knowledge and processes could be broken down into stringent sets of logical rules and algorithmic processes, humans would have no use onboard an SMP; they would be perfectly capable of running themselves. While computer systems are able to follow logical rules and process large amounts of information easily, they lack imaginative and contextual thinking, human experience, and wisdom. Or what Klein terms "tacit knowledge" (Klein, Calderwood and Macgregor, 1989).

Explicit knowledge is objective, rational, and technical. It has a rigid structure, with fixed content, and is independent of context. It can be easily documented and externalised, and is therefore easy to share, transfer and teach.

Tacit knowledge is subjective, cognitive, and personal. It can be context specific, is highly internalised, and is difficult to capture and share. It is explicitly human, and is built up through experience and personal wisdom.

Critical Decision Method interviews have been used in domains such as fire-fighting (Klein and Klinger, 1991), weather prediction (Klein, 2009), and Naval command and control (Kaempf *et al.*, 2006). They focus on a method which tries to capture decision makers' experience with situations involving dynamic and continually changing conditions, and their real-time reactions to these changes (Klein, Calderwood and Macgregor, 1989).

They can be used to try and gain insight into how experts gain situation awareness, referring to the "state of knowledge of perception of elements within the environment, the comprehension of their meaning, and the understanding of their anticipated status in the near future" (M. R. Endsley, 1995; Kaempf *et al.*, 2006).

The CDM interview involves using questions which act as cognitive probes (Klein, Calderwood and Macgregor, 1989), to try and understand the cognitive decision-making processes that experts employ to carry out tasks.

The cognitive probes used in the interview examine goals, cues employed, expectancies and courses of action, whether considered or taken. They also try and search for errors, either committed in the task by the operator, or hypothetical ones, to try and derive the strategies which can overcome them.

Kaempf suggests that decision makers in command-and-control settings use “feature matching” and “story-building” strategies in order to build SA (Kaempf *et al.*, 2006). Feature matching is diagnostic; many incidents have clear and limited feature-sets, which have meaning in the decision-maker’s specific domain. Story building is a diagnostic strategy where decision makers mentally simulate to construct a story of how a state of events may have been caused. When there is not enough information to trigger specific recognition of a situation, the decision maker fits the disparate pieces of information they do have together to try and construct a coherent explanation.

The way in which these decisions are made fits the Recognition Primed Decision model (Klein and Calderwood, 2015). This is a model of decision making, which explains that experts, rather than playing out different courses of action in their head and choosing the best one, identify cues in the moment which relate to their previous experiences and classify the situation based on how they have dealt with it before. They found only 4% of the strategies used in decision-making with Naval officers involved comparing and contrasting courses of action. The decision making usually focused on developing situation awareness. This can be true even when the use-case is difficult.

Klein posits that it is through experience that decision makers gain the ability to evaluate a single course of action through mental simulation, instead of having to generate multiple hypothetical solutions and compare them. This could mean that decision-making aids should have different designs depending on the skill level of the user. Recognition primed decision making is also more

practical when under pressure, or experiencing time constraints, as it results in quicker decision-making.

By identifying how experts make decisions, this can provide insight into how best to design training and decision supports for to aid them in their tasks. This can give insights into the cognitive processes of task performance, and how to better support them through interface designs, instead of focusing solely on physical components and tasks.

5.3: Critical Decision Method Interview Technique

CDM is a semi-structured interview technique which is used to "investigate phenomena that rely on subtle perceptual cues, and assessments of rapidly changing events" (Kaempf *et al.*, 1996). It focuses on trying to elicit subtle aspects of domain knowledge and tasks. Probes focus on goals, hypotheticals, errors, and important cues. Instead of focusing on explicit knowledge, the CDM focuses on "critical incidents", memorable experiences which can be recalled with detail and accuracy. An interviewer tries to get an expert to focus on a previous difficult or challenging experience where they had to apply their skills. The interviewer then uses this experience as a framework to try and probe for decision strategies, pattern recognition, expectancies, errors, and environmental cues (Kaempf *et al.*, 1996).

SME interviews are used as a tool to gain understanding of complex social, technical and information systems, as well as informing suggestions of system design improvements (Barnes, 2003; Dominguez, Long, Miller and Wiggins, 2006; Kaempf *et al.*, 2006; Walker *et al.*, 2010).

CDM interviews are used to try and better understand tacit, as well as explicit, knowledge. They are used in the field of Naturalistic Decision Making to try and understand how experts make decisions or deal with high levels of uncertainty in their specific domains.

In order to better understand the highly cognitive task of contact classification, the ability to extract meaningful insights and classifications from sound recordings and their corresponding spectrograms,

a CDM interview was carried out with an SME. This participant was chosen because they had ample experience with broadband sonar classification, working both as a Sonar Operator and a Sonar Controller during their career, and so had expertise in the interpretation of hydrophone recordings.

5.4: Critical Decision Interview Methodology

The interview methodology is as follows, and has been adapted from (Klein, Calderwood and Macgregor, 1989).

First, a particular event is chosen to focus on during the interview. A scenario which involves challenging decision-making is optimal, usually in situations which are characterised by high time-pressure and high informational content, and are considered to be difficult by the decision-maker.

The interviewer asks for a brief description of the incident, and does not interrupt the participant, allowing them to explain the incident in their own words, in as much detail as they would like to offer.

Then, the scenario is discussed, and a timeline is created from the scenario in conjunction with the interviewee. This is known as the first pass. The timeline contains objectively verifiable information, the sequence and duration of each event reported, as well as thoughts and perceptions reported by the decision-maker. This creates a shared awareness of the facts, from the interviewee's perspective. This also allows for any missing facts or inconsistencies to be identified and corrected, or fleshed out.

Once the timeline has been constructed, the interviewer asks questions about specific areas of the timeline, which are designed to better understand the decision-making processes, known as cognitive probes. The questions require the decision-maker to reflect on their strategies and bases for decisions. Questions can vary in wording and timing, depending on the scenario that has been described, as interviewers have heard a description of the scenario before the probing begins. The

interview is semi-structured, and so has a lenient format, trying to allow natural dialogue to occur as much as possible, to keep the interviewee suitably engaged.

The interviewee is encouraged to draw diagrams to help explain various aspects of the scenario, such as how actors are geographically related to each other, or displays they were using to make the decisions. This gives more context to the situation, and also helps to refresh the decision-maker's memory.

Decision points are identified within the time-line, to understand why a specific course of action was taken at that point. Questions are asked to elicit details around the specific decision points. Probes about options are asked for each decision, including hypothetical ones.

Examples of different probes to be used in the interview process are shown in Table Four.

Table 4: Examples of Critical Decision Method interview probes, from Klein, Calderwood and Macgregor, 1989

Probe Type	**Probe Content**
Cues	What were you seeing, hearing, smelling?
Knowledge	What information did you use in making this decision, and how was it obtained?
Analogues	Were you reminded of any previous experience?
Goals	What were your specific goals at this time?
Options	What other courses of action were considered by or available to you?
Basis	How was this option selected? Why was this option rejected? What rules were being followed?
Experience	What specific training or experience was necessary or helpful in making this decision?
Aiding	If the decision was not the best, what training, knowledge or information could have helped?
Time Pressure	How much time pressure was involved in making this decision?

Situation Assessment	Imagine that you were being asked to describe the situation to a relief officer at this point, how would you summarise the situation?
Hypotheticals	If a key feature of the situation had been different, what difference would it have made in your decision?

A complete set of probe questions and additional questions asked in the CDM interview are provided in Appendix A.

The interview was carried out with an ex-SO, who had also worked as an SC, in person at the University. The focus of the interview was to understand the tasks of classifying and tracking contacts using sonar, and how the inherent uncertainty in the task is managed and mitigated. The interview lasted around two hours. It was recorded, and then transcribed.

The situation chosen was a training exercise conducted in the Clyde straits, with several contacts identified in the area; seven merchant vessels, six fishing vessels, and a single warship.

During the interview, a "timeline" was constructed of the order of events, sequentially and temporally, concerning the actions taken during a specific situation. The timeline was built up during the interview through the first pass of questions, which asked for more detail about the situation, the displays used, the information used, and how this was communicated.

This was then examined in more detail and discussed with the SME. Additional details were then added, and further questioning took place to derive more details about the incident, and the actions taken, using cognitive probe questions.

The SME was eager to share a lot of information about the task, and often would deviate from the questioning to try and give more details about specific actions or protocols. Therefore, additional questions were asked in an attempt to keep the interview focused on the specific event being

discussed, and to go back and discuss specific areas in more detail, when the answers had deviated from this task.

After the interview, the transcription was then analysed for three specific areas of knowledge: Work Rules, relating to the rules and heuristics employed to carry out the task of sonar classification, Mission Objectives, relating to the motivation and reasoning behind the decision-making processes, and Fishing Vessels, identified as contacts which are particularly dangerous and have high levels of uncertainty because of their particular behaviours, such as stopping and starting engines, and staying still in the water, making them particularly challenging to track and classify.

It was then mapped over six layers of abstraction, using a modified AcciMapping technique, to understand the task and its various elements from a socio-technical perspective. This kept the original coding used in the analysis of the transcript of the CDM interview to better visualise how the three different knowledge domains interacted throughout the task, and how they relate to each other, across the levels of the system.

This was first made by going through all of the coded pieces of information identified in the transcript, and writing each piece of text on separate post-it notes, colour coded depending on whether it was a Mission Objective, Work Rule, or information about Fishing Vessels. The diagram was then constructed on a wall, to visually display at what level each of the coded actions and goals identified belong to within the socio-technical system. This was then used to create a digital version of the diagram, which can be seen in Section 4.6.

The diagram covered a lot of information, not only pertaining to the SO, but the interactions between the sound room and control room, the pieces of legislation which govern why some of the actions had taken place, and tasks which occur before and after the actual mission, in order to give context to the task.

The diagram allowed for a specific set of actions and interactions to be identified and isolated, which pertained to the SO, their equipment, goals, and tasks. This was then used to create a timeline of the actions and decisions made by the SO, in conjunction with the timeline derived during the interview. This can be seen in Section 4.7. This was then used to focus on recommendations for autonomous support for the role of an SO, and better inform how the role could be aided with the introduction of autonomous systems.

5.5: Task Description Derived from the CDM Interview

Before a mission can begin, preliminary planning must be conducted. This involves planning a timeline of events, using shipping timetables, intelligence, weather reports and other external intelligence sources to plan a route. Other vessels expected to be in the operational area are plotted on charts, and exit points are established. This planning can take several days, depending on how complex the mission is.

Any information pertinent to the mission must then be learnt, and drills and safety procedures are practiced in advance of the mission starting. Roles must be practiced in relation the Mission Objective (MO). Crew will re-familiarise themselves with emergency procedures.

Once at sea, the SO's main duty is to detect, monitor and classify all contacts. They do this by looking for emerging contacts on-screen. They use a cursor to highlight a potential contact and record what they can hear at that point.

SOs listen to, record, and analyse sounds, using aural characteristics and frequency patterns to identify points of interest in ambient noise (Roberts and Stanton, 2018b). When an SO perceives a point of interest either over their headset or on their display, they begin to look for patterns and characteristics in the sound which could provide an acoustic signature and allow them to derive a classification. The SO will analyse the sound, beginning by working out an estimated engine RPM, allowing them to calculate an estimated speed, when combined with a contact's bearing. The only truths they have are the frequency characteristics of the sound, and the contact bearing; course,

speed and range must be derived from these two key variables, when there is no visual information available (O.Gibson, 2007).

Once a SO has identified a contact's bearing, and attempted to derive the RPM and speed, they make an initial classification of a contact, identifying what kind of vessel it could be, such as a military vessel, fishing vessel, or merchant vessel, for example. Signature sounds, such as "shaft rub" (*"Some of the couplings might be bent, so every time that shaft is rotating around a single point it's going to rub against something. It's like a womp...Womp... Womp. So we know that, and we can get a count off of that…*"), can give indications about how well-engineered the components are of the contact, which can be used to help determine further details of the classification. The SO then alerts the SC of their initial classification of the new contact.

The SC acts as an informational filter, passing relevant information between the Sound Room, via the SOs; and the Command team, via the OPSO and OOW. The OPSO then assesses the classification, and uses external information sources such as shipping charts and timetables, and other sensor information, to build up a profile of the contact. This additional information is used to try and determine more characteristics of the contact, such as its course and range. This information is then passed back to the SC, and the classification is confirmed.

The contact is assigned a tote, is labelled, and added to the SMCS display. It is assigned a contact number, speed, bearing and range.

Passive Sonar systems perform signal analysis such as DEMON and LOFAR. LOFAR estimates vibration, and DEMON analysis allows SOs to extract propeller features, such as the number of blades and shafts (Mello, Moura and Seixas, 2018). These engine features are extrapolated from analysis of the frequency characteristics of a sound on a DEMON display. The characteristics, when combined, provide a unique identifier for a contact, which can allow it to be classified.

To strengthen understanding of the contact, the SO records multiple data tracks, capturing the sound of the contact, which can be sent for TMA. SOs' main duties, as identified in the interview, are to identify, classify and monitor all potential contacts. As repeated data cuts are made, a more accurate TMA solution can be calculated. This can be used to identify the contact's direction, speed, and movement in relation to own ship, to evaluate any potential risk.

Once a contact has been classified, the OOW and OPSO first work out if it is closing or opening – approaching, or moving away from, own-ship. They use all available information to determine what the CPA for the contact should be. The CPA defines a safe distance which must be maintained between own ship and the contact. If a contact enters the CPA, creating a danger of collision, an SO must call what is known as a Close Quarters (CQ) procedure, requiring Command to assess the situation and advise what action should be taken to avoid collision, which could involve diving the SMP to a safer distance away from the contact. Time to CPA is calculated.

The SO must continue to detect, classify, and monitor any emerging or displayed contacts. They iteratively update the bearing of the contact, with this information passed to the SC, who communicates it to the OPSO.

If the contact moves close to the CPA, the SO alerts the SC, who alerts the OPSO, who must then decide whether to begin a CQ procedure. The SMP will then be navigated to safety, and the contacts will need to be re-classified.

Reliability of the contact solution is improved by collecting more data tracks over time. Getting clean, useful data cuts may require the SMP to change course, allowing additional or improved data collection. This decision is made by the OOW, who has to balance maintaining a safe distance from all contacts, positioning the SMP for optimum information acquisition, remaining undetected, and fulfilling mission objectives. The OOW must do this whilst building and updating their 3D mental picture of the operational environment.

Sometimes one contact will obscure another. This can be dealt with by cutting a data track, and monitoring where the contact should reappear based on the TMA solution. In some instances, narrowband sonar can also be used to supplement the picture, and isolate frequencies of interest. When the contact emerges, the aural information can be used to confirm it is the same contact as before, and it will be re-classified. If the contact does not re-emerge, the OOW may decide to alter course to ensure safety, and to try and make the contact visible again.

If a contact suddenly appears, identified in the interview as an “abrupt start”, the frequency band it occupies must be identified. This can be a worrying situation, as fishing vessels often cut their engines whilst fishing. Sometimes, it can be predictable, for example, when close to a port. But when there is potential for the contact to be a fishing vessel, extreme caution must be maintained.

5.5.1: Additional Discussion of Task

From this description of the task, it can be seen that there are multiple inflection points which can introduce uncertainty into the classification process.

The task itself is inherently uncertain, as the variables for course, range and speed must be derived, with only bearing and frequency information for the hydrophone recordings being known. It is difficult to externally validate a contact solution, as often there are no external views available, and very limited data transfer options underwater to communicate and verify the contact solutions.

This uncertainty builds over three layers which all interact with each other. Firstly, basic sensor data is inherently uncertain due to the nature of passive sonar. Sound transmitted through water is affected by temperature, pressure and salinity (Schunn, Kirschenbaum and Trafton, 2013). Ambient noise from the ocean itself, waves, creatures, large vessels, can all create additional noise which can mask and obscure sound sources. High background noise and the high similarity of man-made noises can make it difficult to detect subtle differences in the ambient sound.

On top of this, TMA is inherently uncertain, employing algorithms to extrapolate course, speed and range from bearing direction and rate. A TMA solution is more of a line of best fit, with multiple viable solutions for a contact that can be extrapolated. Being able to distinguish which TMA algorithm to use requires careful consideration and experience.

The third layer of uncertainty is caused by the unpredictability of human action, which can be deliberate in the form of deception. Managing this uncertainty, when tracking many contacts at once, is a gargantuan task.

Information is passed verbally from the SO to the SC, and then this must be conveyed to the Command team. Other studies have identified the SC as a bottleneck for the flow of information in the system (Pope, Stanton and Roberts, 2019), causing the largest disparity in information between them during their trials (Stanton *et al.*, 2020a). The information requirements of the two teams should mean they are interdependent processes and require close informational coupling.

As well as this, limited time to classify contacts, or a high volume of contacts to classify in an operational area, can also impede on an SO's ability to thoroughly verify all contact solutions.

During the interview, fishing vessels were described as "the most dangerous contact". Fishing vessels often will not adhere to the designated fishing zones, and behave in an unpredictable manner, stopping and starting engines to fish. This unpredictable behaviour was discussed at length in the interview. As sonar analysis relies on a contact making noise, when an engine noise ceases, a contact can be "crossed", leaving its position unknown.

Knowledge of fishing areas, shipping routes and timetables are all used to understand a contact's suspected position and classification, and when vessels operate outside of these regulated areas, it becomes harder to predict their movement and actions. This inconsistent behaviour sets fishing vessels apart from, for example, a merchant vessel using auto-pilot to travel in an established shipping lane. Fishing vessels stop and start engines, and raise and lower nets. They can change

direction frequently, and do not follow predictable patterns of movement, making them volatile, with high levels of uncertainty surrounding them.

Fishing vessels have tell-tale characteristics which can help distinguish them from other vessel types. One that was mentioned multiple times in the interview was the presence of “trawl noise”, referring to the sound of clanking chains as fishing nets are trawled through the water.

Trawl noise was highlighted as important to identify quickly, as snagging fishing nets is a real danger, which can be fatal. This was stressed many times, showing it is an important feature which signals danger in the classification process.

A situation where a contact did not present trawl noise, but had similar characteristics as a fishing vessel, were probed many times in the interview. Secondary aural characteristics to help distinguish between small merchant vessels and fishing vessels were mentioned, but trawl noise was stressed to be a main distinguishing factor. This is an unreliable classifier in isolation, because of the discussed unpredictability of fishing vessel contacts; they are not always trawling. This was established in the case study of the *Karen* incident.

5.6: Visualisation of the Tasks Associated with Classification

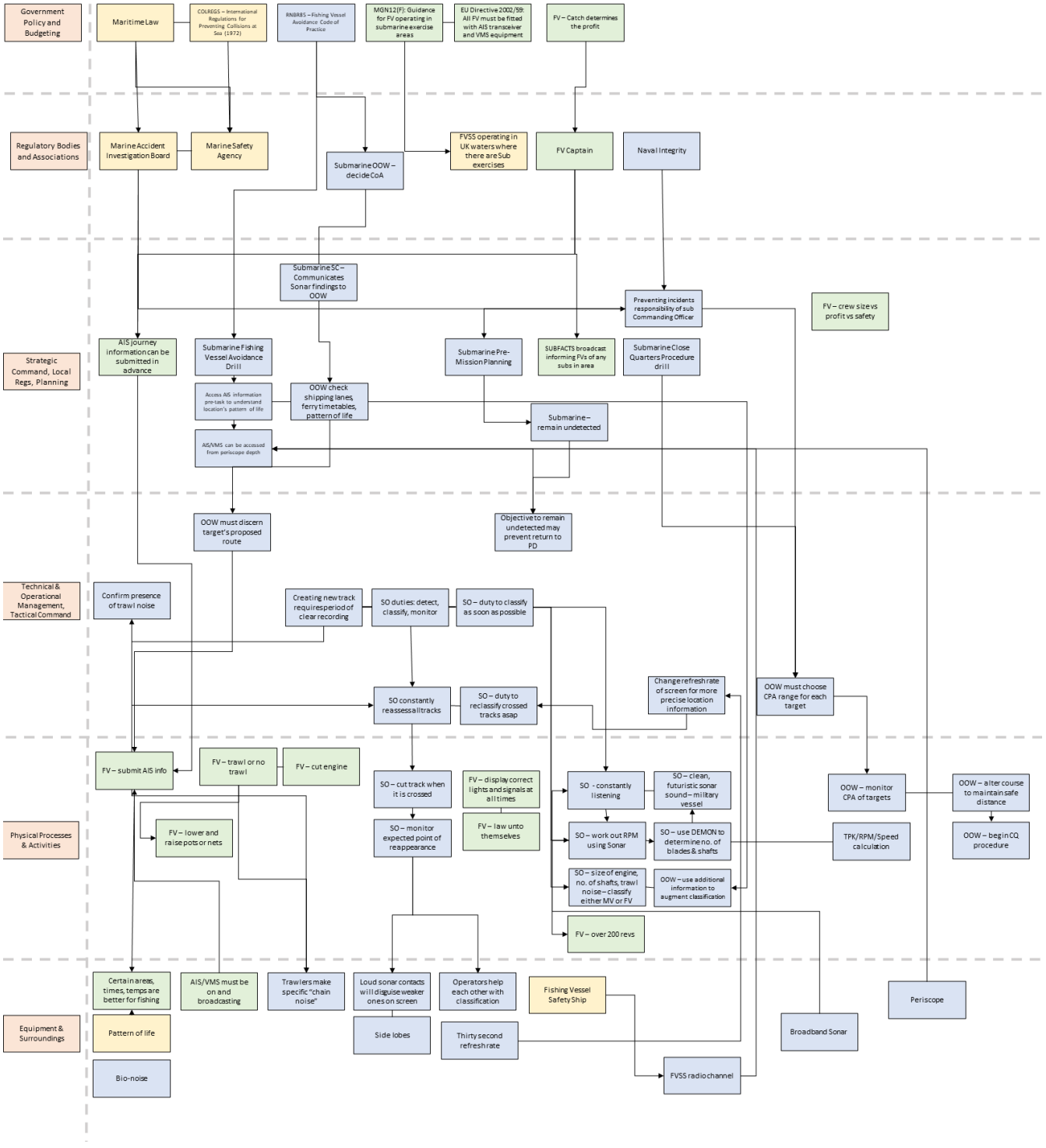


Figure 9: A modified AcciMap for the tasks associated with classification derived from the SME CDM interview

5.7: Timeline and High-Level Task Description Derived from Critical Decision Method Interview

A high-level task description was derived from the CDM interview. This was an attempt to understand how the tasks and duties of the SO are performed; in what order, whether they repeat, and what information is used. This was helpful in identifying specific areas of tasks which could benefit from the introduction of autonomy.

1. **Planning Meeting**
    - Work out a timeline of events
        - Pattern of life – shipping lanes, fishing areas, timetables, intelligence, environmental conditions
    - Establish exit points
    - Plans for dangerous scenarios established – what if, exit plans
    - Predicted behaviour and capability of target discussed
    - Plot other vessels in the area on chart: know what to expect

2. **Preparation**
    - Practice role in relation to MO
    - Learn information relevant to MO
    - CQ drills
    - Emergency procedure refresh/practice/drills: sme95, fishing vessel avoidance

3. **Start Event**

4. **Detect and classify all contacts**

    4a: Classify contact

- Spot track on screen
- Scroll cursor to it
- Listen
    - **Begin initial aural classification**
        - **Use chronograph and DEMON and SONAR to work out vessel speed, engine characteristics**
            - Count revs over six seconds – rough shaft RPM
            - DEMON - Demodulation analysis information – no. of blades, no. of shafts, rev count
            - From this, work out target's speed for TMA
        - Shaft rub – blade count, rev count, how well shafts have been engineered indication of whether vessel is commercial or military
        - Blade slap
        - Diesel noise
        - Engine noise
        - Transience
        - Trawl noise
        - In-out
    - **Classify contact**

4b: Create data track

- Pass new data track to OPSO
- OPSO assesses classification:
    - Uses intelligence, shipping charts, shipping timetables to build up contact profile
    - Work out contact course and range
    - Communicate further information to SC

- **Classification is confirmed**
- Track tote is added to SMC display
    - Classified – colour
    - Assigned a number
    - Speed
    - Bearing
    - Range

4c: Begin TMA solution (OPSO, OOW)

- Determine whether vessel is closing/opening (direction)
- If possible, work out range
    - Define Closest Point of Approach (CPA) OOW
    - Determine when vessel will reach CPA
    - Determine vessel's course, determine SMP's course

Monitor all tracks

- Recursive 4a – constantly reviewing contacts and checking direction, speed
- Detect and classify new contacts
    - Has it changed direction?
    - Has the speed increased?

If bearing rate of contact changes:

- Update contact location and speed info
- SC pass updated info to OPSO
- OPSO decides whether to alter course/call CQ depending on position

If contact moves unpredictably:

- Increase screen update rate
- Isolate contact, focus on contact – may be assigned individual operator
    - SC monitors situation and communicates with OPSO

If contact moves close to CPA:

- Show to SC
- SC reports to OPSO
    - OPSO makes CQ decision
        - CQ procedure begins
    - OPSO may order manoeuvre to ensure safety
        - Reclassify contacts

If contact is obscured by another contact:

- Cut track
- Monitor where contact should reappear based on TMA and predicted movement
- Use narrowband sonar to supplement picture and isolate frequencies of interest
    - Look for associated frequency lines

When contact reappears

- Recut data track
    - Use aural classification to reclassify contacts

Or

When contact does not reappear

- OOW may alter course to ensure safety and make contacts visible
    - Reclassify contacts

If contact abruptly appears, "abrupt start"

- Identify what frequency band it is occupying
    - May be predictable – abrupt starts coming and going from port, for example
- Monitor thickness of line
    - **If suspected FV, THEN:**
        - Trawler identification

5.8: Level and Type of Autonomy Suitable for the Task

The recommendations are considered with regard to automation types (Parasuraman, Thomas B Sheridan and Wickens, 2000a) and degrees of automation (Onnasch *et al.*, 2014a) with applicability to the high-level tasks elicited from the CDM interview shown in the task description, as well as points considered particularly perilous, as highlighted in the interview.

Maintaining situation awareness is of vital importance for contact classification, therefore, the degree of autonomy should be selected with regards to the potential penalties to high-performance automation on the human performance; over-reliance, complacency, and reduced SA (see literature review). **Therefore it is posited that only low LoA and information type automation should be recommended, in order to prevent the trade-off in performance, and to ensure that the "raw" tactical picture is never obscured from the operator, in case of automation failure or erroneous behaviour**.

5.8.1: Information Acquisition

Automation could be useful for highlighting (so as not to obscure the tactical picture):

- New contacts emerging on the display
- When a contact is moving erroneously compared with its solution

- The most likely database entries for TPK
- Areas where a lost contact should re-emerge
- The re-emergence of a lost contact
- Contacts nearing CPA
- A poorly fit TMA emerging

At a higher level, it could prioritise and display:

- Most likely classifications based on speed
- Potential points of emergence for lost contacts
- Cuts that need to be merged
- An unpredictably moving contact
- Contact approaching CPA
- A change in bearing rate

5.8.2: Information Analysis

Automation could project:

- Potential solutions (TMA)
- Fixed points for solutions (TMA)
- CPA

5.8.3: Decision Selection

Although selection appropriateness could be contested because of the safety-critical nature of the task, and the trade-offs in performance if reliable automation fails, it could be beneficial for some of the SC's duties to be assisted through decision selection. One way decision selection automation could be used is by passing information directly from the SO to the OPSO. For example, if a solution is complete. If the SC has a high cognitive load, automating this process could prevent an informational bottleneck. This informational bottleneck has been observed in the literature, where

the SC was not efficiently passing all of the information on in a timely manner, and some things were missed completely

5.8.4: Action Implementation

This could involve:

- Automatically classifying contacts when classification ambiguity is low, and offering these as suggestions
- Deriving speed for contacts, and offering these as suggestions
- Fixing points in TMA solutions, to reduce manual manipulations of the solutions

However, action implementation automation is not recommended, because of the trade-offs described in Chapter 2. This is because, in an always ambiguous environment, 100% performance cannot be guaranteed. This is confounded by the idea that some targets could be implying deceptive technologies to make themselves harder to identify. For this reason, it seems unsafe to recommend more direct forms of automation. It is also prudent to assume that complacency should be prevented, especially when things can change very rapidly; what may appear to be a safe situation could very quickly degrade if contacts stop-start, re-emerge in new places, or a trawler is identified, all scenarios which were identified as particularly uncertain in the CDM interview.

In this sense, action automation should only be carried out alongside an SO's own classification decisions. Automation should not take over the responsibility for classification, but instead, should offer a comparative classification. This also supports the findings of Chapter Two, in that working interdependently on classification as a HAT will optimise trust and performance.

5.9: Conclusion

This chapter has identified a number of specific recommendations for autonomous support in order to aid SOs with broadband sonar classification, helping to answer RQ1. It identifies that low level automation is more appropriate, in order to prevent any obfuscation of real world data from the SO.

It identifies a number of key areas where information could be highlighted, or projected, on current displays. It also provides further justification for why a human is needed in-the-loop, and how autonomy should be integrated in such a way that the Operator and the autonomy can work together interdependently.

It builds on recommendations and informational needs derived from Chapter Three and Chapter Four to offer specific examples of how autonomy could support them, whilst doing so in a way which supports appropriate levels of trust in the autonomy, as outlined in Chapter Two.

Now that concrete ideas for autonomous support have been developed through assessing user requirements, the next chapter will develop an understanding of what the autonomous support could look like to optimise human-autonomy team performance and trust-calibration.

CHAPTER 6: DISPLAY DEVELOPMENT

Parts of this chapter were prepared using the published work "Feature Perception in Broadband Sonar Analysis – Using the Repertory Grid to Elicit Interface Designs to Support Human-Autonomy Teaming". Section 6.4 contains tables from this publication. There is a patent application pending for the method and designs presented in this chapter.

6.1: Repertory Grid Study

In order to design a display which supports the decision-making processes behind contact classification, it was important to establish a better understanding of the mental processes utilised to obtain a classification. The previous chapter identified a number of areas of the task which could benefit from some form of autonomous support, and highlighted the importance of any classification aide being easily understood by an operator, explaining its decisions, in order to facilitate appropriate levels of trust in the information it could provide. In order to be explainable, an autonomous system would need to provide an understandable explanation behind its decision-making processes.

A better understanding of what aural information is used, and how, in the classification process, could provide insight into ways an autonomous system can explain its reasoning to a user in a way which is meaningful to their understanding, and useful for quick evaluation of the classification decision.

It has been established that in order to trust information from an autonomous system, performance is improved when the user has an understanding of – or an explanation for - where that information came from, and the reliability of that information. Transparency is key to trust, as discussed in Chapter Two, and so any uncertainty around a classification must be brought to the attention of a user, as discussed in Chapters Three and Five.

Machine learning classification systems often use classifiers which are very abstract, and not easily understood by humans, sometimes even the humans that created the system. This means there is a disparity between what information the human operator would need to be able to accurately assess the efficacy of a classification decision, and the capability of a machine learning classifier to provide that information. This is integral to establishing trustworthy relationships with autonomous systems, and becomes a sticking point in the introduction of higher-level autonomy. Providing further explanation could even go some way towards making it easier to spot when high-accuracy is behaving erroneously; it would be possible for a human to spot if an explanation does not make sense, for example, potentially adding an additional opportunity for intervention.

Even if a classification system provides the user with a confidence percentage for its classification decisions, this is not enough information to provide a solid foundation of trust. A high confidence percentage would imply that the contact classification decision would be accurate, but with no knowledge of what information was used and how, it is impossible to assess this.

A small study was developed using the Repertory Grid interview technique developed by psychoanalyst Kelley in 1955 (Curtis *et al.*, 2008). The repertory grid technique is a “cognitive mapping technique”, which can be used to understand how “internal representations of subjects’ environments are constructed” (Curtis *et al.*, 2008).

The repertory grid elicits ratings on dichotomous constructs to build a model of cognition. It is an interview technique designed to understand how individuals construct their knowledge, a knowledge acquisition technique. It is borne from "personal construct theory", developed by Kelly. It is built on the epistemological premise of "constructive alternativism", based upon constructivism, which posits that reality does not directly reveal itself to us, but is instead subject to the different constructions we invent (Ford and Bradshaw, 1993).

The questions this study hoped to answer were:

- How do Sonar Operators mentally classify sounds?
- How do their techniques or skills differ from those of other sound professionals and novices?
- How can an autonomous decision aid visually present a credible, understandable, and trustable explanation behind its decisions?
- How might such a presentation differ depending on the role and experience of the end users?

Answering these questions provides solutions to RQ3 and RQ4, developing understanding of how the cognitive classification process is carried out by SOs, in order to inform a display design which can explain its decision-making in a way in which the user can intuitively understand.

6.2: Methodology

DSTL kindly provided a set of hydrophone recordings to be used in this study. From the set, five were selected, each representing a different kind of contact, with different sized vessels, some containing aural features which are used in the classification process. These are summarised in Table 1 below.

Table 5: A table listing the hydrophone recordings used in the repertory grid study

Element Number	**Description**
1	Medium Merchant Vessel
2	Small Merchant Vessel
3	Large Merchant Vessel characterised by “blade slap”, meaning the vessel is light or empty. The propeller of the vessel is not fully submerged as it rides high in the water.
4	Large Merchant Vessel characterised by “shaft rub”, which is the sound of poorly machined propeller shafts, or worn bearings
5	Fishing Vessel. Characteristic “trawl noise”, the sound of trawling fishing nets; heard as tinkling or clunking sounds as the bobbins and chains contact the seabed.

7 second clips were created of the recordings using Logic Pro. From these clips, colour spectrograms were produced, showing frequency information and intensity over time for the recordings.

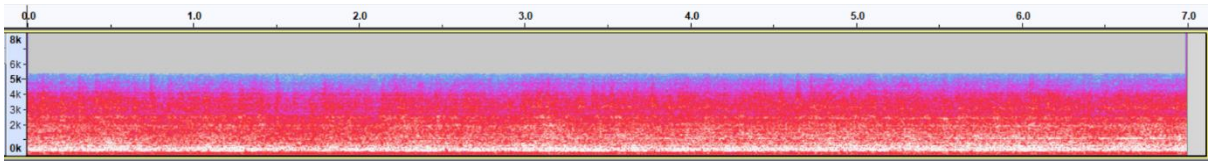


Figure 10: An example of a spectrogram produced for the Medium Merchant Vessel hydrophone recording

In order to better understand the effect which expertise has on the cognitive classification process, different sound professionals were used as participants for this study.

Twelve participants performed the study in total. Five novice participants were used, as well as four sound engineers, with an average of five years' experience of recording, mixing and making music. SME participants were also utilised; an ex-SO, an ex-OOW, and a current RN training professional, who teaches the art of sonar analysis to new recruits at HMS Colingwood, a Sonar Trainer (ST). All of the SME participants had experience performing sonar analysis in a professional capacity. The sound engineer group had no experience of sonar analysis, but did have experience in recording and production of live and recorded sounds in a professional capacity.

The study was repeated with both aural information, the hydrophone recordings, and visual information, using the spectrograms, with each participant.

The sounds (and their corresponding spectrograms) were divided into ten "triads"; unique groupings of three recordings. A triad of sounds would be played to the participant. The participant was then asked to use a word or phrase to describe how two sounds were the same, and distinct from, a third sound. These words or phrases became the "construct" column in the repertory grid.

Once all ten triads had been listened to, producing ten constructs, the participant was then asked to provide an opposing word or phrase for each of the constructs. These did not have to be the literal opposite of the words identified as constructs, but instead, something the participant

conceptualised as an opposing phrase. Therefore, ten bi-polar pairs of constructs and contrasts were derived.

Once the ten bi-polar pairs were established, the participant was played the five sounds individually, and asked if each sound belonged more to the "construct" phrase, or more to the "contrast" phrase for each pair.

This process was then repeated, utilising the spectrograms instead of the hydrophone recordings, therefore each participant produced a repertory grid for aural information, and a repertory grid for visual information. An example of a completed repertory grid can be seen below.

	Element Number						
Construct	**1**	**2**	**3**	**4**	**5**	**Contrast**	**Fla**
Diesel	1	1	0	1	1	Quiet	3
In-out	1	1	0	0	0	Consistent	5
Blade flutter	1	1	1	1	0	Blade slap	3
Whine	1	0	0	1	0	Hum	3
Engine	1	0	0	1	1	Muffled	2
Diesel Engine	1	1	1	0	1	Steam generator	3
Flutter	1	1	0	1	1	Compressed cavitation	3
Cavitation	0	1	1	0	0	Shaft	3
In-out	1	1	0	1	0	Consistent	4
Blade flutter	1	1	1	1	0	Blade slap	3
Total	9	8	4	7	4		
Template	**1**	**1**	**0**	**0**	**0**		

Figure 11: An Example of a Complete Repertory Grid Performed with an SO SME Using Hydrophone Recordings.

The methodology shown in (Baber, 1996, 2015) was used to elicit the cognitive constructs from the repertory grid responses. This is a form of principle component analysis. The responses were split into two roughly equal groups by picking an arbitrary number, allowing the constructs to be translated into a smaller number of groupings, which explain the maximum possible variance (Stanton *et al.*, 2013). Columns below the chosen value were assigned as a "0", and above were labelled as a "1" in the template. The template was then compared to each row, with matching values in the columns totalled and provided in the "fla" column.

When there were 4 or 5 positive matches for a construct, this was taken as a "concept", a grouping of constructs which were related conceptually.

The numbers in the “fla” column were then examined. Any emergent groupings, with the same value in the column, were considered to be their own cognitive concept. These were then removed from the grid, and the process was repeated, until all of the cognitive concepts had been defined.

Looking at Figure 11, it can be seen that there is an emerging concept containing the pairs labelled “in/out, consistent”, to do with the motion of the ship in the water. This became a cognitive concept known as “movement”, and these rows were removed from the table, where the process was then repeated again.

	Element Number						
Construct	**1**	**2**	**3**	**4**	**5**	**Contrast**	**Fla**
Diesel	1	1	0	1	1	Quiet	4
Blade flutter	1	1	1	1	0	Blade slap	4
Whine	1	0	0	1	0	Hum	4
Engine	1	0	0	1	1	Muffled	3
Diesel Engine	1	1	1	0	1	Steam generator	2
Flutter	1	1	0	1	1	Compressed cavitation	4
Cavitation	0	1	1	0	0	Shaft	2
Blade flutter	1	1	1	1	0	Blade slap	4
Total	7	6	4	6	4		
Template	**1**	**1**	**0**	**1**	**0**		

Figure 12: With the first concepts removed, another grouping begins to emerge

All constructs were grouped and labelled as different concepts, showing a unique area of focus for each SME participant.

Table 6: The cognitive concepts derived from the repertory grid, the constructs within them, and an explanation of the concept

Concept	Construct	Contrast	Explanation of Concept Term
MOTION	In-out	Consistent	Relating to the motion of a ship in the water, and the qualities of the noise made as propellers break the surface
SIGNATURE	Diesel	Quiet	Encompassing a range of sounds relating to engine type, vessel's depth in the water, whether the vessel is empty or heavy
	Blade flutter	Blade slap	
	Whine	Hum	
	Flutter	Compressed Cavitation	
LOW CLARITY	Muffled	Engine	Relating to the quality of the sound of an engine
	Cavitation	Shaft	
MECHANISM	Diesel engine	Steam generator	Relating to the type of generator (diesel v steam)

These aural and visual concepts were then displayed in a grid shape. The grid was then coloured depending on how many constructs were present within the concept for each vessel.

Motion	**Signature**	**Clarity**
Mechanism	**Engine Configuration**	**Speed**
Engine Movement	**Propeller**	**Pattern**

Figure 13: A blank conceptual grid containing SO responses

The VINAS grid was then coloured depending on how many constructs were present for each concept. If no constructs were present, the colour was white. If some constructs were present for the concept, it was coloured yellow. If all constructs were present for the concept, it was coloured green. In this way, distinct, visual patterns were created on the grid for each hydrophone recording.

6.3: Discussion

As expected, a participant's unique skill-set and level of expertise affected how they interpreted the sounds and spectrograms. There were profound differences in the responses depending on whether the participant had any experience of sonar analysis. The cognitive concepts that were produced changed depending on the participants' experience of sound analysis.

The participant who had knowledge of sonar analysis had some similarities in their self-identified constructs and contrasts. There were some overlapping constructs. However, the conceptual groupings differed, even when some constructs were repeated, showing that different roles had different informational requirements and meanings behind each construct.

When interpreting the recordings, the group of sound engineers with no professional sonar analysis experience were interested in the recording quality of the clips themselves. They did not derive hidden meanings from the recordings, or identify what was recorded, but focused on the medium itself. They used words in their constructs such as clarity, pitch and rhythm, aligning with their

domain of expertise, which is to produce high-quality audio recordings. This analysis was quite superficial, in the sense that the sound engineers were unable to derive much meaning from the recordings – they identified mechanical sounds, for example, but never the mechanism which produced those sounds. The construct “mechanical” was elicited from 75% of the sound engineers. Participants who were familiar with sonar analysis always produced a construct describing the mechanism identified in the recordings.

Interestingly, the sound engineers produced constructs pertaining to frequency information when only listening to the aural recordings, as opposed to the visual stimuli, unlike any other group of participants.

The SO had starkly different conceptualisations when compared to the sound engineers. Their constructs, as opposed to focusing on sound quality, pitch and frequency, related back to engine characteristics, mechanisms and movement. The SO was able to identify an engine in every recording, and to a finer resolution when compared to the other SMEs. The SO picked out aural features pertaining to engines, mechanisms, propellers and shafts. This was conceptualised as “propulsion”, and is key to being able to construct an initial classification, by determining the unique engine configuration and aural signature of a vessel.

There was overlap between the responses of the SO and the responses of the ST, both noticing specific engine characteristics in the recordings, showing evidence for similarity in their cognitive processes when listening.

The SO identified “diesel” as a construct, contrasting with the word “quiet”, but also “diesel engine”, contrasted with “steam generator”. This shows two distinct uses of the word “diesel”; the former with consideration to safety, as it reveals the presence of a contact which could be moving, which was grouped with the concept of “motion”, and in the latter pair, identifying an engine mechanism, grouped within the concept “engine”.

Clarity was an important conceptualisation made by the SO, containing the bi-polar pair "muffled" and "engine", highlighting the importance of clear recordings for analysis, allowing for details pertinent to classification to become more distinguishable and clearer. This concept differed from the "clarity" conceptualisation identified by the sound engineer group, referring to the clarity of the signal itself and not to a particular aural marker.

Table 7: The derived concepts and their constructs for an OOW listening to hydrophone recordings

Concept	Construct	Contrast
Range	Diesel engine close range	Gas turbine long range
Sound Source	Engine sounds	Biological sounds
	Machinery	Organic
	Drive train	Oars
	Close range	Long range
	Machinery noise	Human noise
	Mechanical	Biological
Mechanism	Other internal systems	Propulsion chain
	Engine	Sails
Identify	Biological	Mechanical

Table 7 shows the conceptual groupings elicited from the OOW responses. The first grouping of the OOW concerned engine type and proximity, conceptualised as range. The OOW is responsible for

overall safety of the vessel, and is able to override navigational decisions to maintain the safety of a vessel (Royal Navy, 2017). The OOW is responsible for avoiding any collisions or risk of grounding. This means they are interested in the physical location of a contact in relation to their vessel, trying to create a three-dimensional mental model of their environment, therefore prioritising range in their conceptualisation, aligned with their primary role.

The OOW, like the SO, was interested in the presence of mechanisms, but to a less detailed degree; they did not identify specific propeller and shaft configurations, but instead focused on identifying the pattern of life, whether the sounds were mechanical, organic, or contained any noise from humans. Broadly, they were interested in whether a noise was derived from a biological or mechanical source, and whether it was close or far away. This makes sense when reviewing their main responsibilities of collision avoidance and safety. They require less granular information about what was actually used in the classification process, but more information pertaining to movement and locality in order to make effective navigational decisions.

Table 8: The derived concepts and their constructs for an ST listening to hydrophone recordings

Concept	Construct	Contrast
Signature	Flutter	Rasp
	Propulsion	Propeller
Motion	Rotation	Stationary
	Diesel	Steam
Mechanism	Diesel Engine	Steam Propulsion
Cavitation	Cavitation	No Cavitation

The ST had some conceptual overlaps when compared with the SO and the OOW (see Table 4). They focused mainly on engine configuration, similarly to the SO. If the SO makes a detailed assessment of configuration, and the OOW is mainly interested in a summative conceptualisation of the configuration, the ST has a middle ground between the two approaches, identifying some engine characteristics, but to a lesser degree of detail when compared to the SO. Both identified noise and characteristics caused by propellers in the water.

Cavitation was a distinct concept identified by the ST, again related to propellers, and also to speed. Engine movement concerned them, similarly to the SO, and the ST constructed the pair "rotation" and "stationary", conveying thought about safety implications for a ship that may still be in the water nearby. The ST's constructs identified specific markers and classifiers that lend themselves to training others how to identify and build up a classification based on the aural signature of the contact, in a similar way to the SO in their conceptualisation of "propulsion".

All SMEs conceptualised an engine type or mechanism, whether diesel or steam. They then deconstructed this mechanism to varying degrees dependent on their role.

With regards to the visual stimuli, the spectrograms, participants again generated distinctly different constructs. Sound engineers were concerned with the smoothness of the signal, and tried to identify whether there was a visual, a repeating pattern, either temporally or in a particular frequency band. The sound engineers described the visual pattern presented to them and did not infer any engine characteristics from the spectrograms.

The SO gained insight in different dimensions when presented with visualised frequency information. Their predominant conceptualisations concerned specific engine characteristics and speed. When they were presented with a way to visually interpret which frequency bands were present in the recordings, the SO built on their identification of specific engine parts by specifying the number of components. The ability to distinguish between specific configurations of propellers

and shafts, supplemented with aural interpretation, allows an SO to make an initial classification of a contact.

The SO further built on this mental model by inferring movement and speed, associated with constructs such as “engine RPM”, “engine firing rate”, and “shaft RPM”. This shows a fusion of visual and aural information in the classification process, utilising both to build a more detailed understanding of the structure of a contact.

The OOW interpreted visual information similarly to the sound engineers, with conceptualisations concerning the shape and rhythm of the presented signal, identifying patterns in its appearance. Like the SO, they gain a better understanding of movement by fusing together the visual and aural information, conceptualising whether rotation is occurring.

The ST identified visual patterns in the spectrograms, focusing on the signal’s behaviour in terms of intensity, diffusion, and fundamentals. They conceptualised the “pattern” of the signal, its “regularity”, “similarity” and “intensity”. This indicates a methodological understanding of the problem, concerned with the presence of a consistent, recognisable pattern, or whether there is obscuration through a lack of clarity, or dissimilarities compared with what they expect to see. The ST commented on the “absence of the normal, the presence of the abnormal” in the classification process, trying to identify incongruencies between what is displayed, and their classification conceptualisation, exemplifying a visual form of abductive reasoning, deriving a conclusion from observation, rather than seeking evidence to support an already formed hypothesis.

The ST, like the SO, conceptualised “speed”, using the construct/contrast pair “fast”, “slow”, gaining a new dimension when presented with the visual spectrogram, concerned with the contact’s movement.

6.4: Results

As shown in Figure 13, the concepts that were elicited through the repertory grid technique were visualised in a grid pattern.

The grid was then coloured according to how many “hits” for the constructs identified within the concepts were recorded for each vessel. This provided a quick, visual way to assess differences in how each vessel had been conceptualised. This is shown in Figure 14, with a grid developed for the role of the SO and using the SO and ST responses. The grid predominantly comprised concepts derived from the SO responses, with the inclusion of “pattern” from the ST, as this concept contained distinct information which could be used in frequency analysis.

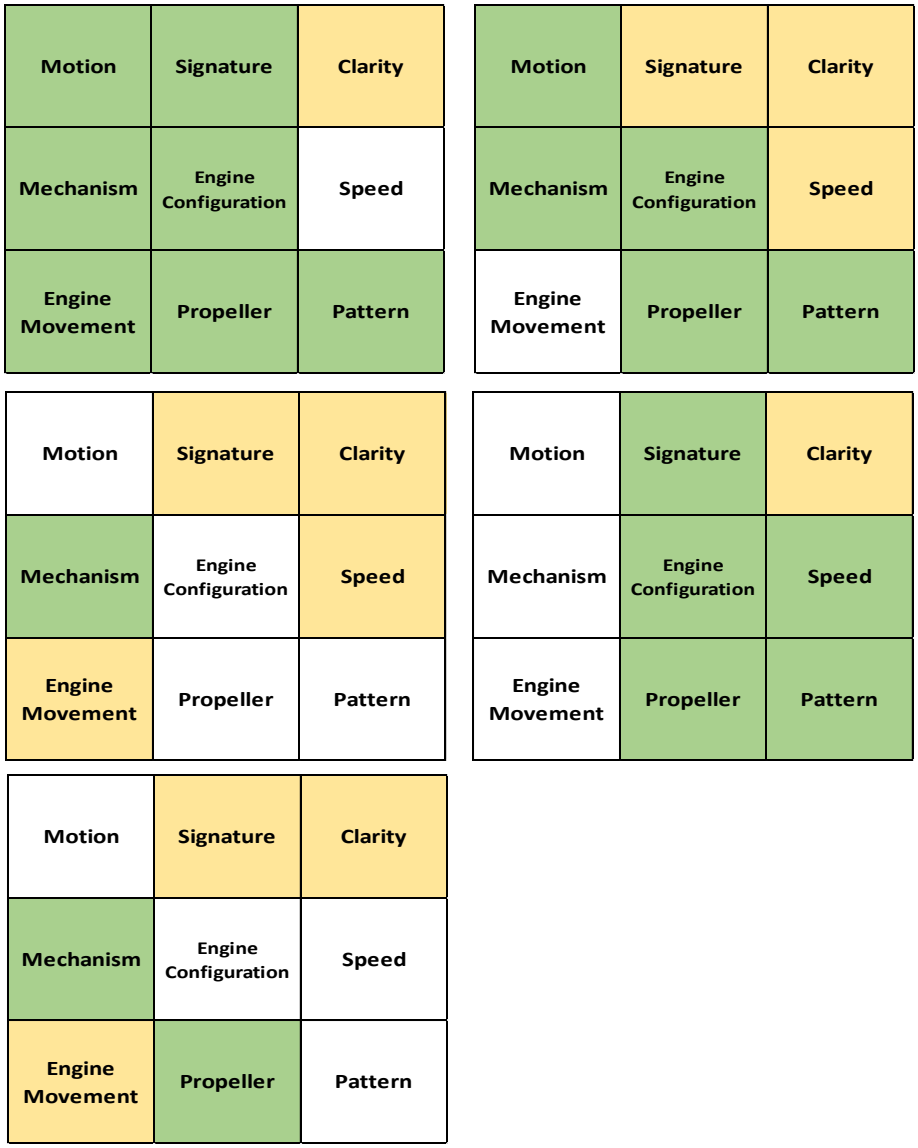


Figure 14: Examples of a VINAS Grid Using SO Responses, Coloured Distinctly for each Vessel Recording

From left to right, coloured VINAS grids are presented for: medium merchant vessel, small merchant vessel (top row), large merchant vessel with characteristic blade slap, large merchant vessel with characteristic shaft rub (middle row), fishing vessel trawling nets (bottom row). Concepts with all constructs present are coloured green. Concepts with some constructs present are coloured yellow. Concepts with no constructs present are not coloured. This created a distinct, visual pattern for each hydrophone recording depending on how many concepts were present for each construct.

As can be observed in Figure 14, each grid was distinctly coloured depending on the vessel being represented. Differences can be visually assessed quickly. Areas with few hits, indicating a lack of

information or ambiguity, are highlighted to an operator, and can be quickly identified and further explored.

In this way, an autonomous decision aid could use a conceptual grid to represent the narrative behind its own classification decisions. Coupled with confidence percentages for different classifications, the grid helps to highlight areas of potential ambiguity, abnormality, or distinction to an operator quickly.

The interpretation of the sounds was directly influenced by the role of the SME and all conceptualised their interpretations of the sounds differently. Therefore this difference in depth of analysis and informational requirements must be reflected in what information is displayed to them.

Both the SO and the ST, for instance, identify specific engine features and rely on these to provide evidence to support initial classification. Therefore an autonomous agent needs a way to signify if these features were observed in its analysis as well as how certain it is that those specific features were observed. This would provide an SO with a comparison to what they observed, a way to quickly spot an absence of a classifier they would expect, or the presence of an abnormality that affects their certainty in the classification. It also provides a starting point for further investigation when there are inconsistencies or contradictions in the classifiers highlighted.

The OOW does not require the same kind of display – they are interested in how close a contact is, whether it is dangerous, whether it is moving towards or away from them, whether it is a vessel or biological – noise associated with marine life or natural processes. For the OOW, information from sonar analysis is only one part of their tactical picture, and so the information they need is at a higher resolution. Command incorporates a larger wealth of information resources into their decision-making process when compared with the sonar team. Frequency characteristics are important for sonar analysis, but not so much for making navigational decisions.

The SO requires that frequency information in order to develop their classification, but the OOW needs to know how close the contact is, directional information and speed, in order to assess any danger presented by the situation and develop a new course of action. However, a VINAS grid could be elicited from the OOW responses and still differentiated between each vessel (see Figure 8).

Figure 15: An Example of VINAS Grids Coloured for Five Vessels from OOW Responses

Pictured in Figure 15 from left to right, are grids for: medium merchant vessel, small merchant vessel (top row), large merchant vessel with characteristic blade slap, large merchant vessel with characteristic shaft rub (middle row), fishing vessel trawling nets (bottom row). The grid was specifically made for the role of OOW, coloured using OOW repertory grid responses. Concepts with all constructs present are coloured green. Concepts with some constructs present are coloured

yellow. Concepts with no constructs present are not coloured. The five vessels can all be distinguished, with less variance compared to the VINAS created from SO responses.

6.6: Early Validation of Design

Early validation of the design was performed by comparing the concepts extracted to a checklist of key features used in classification, provided by HMS Collingwood. Nearly all of the concepts featured in the VINAS grids were represented in the classification checklist, showing high agreeability between the concepts used to train, and the concepts used to explain. A summary of those features can be seen in the table below.

Table 9: Matching key features from training manual to concepts derived through VINAS

Key Feature	**Related to VINAS Concept**
Cavitation	Clarity
Flutter	Signature
Shaft rub	Clarity
Number of blades	Signature
In-out	Motion
Turbine whine	Signature

6.7: VINAS Description

The VINAS grid provides a quick, visually assessable representation of an autonomous decision aid's classification decision. If different vessels make unique patterns on the VINAS grid, and an operator is familiar with the pattern they expect to see for a contact, it is easy for them to visually assess the viability of the autonomously generated contact solution.

If there are differences in the coloured squares compared with the ideal contact solution, this shows a region of uncertainty in the autonomously generated contact solution. By using the cognitive concepts derived from the repertory grid study, the confusion can be understood in terms of the

concept which is mis-coloured, giving an operator a starting point to explore and evaluate the classification decision.

Therefore, the VINAS attempts to provide a narrative behind the classification process, showing which concepts were considered in the classification process, and which concepts have some confusion surrounding them.

6.8: Requirements Evaluation Table

Table 10: Collation of user requirements defined within thesis and how VINAS meets them

Requirement	**Fulfilled by VINAS?**
A system must strive to be transparent (Muir, 1994)	VINAS attempts to show transparency by offering an explanation for its suggested contact solutions. It does this by providing a visual explanation through colour-coding how many constructs are present for each concept represented on the grid.
A system must be predictable (Muir, 1994)	VINAS allows operators to evaluate its contact solutions by providing a visual demonstration of any regions of uncertainty. Predicted pattern can be compared to emerging pattern. Concepts coloured differently than expected show a region of uncertainty around the classification solution.
The process behind a system must be transparent (Lee and See, 2004)	VINAS provides an explanation behind its decision-making to allow any discrepancies to be spotted easily
The purpose of a system must be transparent (Lee and See, 2004)	VINAS has a clear purpose, it shows a visual representation of a classification and tries to facilitate an explanation behind its decision-making

The performance of a system must be transparent (Lee and See, 2004)	Performance can be evaluated by comparing an operator's expected classification's pattern to the pattern which emerges on the VINAS; if an operator is familiar with the patterns they expect to see, this allows quick, visual evaluation of performance
The three Ps must be presented in a way the user can interpret and understand (Lee and See, 2004)	Once a user is familiar with the patterns they expect to see on VINAS, they can easily translate those patterns into classification decisions and understand any regions of uncertainty, quickly and visually
A system should be calibrated to be sensitive in order to minimise the chance of near misses (Rice, 2009)	From the colouring, it will be clear if VINAS is showing any discrepancy in the classification suggested and classification emerging
Human-in-the-loop activities must be maintained in order to ensure adequate SA (Onnasch *et al.*, 2014b)	Both human and VINAS work on classification. Classifications can be compared if required. Human is not removed from the loop. Decisions provide a visual explanation
Automation should be at least 70% accurate to be implemented (Wickens and Dixon, 2007)	VINAS presents a visually assessable way of evaluating classification accuracy, so an operator knows how reliable the automation's decisions are
Automation should not replace a human when doing a task; human and automation should work together as a team to accomplish a task	Both operator and VINAS can work together to classify. VINAS offers a comparison for classification decisions, which can help highlight regions of ambiguity or uncertainty to an operator

Information autonomy would be more prudent to implement rather than action autonomy in a safety-critical domain (Onnasch *et al.*, 2014b)	VINAS does not make a classification decision for an operator, but instead, presents a visual pattern for comparison, and so should not affect an operator's situational awareness, or induce complacency
Information autonomy should seek to highlight areas of a display to an operator, rather than removing any raw information (Parasuraman, Thomas B. Sheridan and Wickens, 2000)	VINAS does not remove any raw information, but may highlight any inconsistencies or regions of uncertainty in a classification
High levels of "team SA" should be encouraged through as much context and transparency behind automation's decisions as possible (Endsley and Kaber, 1999; Schaefer, Evans and Hill, 2015; Mercado *et al.*, 2016; Chen *et al.*, 2018)	Context and transparency is provided through explanation; how concepts are coloured can be interpreted depending on a human's expected classification solution; team SA is supported by providing explanation behind decisions which work towards a common goal.
Facilitate information flow between SOs and TMA Operators	Socio-technical system requirement, outside of the scope of SO performance

Enhanced understanding of surface picture	Socio-technical system requirement, outside of the scope of SO tasks
Minimise ambiguity in classification	VINAS can help highlight areas of ambiguity which need further investigation or cannot be easily confirmed, therefore aiding with the identification of ambiguity
Re-integration of information when it does not match over different sensor systems	VINAS can help highlight regions of uncertainty around a classification, gives a starting point for further investigation
Reduce cognitive load in heavy volume contact situations	If VINAS and operator classification have high levels of agreement, a solution can be accepted quickly through visual assessment
Highlight emerging contacts	Provides potential solution for emerging contacts
Highlight erroneously moving contact	Shows areas of ambiguity for erroneously moving contacts
Highlight most likely TPK database entry	Could help to identify when a solution and speed does not match through visual regions of uncertainty
Highlight area where lost contact should re-emerge	Does not meet this requirement; this requirement is for a different part of the display
Highlight contact nearing CPA	Does not meet this requirement
Prioritise most likely contact based on speed	Outside of scope of VINAS
Prioritise potential point of re-emergence	Outside of scope of VINAS
Prioritise cuts that need to be merged	Outside of scope of VINAS

Prioritise contact nearing CPA	Outside of scope of VINAS
Prioritise change in bearing rate	Outside of scope of VINAS
Project potential solution (TMA)	Different part of the socio-technical system
Pass information between SO/OPSO	Outside of scope of VINAS
Automatically classify contacts when ambiguity is low	VINAS does this by producing a pattern for a contact which is quickly, visually assessable
Derive speed for contact	VINAS could help identify which speed would be most appropriate by providing visualisation of a classification
Fix points in TMA solution	Requirement for different part of socio-technical system

Table Ten summarises the user requirements identified through each chapter of the thesis and offers an explanation of how VINAS supports, or does not support, the requirement.

Requirements defined through Chapter Two, to facilitate trust in a HAT, are generally met by the VINAS. Once the patterns on the grid are understood by an operator, and they know what pattern they expect to see depending on their own classification of a contact, VINAS will either support this classification, or highlight an ambiguity through a differently coloured construct. If a construct is coloured differently, it provides a starting point, or an area of ambiguity, around its contact solution, and an operator will be able to understand what area of the classification it is experiencing that uncertainty for. This gives a point of reference for re-assessing a classification, or highlighting an area where a classification may be ambiguous, and require more information before confirmation.

In this way, VINAS supports appropriate trust building in a number of ways; it makes its performance transparent to an operator; it is designed with a specific purpose; the process by which it derived a

contact solution is also made as transparent to an operator, by offering a visual explanation of what constructs identified in the cognitive classification process of the operator have been identified in the contact solution.

VINAS also supports requirements expressing a need for both a human and the autonomy to work together on a shared task. VINAS does not classify in place of an operator, but instead provides them with additional information with which to make classification decisions. It also does not hide any "raw" information from an operator, so does not negatively impact their SA. It supports the idea of team SA by providing a solution which is explained, which an operator's own assessment can be compared to.

Some of the requirements listed focus on different parts of the socio-technical system, such as TMA solutions, or facilitating information transfer. As the VINAS is a display which has been specifically designed to aid an SO in their task, these requirements do not apply to the VINAS display, and so it does not necessarily meet those requirements.

6.9: Conclusion

The VINAS design fulfils many of the user requirements identified in previous chapters of the thesis, by offering transparency and explanation behind its decision-making, and by offering support to an SO's classification tasks. This is summarised in Table Ten.

The VINAS attempts to provide a translation between the classifiers that could be used by a machine learning algorithm, which may not be very understandable to a human, into a visual display which uses classifiers derived from the human's classification process in order to provide an explanation behind its decision-making. Depending on how the squares of the VINAS grid are coloured, it is possible to identify areas of the classification which may warrant further investigation by the human Operator.

The VINAS grid itself provides a narrative behind an autonomous systems' decision-making process which can provide supporting evidence behind a classification decision.

The next steps, shown in Chapter Seven, are to verify that the VINAS does positively impact trust and performance during the classification process.

Another need which was identified during the accident case studies was support for when incongruent information is offered during the classification process, and so how VINAS is used in those situations is evaluated.

CHAPTER 7: DISPLAY TESTING

7.1: Introduction

Two experiments were designed to test how a VINAS would affect performance in a classification task when using a mock-autonomous classification aide. The experiments also attempted to understand what effect VINAS would have on the trust of an operator when utilising an autonomous classification aide.

Therefore, this chapter attempts to answer RQ5 and RQ6:

> RQ5: Can a Visual, Intelligent Narrative of Autonomous Systems (VINAS) improve performance in a classification task utilising an autonomous classifier?
>
> RQ6: Can a VINAS improve trust in an autonomous classifier when conducting a classification task?

For both of these experiments, five different hydrophone recordings were used, provided by DSTL. A description of each of the recordings can be seen in Table Eleven:

Table 11: An overview of the five sounds used in Experiments One and Two with descriptions

Recording	**Description**
01	This example is of dolphins and whales. The high-pitched cries and squeals are from dolphins, the low-pitched groans and grunts are from a humpback whale
02	Recording of a fishing vessel. Trawl noise, often heard as clinking and tinkling sounds as the bobbins and chains contact the seabed, can be heard
03	Large Merchant Vessel characterised by “blade slap”, meaning it is light/empty of cargo. The vessel rides high in the water with the propeller not fully submerged.
04	Small Merchant Vessel
05	Medium Merchant Vessel

The recordings were all clipped to seven seconds in length.

A spectrogram for each of the seven second recordings was produced in audio software.

7.2: Experimental Interface

The different features of the experimental testbed can be seen in Figure 16 (see next page).

At the top of the screen, a key for interpreting the different patterns presented on the VINAS grid was always shown.

Underneath this is a spectrogram of one of the recordings used. This spectrogram, and the sound itself, always match and are always correct, representing a ground truth. These stimuli match the stimuli used in the repertory grid experimentation shown in Chapter Seven.

Below the spectrogram on the left-hand side was a VINAS grid produced for each sound. In Experiment One, this was absent for half of the conditions.

To the right of the VINAS was a bar chart showing suggested classifications, with a confidence percentage generated by the system. In the bottom right corner is a suggested classification for the sound, which the participant can accept or reject by clicking the respective button. The suggested classification matched the classification on the bar chart with the highest confidence percentage. This visualisation has been used for explaining the confidence of different developments of autonomous classifiers. However, it does not offer any explanation for its confidence, and so this research posits it does little to provide appropriate trust calibration in the classifier's decision-making abilities.

In the middle of the screen were controls for playing the audio recording.

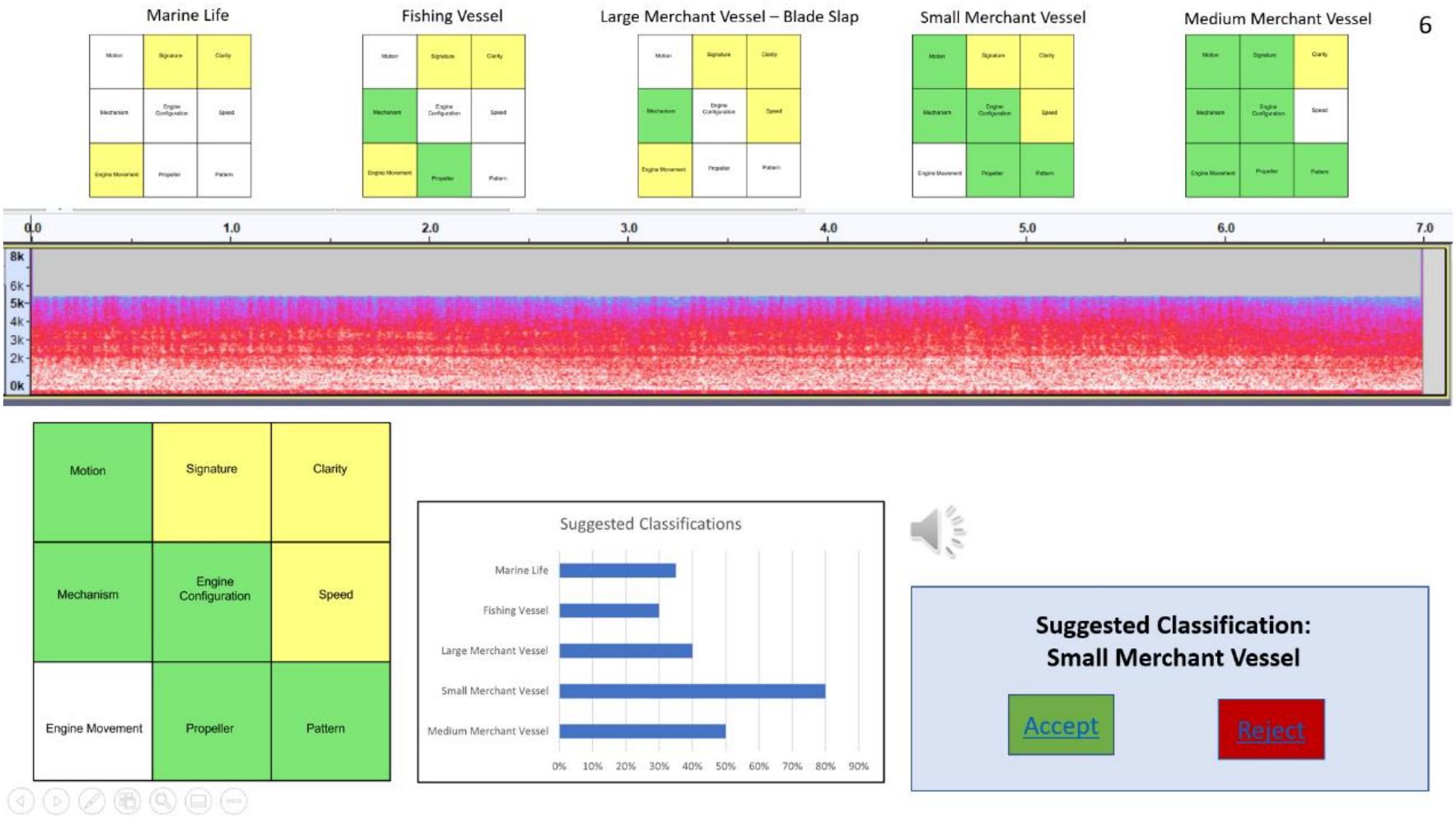


Figure 16: A picture of the experimental test bed, showing a VINAS and suggested classification for a Small Merchant Vessel

After each classification decision was made, the participant would be asked to rate their self-confidence in their classification decision on a five-point Likert scale, being asked, “How confident are you in your decision?” With ratings from 1: Very low, to 5: Very high.

Trust and self-confidence are inherently related. Self-confidence has been used in other studies as a measure of dispositional trust (Hoff and Bashir, 2015). Self-confidence has been shown to be mediated by trust in a system, and fluctuations in self-confidence can be related to automation performance, and trust in the automation (Lee and Moray, 1992, 1994).

Figure 17: A picture of the screen shown after each classification decision, asking a participant to rate how confident they were in their decision

7.3.1: Experiment One Overview

Experiment One sought to test how the inclusion of a VINAS would impact a user’s performance, trust, confidence, and perceived workload, when using an autonomous classifier which suggested a classification for a hydrophone recording. In half of the conditions, no VINAS was shown to the participant. This was to understand whether the inclusion of the VINAS would impact any of the measures in a meaningful way.

The hypotheses for Experiment One are as follows:

H1: The inclusion of a VINAS would improve performance when accepting or rejecting classification decisions

H2: The inclusion of a VINAS would improve trust in an autonomous system when accepting or rejecting classification decisions

H3: The inclusion of a VINAS would improve self-confidence when accepting or rejecting classification decisions

H4: The inclusion of a VINAS would lower perceived workload when accepting or rejecting classification decisions

H4a: The inclusion of a VINAS would lower perceived frustration when accepting or rejecting classification decisions

The reasoning behind the hypotheses for Experiment One is as follows.

VINAS presents an understandable, visual explanation behind an autonomously generated classification. As VINAS is derived directly from a recording, it gives an accurate visual representation of the features of that recording which are pertinent to classification.

It uses classifiers derived from the cognitive classification process of an SME in sonar classification, and so provides a high-level description of the features of the recording which supports the classification decision-making process.

Therefore, it should provide evidence to support a classification decision, which can help improve performance [H1] and trust [H2] by making a classifier's decision more transparent, and understandable, to a user, both key factors for developing trust in automation.

Having supporting evidence for a classification should make it clearer to a user whether a classification is viable, therefore boosting their self-confidence in their decision making [H3].

Providing evidence which can support a classification decision should make it easier to understand and evaluate, lowering perceived workload [H4] and frustration [H4a].

7.3.2: Experiment One Methodology

7.3.2.1: Participants

Thirteen participants, ten male, and three female, (age mean = 35, range = 27-52) took part in the experiment. Participants came from a variety of backgrounds, 5 office workers, 5 with physically demanding jobs, 2 students, and 1 unemployed. None had previous experience with sonar sounds or imagery. There was no access to expert Naval participants at this time, and so naïve participants were used. Because of the small and distinct range of samples, using expert participants for this study may not have been preferrable, as it would be very easy for an expert to be able to identify classifications without having to utilise VINAS for these experiments, so in a sense, using naïve participants allowed for a more accurate reflection of how VINAS would be used when there was uncertainty around a classification decision.

7.3.2.2: Conditions

There were two conditions for the experiment. Classification with VINAS, and Classification without VINAS.

7.3.2.3: Measures

Measures were taken for workload, trust, self-confidence, and performance.

To measure trust, the Checklist for Trust between People and Automation (Jian, Drury and Bisantz, 2000) was used. It is a widely-used measure that assesses beliefs in automation's trustworthiness and its capabilities. It is a 12-item measure with a 7-point Likert rating scale for each question. Overall scores range from 7- 84. Higher scores indicate a higher level of trust in automation.

The Checklist was statistically derived based on conceptualisations of human-autonomy trust such as reliability, integrity, familiarity and honestly through cluster analysis of words associated with trust, and has been independently validated for efficacy by several other studies.

A NASA TLX was used to measure workload. The NASA TLX is a subjective, multi-dimensional assessment tool which is widely used to rate perceived workload. This can be used to assess a system's effectiveness, as well as other aspects of performance. It rates performance across six dimensions to determine an overall workload rating. The ratings can also be used individually to assess perceived mental workload, physical workload, temporal workload, perceived performance, perceived effort and perceived frustration with a task or system.

Copies of both questionnaires are included in Appendix B and Appendix C respectively.

Self-confidence was measured on a 5-point Likert scale asking, "How confident were you in your decision?" (see Fig. 15), with one being very low, and five being very high.

Performance was measured by the number of correct classifications that were accepted.

After all of the classification decisions were complete, a semi-structured interview was performed with each participant, designed to better understand what strategies they employed to make their decisions. Five questions were asked in total, which can be seen in Appendix D.

7.3.2.3: Protocol

The experiment was a repeated measures design. Participants classified ten sounds in total over both of the conditions. The conditions were randomised between each participant.

Confidence was measured after each classification decision. NASA TLX and the Checklist between People and Automation were both issued after each participant had made five classification decisions. Once all classification decisions were made, a semi-structured interview was conducted with the participants.

Participants completed the experiment at the University of Birmingham and were seated comfortably in front of a laptop. They completed the classifications using the experimental interface shown in Figure 16.

7.3.2.5: Data Analysis

Statistical significance for differences between the conditions for all measures were assessed using paired-samples t-tests in IBM SPSS V29. Significance was set at $p<0.05$. Effect Size was calculated using Cohen's *d*.

Cohen's *d* is calculated by:

$$D = \frac{M_1 - M_2}{S_p}$$

Where M_1 and M_2 are sample means for groups 1 and 2, and S_p represents the estimated population standard deviation.

The following measures were compared:

- Performance score in the VINAS present and VINAS absent conditions to test H1
- Trust between People and Automation Checklist scores in the VINAS present and VINAS absent conditions to test H2
- Self-confidence ratings in the VINAS present and VINAS absent conditions to support H3
- Overall workload scores from the NASA TLX ratings in the VINAS present and VINAS absent conditions to test H4
- Frustration scores from the NASA TLX ratings in the VINAS present and VINAS absent conditions to test H4a

7.3.3: Experiment One Results

7.3.3.2: Performance

For Performance Score, VINAS Present [(M = 4.23), (SD = 0.927)] was significantly higher than for the VINAS Absent [(M = 3.38), (SD = 1.193)] condition: [t(12)=3.811, p=0.001, d=0.801]. This supports H1, showing better performance with a VINAS included than without a VINAS (see Figure 15).

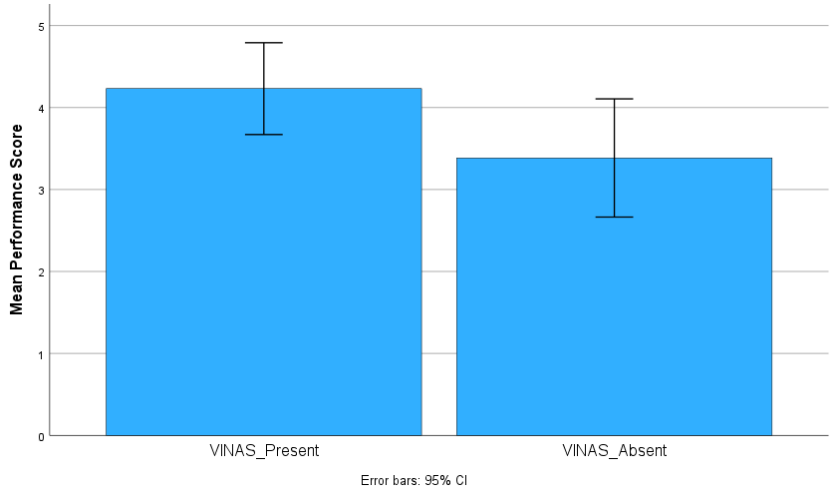


Figure 18: Bar graph with error bars comparing mean performance score for VINAS_present and VINAS_absent conditions

7.3.3.3: Trust

For Trust between People and Automation Checklist score, VINAS Present [(m = 53.08), (SD = 15.163)] was significantly higher than for the VINAS Absent [(M = 44.62), (SD= 12.997)] condition: [t(12)=0.003, p=0.034, d=.555]. This supports H2, showing higher trust ratings with a VINAS included than without a VINAS (see Figure 16).

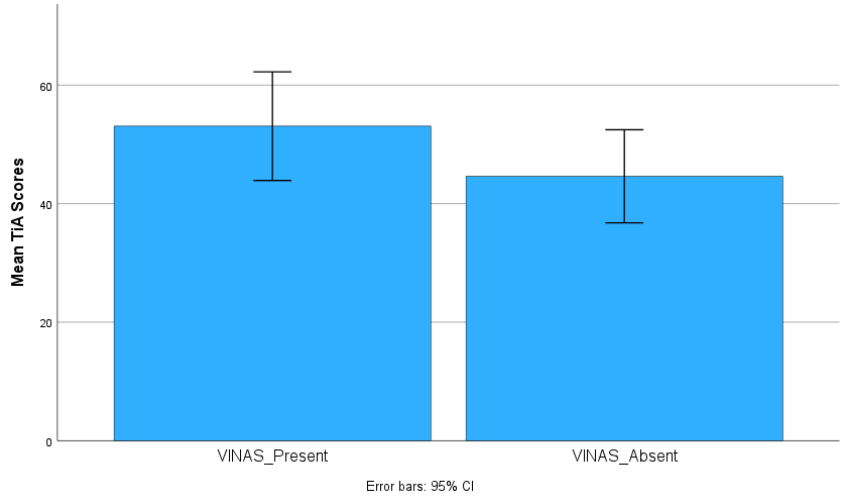


Figure 19: Bar chart with error bars for mean TiA score for VINAS_Present and VINAS_Absent conditions

3.3.3.4: Confidence

There was not a significant difference in the confidence scores for VINAS Present [(M = 18.23), (SD = 3.767)] and VINAS Absent [(M = 17.62), {SD = 3.097)] conditions: [t(12) = 0.519, p = 0.307, d = 0.144]. This does not support H3, showing no significant difference in confidence ratings with a VINAS included and without a VINAS.

3.3.3.5: Workload

A summary of the results for the paired t-tests for the NASA TLX total score, and individual scores, can be seen below:

Table 12: Results of the Paired t-tests for NASA TLX comparing VINAS present and VINAS absent

Workload Variable	Condition	Mean	SD	t-value	Cohen's *d*	p-value
Mental Demand	VINAS Present	5.23	4.512	-0.56	-0.155	0.293
	VINAS Absent	6	4.915			
Physical Demand	VINAS Present	3.23	3.491	0.143	0.04	0.444
	VINAS Absent	3.153	3.184			
Temporal Demand	VINAS Present	3.384	3.524	0.662	0.184	0.26
	VINAS Absent	2.846	2.339			

Performance	VINAS Present	7.846	4.651	-0.636	-0.176	0.268
	VINAS Absent	8.615	5.59			
Frustration	**VINAS Present**	**5.538**	**5.538**	**-1.833**	**-0.508**	**0.046**
	VINAS Absent	**8.076**	**8.076**			
Effort	VINAS Present	5.384	4.073	-0.867	-0.24	0.202
	VINAS Absent	6.461	4.701			
Total	VINAS Present	5.988	3.111	-5.87	-0.163	0.284
	VINAS Absent	6.488	2.751			

For Frustration Score, VINAS Present [(M = 5.538), (SD = 5.125)] was significantly lower than VINAS Absent [(M = 8.077), (SD = 5.937)]: [t(12) = -1.833, p = 0.046, d = 0.508]. This supports H4a, showing a significant difference in perceived frustration with a VINAS included compared with no VINAS.

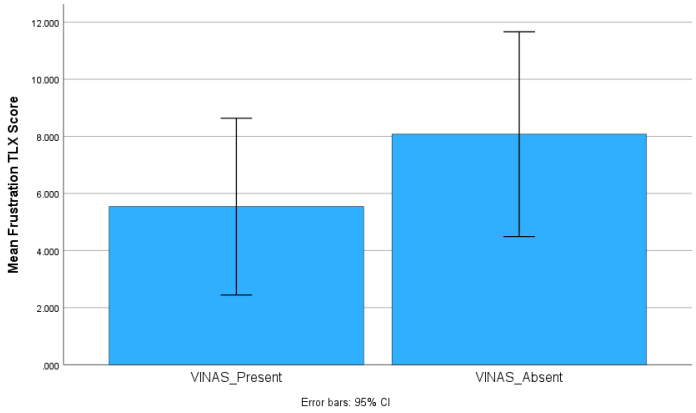


Figure 20: Bar chart with error bars for mean Frustration TLX score for VINAS_Present and VINAS_Absent conditions

Overall workload score was not found to be significant.

7.3.4: Experiment One Discussion

The presence of a VINAS significantly improved both performance and Trust in Automation (TiA) score, thus supporting both H1 and H2. Based on the Cohen's *d* value, the inclusion of a VINAS had a large effect on performance (*d* = 0.8) and a medium effect (*d* = 0.5) on trust.

Although the difference was not significant, the inclusion of VINAS did also impact the self-reported confidence scores, with the scores being slightly higher for the VINAS Present condition.

In Experiment One, the suggested classification was always correct, and yet participants still rejected the classification suggestion at times. This could be seen as evidence as to why some kind of explanation or narrative behind an autonomous decision maker's choices is vital for fostering trust; even in perfect conditions, participants are still wary of the output of a classification aide which does not afford them any opportunity to understand the processes by which it works. Experiment One demonstrates that by providing some kind of understandable background to an autonomous decision, an increase in performance and trust can be observed.

This could also account for the significantly reduced frustration scores observed when a VINAS is present. By providing some kind of context to the otherwise opaque suggested classification scores, participants may feel more comfortable in accepting a classification decision.

Although the results were not significantly different, perceived mental workload was also lower for the VINAS Present condition [(M = 5.23), (SD = 4.51)] in comparison to the VINAS Absent [(M = 6.0), (SD = 4.92)] condition.

There was also a reduction in perceived effort when a VINAS was present [(M=5.38), (SD = 4.07)] compared with when the VINAS was absent [(M = 6.46), (SD = 4.7)], although this result was also not significant.

Overall, the inclusion of a VINAS has many positive effects when compared with the absence of one. Experiment One's results are encouraging, and although they are somewhat limited by the small sample size and number of participants, they warrant further research, showing the inclusion of VINAS has a positive impact, even on such a small scale.

7.4.1: Experiment Two Overview

Experiment Two sought to understand how the inclusion of a VINAS would impact a user's performance, trust, confidence, and perceived workload, when using an autonomous classifier which suggested a classification for a hydrophone recording.

In this experiment, the accuracy of the suggested classification was manipulated, providing an incorrect classification suggestion for half of the classification decisions. This was to better understand the effects of the inclusion of VINAS when there was incongruent, or mis-matched information sources to use when accepting or rejecting a classification decision, creating uncertainty around classification decisions, and whether this would impact a user's performance, trust in the automation, self-confidence, or perceived workload.

High levels of uncertainty have been shown in the literature to have a negative effect on trust in autonomy (Kirschenbaum, 2002). It has also been demonstrated to have a negative impact on perceived workload, situational awareness, and performance (Loft *et al.*, 2015).

Experiment Two seeks to understand the effect of incongruent pieces of information on a classification decision. This is done by manipulating the reliability of the suggested classification in order to induce information incongruency.

A salient choice was made not to manipulate the reliability of the VINAS, as this should always be based on the ground truth, which is the recording itself. Therefore, it would be unrealistic to present an unreliable VINAS. Instead, suggested classification is manipulated. This is because in reality, a classification decision could be based upon unreliable or faulty information, incorporate a source of information with low accuracy or reliability, or noisy information from fused information or sensor sources, for example.

The hypotheses for Experiment Two are as follows:

H1: Trust will be higher when VINAS and the suggested classification are congruent

H2: Performance will be higher when VINAS and the suggested classification are congruent

H3: Self-confidence will be higher when VINAS and the suggested classification are congruent

H4: Workload will be lower when VINAS and the suggested classification are congruent

H5: Frustration will be lower when VINAS and the suggested classification are congruent

H6: Effort will be lower when VINAS and the suggested classification are congruent

When faced with uncertainty around a decision, trust, workload, and self-confidence will decrease. Frustration and effort will increase. Therefore, when all of the information on the screen is congruent, it should be expected that trust [H1], performance [H2] and self-confidence [H3] will score higher.

Trying to work out what piece of information to use to make a decision may add to cognitive load, and therefore it could be expected that in the congruent condition, workload [H4], frustration [H5] and effort [H6] will be lowered.

7.4.2: Experiment Two Methodology

Participants, measures, and protocol were the same for Experiment Two as they were for Experiment One. Data analysis was conducted in the same way (see Section 7.3).

7.4.2.1: Conditions

There were two conditions for Experiment Two, VINAS and Suggestion congruent, and VINAS and Suggestion incongruent. In the incongruent condition, the suggested classification was always incorrect. There were five classifications for each condition.

7.4.3: Experiment Two Results

7.4.3.1: Performance

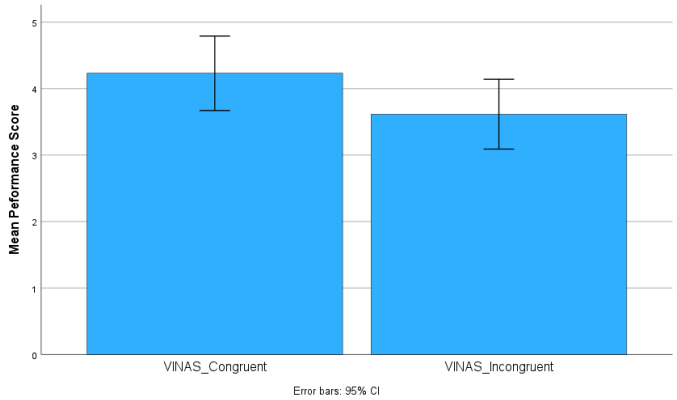


Figure 21: Bar chart with error bars comparing mean performance score for VINAS_Congruent, VINAS_Incongruent

For Performance Score, VINAS Congruent [(M = 4.23), (SD = 0.927)] was significantly higher than VINAS Incongruent [(M = 3.62), (SD = 0.87)]: [t(12) = 2.125, p = 0.027, d = 0.59]. This supports hypothesis H1, that performance would be higher when the information presented was congruent.

7.4.3.2: Trust

For TiA score, VINAS Congruent [(M = 53.08), (SD = 15.163] was significantly higher than VINAS Incongruent [(M = 41.92), (SD =13.009)]: [t(12) = 2.782, p = 0.008, d = 0.722]. This supports hypothesis H2, that trust in automation would be higher when the information presented was congruent.

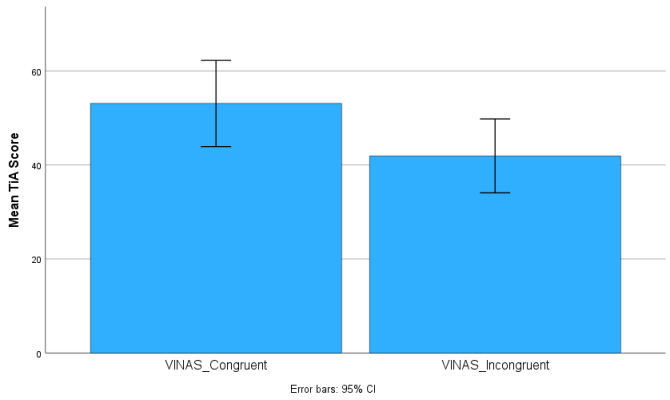


Figure 22: Bar chart with error bars comparing mean performance for VINAS_Congruent, VINAS_Incongruent

7.4.3.3: Self-Confidence

For self-confidence rating, VINAS Congruent [(M = 18.23), (SD = 3.767] was not significantly different to VINAS Incongruent [(M = 19.08), (SD =2.565)]. This does not support hypothesis H3.

7.4.3.4: Workload

A summary of the workload ratings, including total workload, can be seen below.

Table 13: A table comparing t-test results for the Congruent and Incongruent conditions

Workload Variable	Condition	Mean	SD	t-value	Cohen's *d*	p-value
Mental Demand	Congruent	5.23	4.512	-2.658	-4.6	0.062
	Incongruent	7.307	5.0888			
Physical Demand	Congruent	3.23	3.491	1.594	0.442	0.068
	Incongruent	2.846	3.236			
Temporal demand	Congruent	3.384	3.524	-1.389	-0.385	0.95
	Incongruent	4.076	3.882			
Performance	Congruent	7.846	4.651	-1.771	-0.491	0.051
	Incongruent	10.307	4.441			
Frustration	**Congruent**	**5.538**	**5.125**	**-3.007**	**-0.834**	**0.005**
	Incongruent	**8.856**	**4.913**			
Effort	Congruent	5.384	4.073	-1.379	-0.382	0.097
	Incongruent	6.846	4.219			
Total	**Congruent**	**5.988**	**3.111**	**-1.809**	**-0.502**	**0.048**
	Incongruent	**6.77**	**2.69**			

It can be seen from the table that overall workload, as well as frustration, were significantly lower when the VINAS and suggestion were congruent.

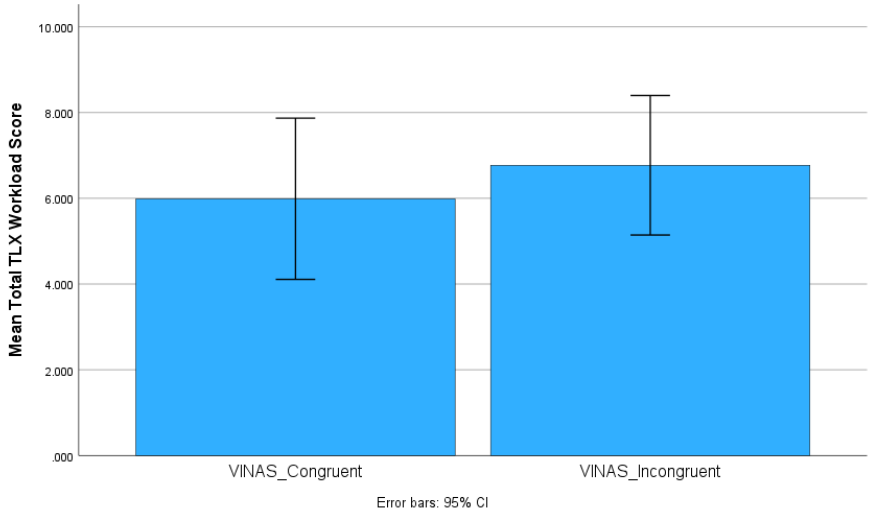


Figure 23: Bar chart with error bars comparing mean total workload score for VINAS_Congruent and VINAS_Incongruent conditions

For the total workload score, VINAS Congruent [(M = 5.988), (SD = 3.111] was significantly lower than VINAS Incongruent [(M = 6.77), (SD =2.69)]: [t() = 12, p = 0.048, d = -0.502]. This supports hypothesis H4, that workload would be lower when information is congruent.

7.4.3.5: Frustration

For Frustration TLX score, VINAS Congruent [(M = 5.538), (SD = 5.125] was significantly lower than VINAS Incongruent [(M = 8.846), (SD =4.914)]: [t(12) -3.007, p = 0.005, d = 0.834)

Incongruency causes a heavy penalty in terms of perceived frustration. This supports hypothesis H5, which stated frustration would be lower when information is congruent.

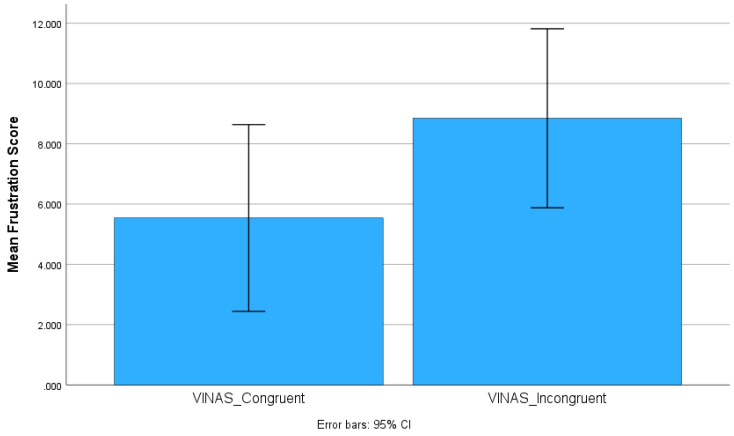


Figure 24: Bar chart with error bars comparing frustration TLX score for the VINAS congruent and VINAS incongruent conditions

7.4.3.6: Effort

There was so significant difference for effort when information was congruent. This does not support hypothesis H6.

7.4.3.4: Experiment Two Discussion

Results show that when the information was congruent, participants had higher performance and higher trust. This supports the literature, which states that uncertainty can degrade performance and trust.

Surprisingly, participants were no more confident when the information was congruent, than when it was incongruent, disproving hypothesis H3. This could be for a number of reasons. If participants have some automation bias, they may have their self-confidence inflated by feeling that they can "outsmart" the autonomy. However, the difference in performance scores makes it seem like there was a disparity between the participant's perceived performance, and their actual performance. This is supported by looking at the TLX performance scores, as participants thought they were performing better in the incongruent condition.

Frustration was a lot higher when the information did not match. This could imply that the participants were struggling to identify which piece of information to use to make their decisions.

Although effort was not significantly lower as was put forward in H6, it was lower for the congruent information condition. This was the same for temporal and mental demand, which also were perceived to be lower when all information matched.

In conclusion, this goes some way to answering RQ1, RQ4, RQ5 and RQ6.

VINAS helps to maintain appropriate levels of trust and improves performance, even when faced with incongruent information. It provides a credible, understandable, and trustable explanation behind its decisions.

7.5: Conclusion

In conclusion, from the experimentation outlined above, it can be seen that the inclusion of a VINAS has many beneficial effects. VINAS helps to improve both performance and trust in automation in a significant way. It also was shown that VINAS can help to reduce perceived workload and frustration.

However, the author acknowledges that the number of participants for both experiments were small, and also that the participants lacked familiarity with classification of hydrophone recordings.

Using non-expert participants may have been beneficial for these experiments. This is because only a small number of hydrophone recordings were used in the experimentation, and these recordings were quite distinct, with clear characteristics which could be used in the classification process audible in the recordings. Using participants with experience of sonar classification would not have provided insight into how the VINAS was used, as they would have potentially been able to classify the recordings without using the additional information sources. Therefore, using non-expert participants may actually have provided more insight into how the VINAS was used, especially when there was uncertainty around a particular classification decision.

Future work utilising more participants may be beneficial, and provide more rigorous results. A larger sample size would allow for these preliminary results to be validated.

CHAPTER 8: FINDINGS AND LIMITATIONS

8.1: Discussion

The aim of the thesis as a whole was to determine how human-autonomy teaming could be utilised in future Naval defence, specifically in the maritime SMP domain, in a trustworthy, safe, useful and advantageous way.

This research attempted to summarise best practice in the field of trust in autonomous systems with regards to how, what kind, and what level of autonomous system would be appropriate to introduce, especially with regards to the unique properties of the socio-technical system of an SMP.

It presented research into the decision-making schema employed by operators in command-and-control. It did this in three ways. Firstly, by reviewing the interactions and teamwork within the SMP as a form of a socio-technical system, and reviewing research into how information is shared and used throughout that system to facilitate the achievements of goals and development of the tactical picture.

It secondly introduced a decision-making model, called the Recognition Primed Decision model, which has been shown in the literature to be representative of the way in which skilled operators make decisions when under heavy time penalties and high cognitive workload.

It further explored whether this model was applicable to Sonar Operators through performing a Critical Decision Method interview with a SME, in order to understand not only the physical tasks they perform in order to classify, but also how they mitigate and manage the inherent uncertainties of the task. This is shown in Chapter Five.

Thirdly, it then built on this by attempting to elicit some of the cognitive processes and heuristics which operators employ in the process of classification by conducting a study using the repertory grid interview technique. This is shown in Chapter Six.

This modelling of how operators construct their understanding and situation cognitively is important: the literature shows a need for explainability and inter-communication between operator and autonomy, as demonstrated in Chapter Two. Even if a system behaves at a level of perfect reliability, it can still be mis-used; and overall performance, even at high- or perfect- levels of reliability introduce new and dangerous opportunities for things to go wrong.

Therefore, as posited by Chapter Two, it is vitally important for systems to produce explanation and transparency behind their decision-making. Understanding how operations make their decisions is therefore important to facilitate design features which support this.

The research then presented an initial idea for a display which could help aid Sonar Operators with their classification process, known as the VINAS, developed to both aid with contact classification when there are high volumes of contacts to classify quickly, a need identified in Chapter Four through accident analysis, and to increase the explainability of an autonomous classifier, as shown to be required in Chapter Two.

Finally, initial testing of the VINAS to evaluate whether it had a positive impact on trust and performance was performed in Chapter Seven, showing promising preliminary results.

8.2: Thesis Objectives

The first objective of the research was to formulate an understanding of the concepts of trust and autonomy and how they relate to each other. This was done in Chapter Two, with several models being discussed and demonstrated in the literature, and the relationship between level of autonomy, type of autonomy, the autonomy's level of performance, trust, overall performance, and overall situation awareness being explored and established.

The second objective of the thesis was to present an understanding of how a SMP works as a socio-technical system. This was done in order to understand the informational requirements of each component and how they could potentially be facilitated or supported through the integration of

forms of autonomy. This was demonstrated in Chapter Three, where an explanation of how information is communicated was presented, along with suggestions and recommendations from the literature which had been developed through observational study and simulation.

The third objective was to understand how SMP accidents can occur from a systemic perspective. Two accidents which had occurred since 2015 were analysed using the AcciMapping method in Chapter Four, identifying key areas which could be supported with intelligent information processing.

These were, firstly, the planning of navigation during the pre-mission stages of operation, where autonomous systems could be beneficially used to better plan navigation routes depending on predicted oceanic traffic levels, as well as increase understanding of potential hazards in the operational environment, in order to facilitate safer operations. Secondly, it was identified that when there are large numbers of contacts to classify in periods with heavy time constraints, Sonar Operators could benefit from systems to help mitigate their cognitive load.

The fourth objective of the thesis was to understand better the identified user for that use case, the Sonar Operator. Chapter Five, Chapter Six, and Chapter Seven aimed to do this by developing understanding of how the Sonar Operator performs their tasks, both physically and cognitively, in order to inform the design of autonomous systems which could aid in their classification process. This was done through modelling their task and their decision making using the Critical Decision Method interview technique in Chapter Five, and trying to develop sets of key cognitive constructs which they used in their mental classification heuristics using the Repertory Grid interview technique in Chapter Six.

The fifth objective of the research was to develop a visualisation using the elicited user requirements and informational needs of the SO to support their classification methodology and help manage their uncertainty and cognitive load. This was developed through Chapter Six, with the VINAS visualisation being produced.

8.3: Thesis Research Questions

The research questions of the thesis were as follows:

RQ1: What Level and Type of Autonomy could be suitably applicable to tasks caried out in the process of broadband sonar classification, whilst maintaining appropriate levels of trust in the automation?

The thesis presents a strong argument for low- to mid-level automation being the most suitable for aiding in the task whilst still maintaining appropriate levels of trust. It posits that because of the dangers of over-reliance and complacency, and also potential biases in users' propensity to trust (in general, not just SOs), as identified in the literature in Chapter Two, that suggestive automation, which offers solutions without hiding the real-world picture, would be more beneficial to SOs than decisive information. This is especially true because of the safety-critical nature of the work, and also for the need to develop as accurate distributed SA as possible, which cannot be achieved without human-in-the-loop interactions, rather than human-on- or human-outside- the loop.

The thesis posits that *information* automation is the **most beneficial** type of automation to be used in the task of sonar classification. It suggests that some *action* automation could be beneficial, but only ever along-side an SO's own actions, so at low levels of automation. This was assessed through evaluation of the literature, showing information automation suffers less of a penalty to SA and reliance when it fails (Parasuraman, Thomas B. Sheridan and Wickens, 2000; Onnasch *et al.*, 2014b).

It proposes that:

Moderate levels of *information acquisition automation* would be appropriate, as certain types of contact (emerging, behaving unusually) and areas of the display (waiting for a contact to re-emerge) could be highlighted to an operator. This does not disguise the real-world picture, but could also be beneficial to draw attention to areas when there are heavy cognitive loads on an operator (Parasuraman, Thomas B. Sheridan and Wickens, 2000).

Low levels of *information analysis automation* may be beneficial to the process of TMA. This could involve drawing projections of the most optimal solutions, in order to reduce workload for TMA operators.

The thesis argues that *decision selection* automation could be used to facilitate more effective information communication by sending solutions directly from the sound room to the control room. However, as the separation of the sound room and control room is not necessarily a design choice which will be carried further into future SMP designs, this may not be needed. Discussion of positioning TMA and sonar displays closer to each other to facilitate this can be found in Chapter Three.

The thesis also suggests low level *action automation* can be beneficial, but must be carefully mediated. It suggests this in three ways; by deriving suggested speeds for contacts from the information available, and offering these as suggestions, and by fixing points in TMA solutions, to reduce the manual manipulations of these solutions.

It also posits that when a SO is experiencing heavy cognitive load, low-level action automation could be beneficial by automatically classifying contacts, and offering classification solutions, but only as *suggestions*, and only when classification ambiguity is low.

VINAS would offer an extension of this, as it provides the explainability behind the classification decision which would allow operators to evaluate the solution's viability, therefore, mediating the action automation by confining it to suggestion, but allowing it to make its own decisions independently of an SO, would allow the beneficial effects of this type of automation to be experienced with less of a penalty through loss of SA, or automation induced complacency.

When evaluating VINAS with regards to the review of XAI in Chapter Two, it can be seen that VINAS provides a simple, visually assessable way to evaluate a system's decision-making. It does not rely on text, or anthropomorphised agents in order to provide an explanation, which can introduce

additional clutter and take up screen-space, as well as drawing attention away from the tactical picture. It does not simply list weighted classifiers, or present a large tree, which can be cumbersome and slow to explore and find relevant information for evaluation. Instead, it tries to stick with design principles which make it simplistic, eye-catching, easy to assess, and easily interpretable (if familiar with classifiers used in traditional sonar classification).

RQ2: How can the causes of previous SMP accidents be made safer through the introduction of autonomy?

The thesis highlights two key areas where autonomy could be beneficial, as identified by the AcciMap analysis in Chapter Four. As an aid to planning routes and understanding traffic in operational areas in pre-mission planning, and supporting SOs with classification when they have a high volume of contacts to classify, as described above.

RQ3: How do Sonar Operators cognitively classify sounds?

The thesis explores this research question through different types of interviews with SMEs and presents the results through the cognitive constructs used in the VINAS display, as shown in Chapter Six.

RQ4: How can an autonomous decision aid visually present a credible, understandable, and trustable explanation behind its decisions?

RQ4 is demonstrated through the VINAS, which offers an explanation based on the cognitive classifiers derived from the two types of interview with SMEs. It tries to do this in a number of ways. Firstly, it offers a quick, visual way to assess a classification decision once an operator is familiar with the coloured patterns that emerge on a VINAS for each type of classification.

Secondly, because the display uses specific constructs derived from SO classification decisions, if a region of the VINAS is coloured in an unexpected way, it becomes easier for an operator to understand which area of the classification has some ambiguity surrounding it. This supports

interdependent classification, as then an SO can apply their expertise in a way which is consistent with the RPD model of decision-making, looking for certain clues and elements which match with their previous experience to derive a classification.

This facilitates trust in the autonomous system by offering an explanation behind its decisions which a human can understand. It also supports their decision-making processes by highlighting specific points of ambiguity embedded in a context of clues which they can relate back to their cognitive classification process.

RQ5: Can a VINAS improve performance in a classification task utilising an autonomous classifier?

RQ6: Can a VINAS improve trust in an autonomous classifier when conducting a classification task?

Questions RQ5 and RQ6 are tested in the experimentation outlined in Chapter Seven. Preliminary experimentation in this area seems to show that VINAS does both improve performance and increase trust in an autonomous classifier, even when there is incongruent information offered by a system.

8.4: Industry and Public Engagement

The PhD was an ICASE studentship performed in collaboration with BAE Systems. ICASE studentships aim to facilitate industrially relevant and applied research, offering the researcher expertise outside of an academic setting.

This meant there was some access to a SME during the earlier stages of the PhD, which would have been extremely difficult to achieve otherwise. Input from BAE Systems has enabled the research to be better grounded in a realistic understanding of the maritime domain.

In the later stages of the PhD it became possible to have access to examples of future-focused classification systems, which provided some confirmation that the research is aligned with what the future capabilities in sonar classification may consist of. It also allowed the researcher to have direct

input into experimentation carried out in this domain, which again shows that the research is well-aligned with future capabilities.

As well as this, the demonstration of the work at both academic and industrial conferences was well received. Of particular note is the presentation of the work on VINAS at the Underwater Defence Technology 2022, where naval industry figureheads from around the world showed very positive engagement towards the research. The Ministry of Defence representatives in particular gave positive feedback towards the research, showing it is both pertinent and relevant to UK naval defence research.

The work has also been accepted and demonstrated at various academic conferences, including the Naturalistic Decision-Making conference of 2019, the International Conference for Multimodal Interaction in 2021, and the virtual Human Factors and Ergonomics Society conference of 2020. This shows that the research is poignant, accepted and validated by the wider academic community, and aligned well with current research in these areas.

8.5: Limitations and Future Work

There are some limitations to the work. Only one Sonar Operator was involved in the development of the research (although two other SMEs were utilised in the repertory grid study), which is unfortunate.

The original plan for the research was to incorporate large-scale observational studies of users, through observing training exercises of either Naval students or operational personnel. Unfortunately, because of the COVID-19 pandemic, this part of the research was never able to be carried out. This does put limitations on how widely applicable the principles derived are. However, the research has been supported by BAE Systems throughout the PhD, which does give it some credibility. As well as this, it has been published and demonstrated at several prestigious conferences, where the work has been peer-reviewed, which means that it does have some integrity and merit, despite the smaller sample sizes used in the initial research.

Participant numbers were generally low for all of the experimentation. This was due to a number of reasons. Firstly, the COVID-19 pandemic severely limited opportunities for in-person observation and experimentation for the majority of the time the research was being conducted. Secondly, due to the nature of the activities being studied, it was difficult to recruit participants who had expertise in the subject area. This was further confounded by the pandemic making it difficult to visit personnel.

Therefore, the experimentation outlined in Chapter Five and Chapter Six could have benefited from better access to SMEs, to increase the number of experts being interviewed, and to reassert the validity and applicability of the research. It is unfortunate that this was not achieved, but understandable, considering the global pandemic severely limiting the opportunity for observation and networking for half of the time allotted to carrying out the PhD, and also, the fundamental restrictions because of the area of research being so heavily related to national defence.

The experimentation outlined in Chapter Seven is somewhat limited as it was performed with participants who had no experience of sonar data. It also suffered from a small sample sizing and limited variables. However, the purpose of the thesis was not to specifically develop and evaluate an interface for classification, but instead was to inform how the introduction of automation into a maritime defence task could occur. The thesis has done this by presenting a unique methodology for deriving cognitive methodologies employed by experts, and the comprehensive literature review allowed for well-informed recommendations to be made.

Further experimentation, engaging with participants with experience of sonar classification, would build on the results shown here, and show their replicability. Repeating the experimentation conducted in Chapter Seven with a larger number of participants would be beneficial, and help to show how robust the results are. Repeating the experimentation outlined in Chapter Seven with expert participants and a larger variety of hydrophone recordings would help to build on this foundation of research, showing how VINAS could work in more realistic scenarios.

REFERENCES

AIS Transponders (no date) *International Maritime Organisation*. Available at: https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx.

Allen, G. and Taniel, C. (2017) *Artificial Intelligence and National Security*.

Antão, P., Grande, O., Trucco, P.A.O.L.O., Soares, C.G., Martorell, S. and Barnett, J., 2009. Analysis of maritime accident data with BBN models. *Safety, reliability and risk analysis: theory, methods and applications*, *2*, pp.3265-3274..

Asplin, R.G. and Christensson, C.G., 1988, October. A new generation side scan sonar. In *OCEANS'88.'A Partnership of Marine Interests'. Proceedings* (pp. 329-334). IEEE..

Athos, A.G., Gabarro, J.J. and Holtz, J.L. (1978) *Interpersonal behavior : communication and understanding in relationships*. Englewood Cliffs (N.J.) : Prentice-Hall.

Baber, C. (1996) 'Repertory Grid and its Application to Product Evaluation', in P. Jordan et al. (eds) *Usability Evaluation in Industry*, pp. 157–166.

Baber, C. (2015) 'Repertory Grid for Product Evaluation', in *Handbook of Human Factors and Ergonomics Methods*. London: Taylor and Francis, pp. 31.1-31.7

Bainbridge, L., 1983. Ironies of automation. In *Analysis, design and evaluation of man–machine systems* (pp. 129-135). Pergamon.

Banks, V.A., Plant, K.L. and Stanton, N.A., 2020. Leaps and shunts: designing pilot decision aids on the flight deck using Rasmussen's ladder. *Contemporary ergonomics and human factors*.

Barnes, M.J., 2003. *The Human Dimensions of Battlespace Visualization: Research and Desing Issues*. Army Research Laboratory.

Bartlett, M.L. and McCarley, J.S., 2021. Ironic efficiency in automation-aided signal detection. *Ergonomics*, *64*(1), pp.103-112.

Blockeel, H., Devos, L., Frénay, B., Nanfack, G. and Nijssen, S., 2023. Decision trees: from efficient

prediction to responsible AI. *Frontiers in Artificial Intelligence*, *6*.

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), pp.54-61..

Bradshaw, Jeffrey M *et al.* (2013) 'The Seven Deadly Myths of "Autonomous Systems"', *IEEE Intelligent Systems*, (IS-28-03-HCC), pp. 2–9.

Branford, K., Hopkins, A. and Naikar, N., 2009. Guidelines for AcciMap analysis. In *Learning from high reliability organisations*. CCH Australia Ltd.

Breznitz, S., 2013. *Cry wolf: The psychology of false alarms*. Psychology Press.

Brinkmann, K. and Hurka, J., 2009. Broadband passive sonar tracking. *Informatik 2009–Im Focus das Leben*.

Butler Jr, J.K. and Cantrell, R.S., 1984. A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological reports*, *55*(1), pp.19-28.

Carrigan, G.P., 2009. *The design of an intelligent decision support tool for submarine commanders* (Doctoral dissertation, Massachusetts Institute of Technology).

Chancey, E.T., Bliss, J.P., Yamani, Y. and Handley, H.A., 2017. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors*, *59*(3), pp.333-345.

Chandra, D., Zuschlag, M., Helleberg, J. and Estes, S., 2009, October. Symbols for cockpit displays of traffic information. In *2009 IEEE/AIAA 28th Digital Avionics Systems Conference* (pp. 5-B). IEEE.

Chauvin, C., 2011. Human factors and maritime safety. *The Journal of Navigation*, *64*(4), pp.625-632.

Chauvin, C., Lardjane, S., Morel, G., Clostermann, J.P. and Langard, B., 2013. Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS. *Accident Analysis & Prevention*, *59*, pp.26-37.

Chen, J.Y., 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), pp.259-282.

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

Chen, M., Nikolaidis, S., Soh, H., Hsu, D. and Srinivasa, S., 2018, February. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction* (pp. 307-315).

Chien, S.Y., Lewis, M., Sycara, K., Liu, J.S. and Kumru, A., 2018. The effect of culture on trust in automation: reliability and workload. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), pp.1-31.

Chien, S.Y., Lewis, M., Sycara, K., Liu, J.S. and Kumru, A., 2016, October. Influence of cultural factors in dynamic trust in automation. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 002884-002889). IEEE.

Chien, S.Y., Lewis, M., Sycara, K., Kumru, A. and Liu, J.S., 2019. Influence of culture, transparency, trust, and degree of automation on automation use. *IEEE Transactions on Human-Machine Systems*, *50*(3), pp.205-214.

Christoffersen, K. and Woods, D.D., 2002. How to make automated systems team players. In *Advances in human performance and cognitive engineering research* (pp. 1-12). Emerald Group Publishing Limited.

Clarke, J., 2023. A proposed submarine electronic chart display and information system.

Cohen, T.N., Wiegmann, D.A. and Shappell, S.A., 2015. Evaluating the reliability of the human factors analysis and classification system. *Aerospace medicine and human performance*, *86*(8), pp.728-735.

Colquitt, J.A., Scott, B.A. and LePine, J.A., 2007. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology*, *92*(4), p.909.

Cunningham, A. and Thomas, B., 2005, January. Target motion analysis visualisation. In *ACM International Conference Proceeding Series* (Vol. 109, pp. 81-90).

Curtis, A.M., Wells, T.M., Higbee, T. and Lowry, P.B., 2008. An overview and tutorial of the repertory grid technique in information systems research. *Communications of the Association for Information*

Systems (CAIS), *23*(3), pp.37-62.

Defense Science Board (2016) *Report of the Defense Science Board Summer Study on Autonomy, Department of Defense*. Washington, DC. Available at: https://doi.org/10.1162/leon_r_01374.

Dieber, J. and Kirrane, S., 2020. Why model why? Assessing the strengths and limitations of LIME. *arXiv preprint arXiv:2012.00093*.

Dietterich, T., 1995. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, *27*(3), pp.326-327.

Dietterich, T.G. and Kong, E.B., 1995. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms.

Dijkstra, J.J., Liebrand, W.B. and Timminga, E., 1998. Persuasiveness of expert systems. *Behaviour & Information Technology*, *17*(3), pp.155-163.

Dixon, S.R. and Wickens, C.D., 2006. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human factors*, *48*(3), pp.474-486.

Dixon, S.R., Wickens, C.D. and McCarley, J.S., 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses?. *Human factors*, *49*(4), pp.564-572.

Dominguez, C., Long, W.G., Miller, T.E. and Wiggins, S.L., 2006, June. Design directions for support of submarine commanding officer decision making. In *Proceedings of 2006 Undersea HSI symposium: research, acquisition and the warrior* (pp. 6-8).

Doney, P.M., Cannon, J.P. and Mullen, M.R., 1998. Understanding the influence of national culture on the development of trust. *Academy of management review*, *23*(3), pp.601-620.

Driskell, J.E., Salas, E. and Hughes, S., 2010. Collective orientation and team performance: Development of an individual differences measure. *Human factors*, *52*(2), pp.316-328.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G. and Beck, H.P., 2003. The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), pp.697-718.

Ehsan, U., Harrison, B., Chan, L. and Riedl, M.O., 2018, December. Rationalization: A neural

machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 81-87).

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B. and Riedl, M.O., 2019, March. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 263-274).

Endsley, M.R., 1995. Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*(1), pp.32-64.

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), pp.462-492.

Erikson, E.H., 1950. *Childhood and society*. WW Norton & Company.

Fay, D., Stanton, N.A. and Roberts, A.P., 2019. All at sea with user interfaces: from evolutionary to ecological design for submarine combat systems. *Theoretical Issues in Ergonomics Science*, *20*(5), pp.632-658.

Fillinger, L., de Theije, P., Zampolli, M., Sutin, A., Salloum, H., Sedunov, N. and Sedunov, A., 2010, November. Towards a passive acoustic underwater system for protecting harbours against intruders. In *2010 International WaterSide Security Conference* (pp. 1-7). IEEE.

Ford, K.M., Bradshaw, J.M., Adams-Webber, J.R. and Boose, J.H., 1993. Beyond the repertory grid: new approaches to constructivist knowledge acquisition tool development. *International Journal of Intelligent Systems*, *8*(2), pp.287-333.

Freitas, A.A., 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, *15*(1), pp.1-10.

Garske, J.P., 1976. Personality and generalized expectancies for interpersonal trust. *Psychological Reports*, *39*(2), pp.649-650.

Genç, H.M., 2010, July. A new solution approach for the bearing only target tracking problem. In *4th International Workshop on Soft Computing Applications* (pp. 95-100). IEEE.

Gimse, H., 2017. *Classification of marine vessels using sonar data and a neural network* (Master's

thesis, NTNU).

Global Fishing Watch (2019) *Global Fishing Watch - About Us, globalfishingwatch.org*. Available at: https://globalfishingwatch.org/about-us/ (Accessed: 1 November 2019).

Global Fishing Watch and Dicaprio, L. (2018) 'Sustainability through transparency'.

Glomsrud, J.A., Ødegårdstuen, A., Clair, A.L.S. and Smogeli, Ø., 2019, September. Trustworthy versus explainable AI in autonomous vessels. In *Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC)* (Vol. 37).

Gutzwiller, R.S. and Reeder, J., 2021. Dancing with algorithms: Interaction creates greater preference and trust in machine-learned behavior. *Human Factors*, *63*(5), pp.854-867.

Hagberg, S., 1997. Edwin Hutchins, cognition in the wild.

Hall, T.J. and Barrett, F.J., 2012. *A case study of innovation and change in the US Navy Submarine fleet*. Monterey, California. Naval Postgraduate School.

Hamim, O.F., Hoque, M.S., McIlroy, R.C., Plant, K.L. and Stanton, N.A., 2019. Applying the AcciMap methodology to investigate the tragic Mirsharai road accident in Bangladesh. In *MATEC Web of Conferences* (Vol. 277, p. 02019). EDP Sciences.

Hawley, J.K., 2017. Patriot wars: automation and the Patriot air and missile defense system.

Hawley, J.K. (2017b) *Patriot Wars | Center for a New American Security, CNAS*. Available at: https://www.cnas.org/publications/reports/patriot-wars (Accessed: 25 January 2019).

Hetherington, C., Flin, R. and Mearns, K., 2006. Safety in shipping: The human element. *Journal of safety research*, *37*(4), pp.401-411.

Hillesheim, A.J., Rusnock, C.F., Bindewald, J.M. and Miller, M.E., 2017, September. Relationships between user demographics and user trust in an autonomous agent. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 314-318). Sage CA: Los Angeles, CA: SAGE Publications.

Hoff, K.A. and Bashir, M., 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), pp.407-434.

Hopko, S.K. and Mehta, R.K., 2021. Neural correlates of trust in automation: Considerations and generalizability between technology domains. *Frontiers in Neuroergonomics*, *2*, p.26.

Huf, S. and Brolese, A., 2006, October. Visualizing uncertainty to improve operators' spatial proximity judgments in uncertain surroundings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 3, pp. 294-298). Sage CA: Los Angeles, CA: SAGE Publications.

Hughes, D.T., Sildam, J., Arnold, B., Ryan, K. and Haun, J., 2010, September. Passive acoustic monitoring during the SIRENA 10 cetacean survey. In *OCEANS 2010 MTS/IEEE SEATTLE* (pp. 1-10). IEEE.

Hollan, J., Hutchins, E. and Kirsh, D., 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), pp.174-196.

Jenssen, G.D., Moen, T. and Johnsen, S.O., 2019, October. Accidents with Automated Vehicles-Do self-driving cars need a better sense of self?. In *Proceedings of the 26th ITS World Congress, Singapore* (pp. 21-25).

Jian, J.Y., Bisantz, A.M. and Drury, C.G., 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, *4*(1), pp.53-71.

Jian, J.-Y., Drury, C.G. and Bisantz, A.M. (2000) 'Foundations for an Empirically Determined Scale of Trust in Automated Systems', *International Journal of Cognitive Ergonomics*, 4(1), pp. 73–86.

Johnson, M., Bradshaw, J.M., Feltovich, P.J., Jonker, C.M., Van Riemsdijk, M.B. and Sierhuis, M., 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, *3*(1), pp.43-69.

Kaempf, G.L., Klein, G., Thordsen, M.L. and Wolf, S., 1996. Decision making in complex naval command-and-control environments. *Human factors*, *38*(2), pp.220-231.

Kemper, G., Ponce, D., Telles, J. and Del Carpio, C., 2019. An algorithm to obtain boat engine RPM

from passive sonar signals based on DEMON processing and wavelets packets transform. *Journal of Electrical Engineering & Technology*, *14*(6), pp.2505-2521.

Kirschenbaum, S.S. and NAVAL UNDERSEA WARFARE CENTER DIV NEWPORT RI, 2002. Uncertainty and automation. In *Proceedings of RTO Human Factors and Medicine Panel (HFM) Symposium*.

Kirschenbaum, S.S., Trafton, J.G., Schunn, C.D. and Trickett, S.B., 2014. Visualizing uncertainty: The impact on performance. *Human Factors*, *56*(3), pp.509-520.

Klein, G.A., 2011. *Streetlights and shadows: Searching for the keys to adaptive decision making*. MIT Press.

Klein, G.A., Calderwood, R. and Macgregor, D., 1989. Critical decision method for eliciting knowledge. *IEEE Transactions on systems, man, and cybernetics*, *19*(3), pp.462-472.

Klein, G. and Calderwood, R. (2015) 'The Recognition-Primed Decision (RPD) Process Your 10-Minute Guide to Making Good Decisions When you Don ' t Have Much Time', *MindTools Corporate* [Preprint].

Klein, G. and Klinger, D. (1991) 'Naturalistic Decision Making', *Human Systems IAC*, 11(1), pp. 769–773. Available at: https://doi.org/10.7314/APJCP.2015.16.2.769.

Klein, G.A., Calderwood, R. and Macgregor, D., 1989. Critical decision method for eliciting knowledge. *IEEE Transactions on systems, man, and cybernetics*, *19*(3), pp.462-472.

Klippenstein, K. (2023) *Exclusive: Surveillance footage of Tesla crash on SF's Bay Bridge hours after Elon Musk announces 'self-driving' feature*, *The Intercept*. Available at: https://theintercept.com/2023/01/10/tesla-crash-footage-autopilot/.

Kohn, S.C., de Visser, E.J., Wiese, E., Lee, Y.C. and Shaw, T.H., 2021. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, *12*, p.604977.

Krippl, M. *et al.* (2016) 'Human-Computer Interaction: Theory, Design, Development and Practice: Part One', in M. Kurosu (ed.) *Lecture Notes in Computer Science (including subseries Lecture Notes in*

Artificial Intelligence and Lecture Notes in Bioinformatics). Springer. Available at: https://doi.org/10.1007/978-3-319-39510-4_29.

Larouzee, J. and Le Coze, J.C., 2020. Good and bad reasons: The Swiss cheese model and its critics. *Safety science*, *126*, p.104660.

Latoschik, M.E., Roth, D., Gall, D., Achenbach, J., Waltemate, T. and Botsch, M., 2017, November. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM symposium on virtual reality software and technology* (pp. 1-10).

Lay, S., Brace, N., Pike, G. and Pollick, F., 2016. Circling around the uncanny valley: Design principles for research into the relation between human likeness and eeriness. *i-Perception*, *7*(6), p.2041669516681309.

Lee, J. and Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), pp.1243-1270.

Lee, J.D. and Moray, N., 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), pp.153-184.

Lee, J.D. and See, K.A., 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), pp.50-80.

Lewis, M., 1998. Designing for human-agent interaction. *AI Magazine*, *19*(2), pp.67-67.

Li, R.H. and Belford, G.G., 2002, July. Instability of decision tree classification algorithms. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 570-575).

Loft, S., Bowden, V., Braithwaite, J., Morrell, D.B., Huf, S. and Durso, F.T., 2015. Situation awareness measures for simulated submarine track management. *Human factors*, *57*(2), pp.298-310.

Loft, S., Morrell, D.B., Ponton, K., Braithwaite, J., Bowden, V. and Huf, S., 2016. The impact of uncertain contact location on situation awareness and performance in simulated submarine track management. *Human Factors*, *58*(7), pp.1052-1068.

Lunsky, L.L., 1966. Identity and the Life Cycle. *Archives of Internal Medicine*, *118*(3), pp.288-289.

Ly, T., Huf, S. and Henley, P. (no date) 'Design for Submarine Command and Control in the 21st Century'.

Lyons, J.B., Sadler, G.G., Koltai, K., Battiste, H., Ho, N.T., Hoffmann, L.C., Smith, D., Johnson, W. and Shively, R., 2017. Shaping trust through transparent design: theoretical and experimental guidelines. In *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA* (pp. 127-136). Springer International Publishing.

Lyons, J.B. and Guznov, S.Y., 2019. Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, *20*(4), pp.440-458.

Lyons, J.B. and Havig, P.R., 2014. Transparency in a human-machine context: Approaches for fostering shared awareness/intent. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 6* (pp. 181-190). Springer International Publishing.

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), pp.277-301.

Madhavan, P., Wiegmann, D.A. and Lacson, F.C., 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors*, *48*(2), pp.241-256.

MAIB (2004) 'Bridge watchkeeping safety study.', *Southampton: Marine Accident Investigation*, (July), pp. 1–24.

Marine Accident Investigation Branch (2015) *MAIBInvReport 20_2016 - Karen - Serious Marine Casualty*.

Marine Accident Investigation Branch (2020) *MAIB Report - Stena Superfast VII and Royal Navy Submarine*. Southampton.

MaritimeEuropeanAgencySafety (2010) 'MARITIME ACCIDENT REVIEW 2010 Work Programme

2010', *Maritime Accident Review 2010,* pp. 1–32.

Matthews, M.L., Bos, J., Crebolder, J.M. and McFadden, S., 2005, September. Modeling the effectiveness of tools to assist sonar operators. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 12, pp. 1124-1128). Sage CA: Los Angeles, CA: SAGE Publications.

Mayer, R.C., Davis, J.H. and Schoorman, F.D., 1995. An integrative model of organizational trust. *Academy of management review*, *20*(3), pp.709-734.

dos Santos Mello, V., de Moura, N.N. and de Seixas, J.M., 2018, July. Novelty detection in passive sonar systems using stacked autoencoders. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D. and Procci, K., 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors*, *58*(3), pp.401-415.

Merritt, S.M. and Ilgen, D.R., 2008. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, *50*(2), pp.194-210.

Meyer, J., 2001. Effects of warning validity and proximity on responses to warnings. *Human factors*, *43*(4), pp.563-572.

Barnes, M.J., Chen, J.Y., Jentsch, F., Oron-Gilad, T., Redden, E., Elliott, L. and Evans III, A.W., 2014. Designing for humans in autonomous systems: Military applications. *DTIC Document, January*.

Molloy, R. and Parasuraman, R., 1996. Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, *38*(2), pp.311-322.

Mori, M., MacDorman, K.F. and Kageki, N., 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), pp.98-100.

Mou, J.M., Van der Tak, C. and Ligteringen, H., 2010. Study on collision avoidance in busy waterways by using AIS data. *Ocean Engineering*, *37*(5-6), pp.483-490.

De Moura, N.N., De Seixas, J.M. and Ramos, R., 2011. Passive sonar signal detection and

classification based on independent component analysis. In *Sonar Systems* (pp. 93-103). Makati, Philippines: InTech.

Muir, B.M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), pp.1905-1922.

Muir, B.M. and Moray, N., 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), pp.429-460.

Murphy, M. (2000) 'Dead-Reckoning Systems', in *Operations Specialist, Volume 1*. Pensacola, FL: Naval Education and Training Professional Development Center.

Murphy, R.R., 2000. Marsupial and shape-shifting robots for urban search and rescue. *IEEE Intelligent Systems and their applications*, *15*(2), pp.14-19.

Nass, C. and Moon, Y., 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), pp.81-103.

Norman, D.A., 1984. Cognitive engineering principles in the design of human-computer interfaces. *Human-Computer Interaction. Amsterdam: Elsevier*, pp.11-16.

National Transportation Safety Board, 1997. Grounding of the Panamanian passenger ship Royal Majesty on Rose and Crown shoal near Nantucket, Massachusetts, June 10, 1995. *Marine Accident Report*.

O.Gibson, H. (2007) *Displaying Uncertainty: A Comparison Between Submarine Subject Matter Experts, the Total Army Competitive Category Optimization Model: Analysis of U.S. Army Officer Accessions and Promotions*.

Oleson, K.E., Billings, D.R., Kocsis, V., Chen, J.Y. and Hancock, P.A., 2011, February. Antecedents of trust in human-robot collaborations. In *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (pp. 175-178). IEEE.

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), pp.476-488.

Ososky, S., Sanders, T., Jentsch, F., Hancock, P. and Chen, J.Y., 2014, June. Determinants of

system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Unmanned systems technology XVI* (Vol. 9084, pp. 112-123). SPIE.

Parasuraman, R. and Manzey, D.H., 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors*, *52*(3), pp.381-410.

Parasuraman, R., Molloy, R. and Singh, I.L., 1993. Performance consequences of automation-induced'complacency'. *The International Journal of Aviation Psychology*, *3*(1), pp.1-23.

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), pp.286-297.

Parasuraman, R., Sheridan, T.B. and Wickens, C.D., 2008. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making*, *2*(2), pp.140-160.

Perneger, T.V., 2005. The Swiss cheese model of safety incidents: are there holes in the metaphor?. *BMC health services research*, *5*, pp.1-7.

Pope, K., Roberts, A. and Stanton, N., 2019. Investigating Temporal Implications of Information Transition in Submarine Command Teams. In *Advances in Human Aspects of Transportation: Proceedings of the AHFE 2018 International Conference on Human Factors in Transportation, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9* (pp. 243-253). Springer International Publishing.

Press, T.A. and Estacio, T. (2022) *Thanksgiving Bay Bridge crash may have involved Telsa in self-driving mode, KRON4*. Available at: https://www.kron4.com/news/bay-area/thanksgiving-bay-bridge-crash-may-have-involved-telsa-in-self-driving-mode/ (Accessed: 21 December 2023).

Rasmussen, J., 1997. Risk management in a dynamic society: a modelling problem. *Safety science*, *27*(2-3), pp.183-213.

Rasmussen, J. and Svedung, I. (2000) *Proactive Risk Management in a Dynamic Society*, *Proactive Risk Management In a Dynamic Society*.

Reason, J., 1995. A systems approach to organizational error. *Ergonomics*, *38*(8), pp.1708-1721.

Reason, J. (1997) *Managing the Risks of Organizational Accidents*. Routledge.

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Rice, S., 2009. Examining single-and multiple-process theories of trust in automation. *The Journal of general psychology*, *136*(3), pp.303-322.

Rice, S. and Geels, K., 2010. Using system-wide trust theory to make predictions about dependence on four diagnostic aids. *The Journal of general psychology*, *137*(4), pp.362-375.

Rieger, T. and Manzey, D., 2022. Human performance consequences of automated decision aids: The impact of time pressure. *Human factors*, *64*(4), pp.617-634.

Roberts, A., Stanton, N. and Fay, D., 2015. The command team experimental test-bed stage 1: design and build of a submarine command room simulator. *Procedia Manufacturing*, *3*, pp.2800-2807.

Roberts, A.P. and Cole, J.C., 2018. Naturalistic decision making: taking a (cognitive) step back to take two steps forward in understanding experience-based decisions. *Journal of applied research in memory and cognition*, *7*(1), pp.70-81.

Roberts, A.P. and Stanton, N.A., 2018. Macrocognition in submarine command and control: a comparison of three simulated operational scenarios. *Journal of applied research in memory and cognition*, *7*(1), pp.92-105.

Roberts, A.P., Stanton, N.A. and Fay, D.T., 2018. Go Deeper, Go Deeper: Understanding submarine command and control during the completion of dived tracking operations. *Applied ergonomics*, *69*, pp.162-175.

Roberts, M. *et al.* (2004) 'EMC Analysis of Universal Automatic Identification and Public Correspondence Systems in the Maritime VHF Band.', *Report for the US Coast Guard/G-SCT-2 under contract DCA100-00-C-4012* [Preprint].

Rotter, J.B., 1954. Social learning and clinical psychology.

Rotter, J.B., 1967. A new scale for the measurement of interpersonal trust. *Journal of personality*.

Rousseau, D.M., Sitkin, S.B., Burt, R.S. and Camerer, C., 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review*, *23*(3), pp.393-404.

Rovira, E., McGarry, K. and Parasuraman, R., 2007. Effects of imperfect automation on decision making in a simulated command and control task. *Human factors*, *49*(1), pp.76-87.

Royal Navy (2017) 'BRD2: The Queens Regulations For the Royal Navy', in, pp. 1–11.

Salas, E., Dickinson, T.L., Converse, S.A. and Tannenbaum, S.I., 1992. Toward an understanding of team performance and training.

Salmon, P.M., Stanton, N.A., Walker, G.H., Baber, C., McMaster, R., Jenkins, D., Beond, A., Sharif, O., Rafferty, L. and Ladva, D., 2006, October. Distributed situation awareness in command and control: A case study in the energy distribution domain. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 3, pp. 260-264). Sage CA: Los Angeles, CA: SAGE Publications.

Schaefer, K.E., Chen, J.Y., 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), pp.377-400.

Schaefer, K.E., Evans III, A.W. and Hill, S.G., 2015. Command and control in network-centric operations: trust and robot autonomy. In *20th International Command and Control Research and Technology Symposium, Annapolis, MD*.

Schaekermann, M., Beaton, G., Sanoubari, E., Lim, A., Larson, K. and Law, E., 2020, April. Ambiguity-aware ai assistants for medical data analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Schlabach, J.L., Hayes, C.C. and Goldberg, D.E., 1999. FOX-GA: A genetic algorithm for generating and analyzing battlefield courses of action. *Evolutionary Computation*, *7*(1), pp.45-68.

Schunn, C.D., Kirschenbaum, S.S., Trafton, J.G. and Schunn, C., 2003. The ecology of uncertainty: Sources, indicators, and strategies for informational uncertainty. *Manuscript submitted for publication*.

Shar, P. and Li, X.R., 2000. Passive sonar fusion for submarine c/sup 2/systems. *IEEE Aerospace and Electronic Systems Magazine*, *15*(3), pp.29-34.

Sheridan, T.B. and Hennessy, R.T., 1984. Research and modeling of supervisory control behavior. Report of a workshop. *National Research Council Washington DC Committee on Human Factors, Tech. Rep*.

Siddiqui, F. and Merrill, J.B. (2022) *NHTSA launches probe into Tesla's 'Phantom Braking - The Washington Post, The Washington Post*. Available at: https://www.washingtonpost.com/technology/2022/02/17/tesla-phantom-braking/.

Solberg, E., Kaarstad, M., Eitrheim, M.H.R., Bisio, R., Reegård, K. and Bloch, M., 2022. A conceptual model of trust, perceived risk, and reliance on AI decision aids. *Group & Organization Management*, *47*(2), pp.187-222.

Sotiralis, P., Ventikos, N.P., Hamann, R., Golyshev, P. and Teixeira, A.P., 2016. Incorporation of human factors into ship collision risk models focusing on human centred design aspects. *Reliability Engineering & System Safety*, *156*, pp.210-227.

Stanton, N., Baber, C. and Harris, D., 2008. *Modelling command and control: Event analysis of systemic teamwork*. Ashgate Publishing, Ltd..

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), pp.43-68.

Stanton, N.A. *et al.* (2013) 'Repertory Grid Analysis', in *Human Factors Methods: A Practical Guide for Engineering and Design*. Second. Ashgate Publishing, pp. 458–464.

Stanton, N.A., 2014. Representing distributed cognition in complex systems: how a submarine returns to periscope depth. *Ergonomics*, *57*(3), pp.403-418.

Stanton, N.A., Jenkins, D.P., Salmon, P.M., Walker, G.H., Revell, K.M. and Rafferty, L.A., 2017. *Digitising command and control: a human factors and ergonomics analysis of mission planning and battlespace management*. CRC Press.

Stanton, N.A., Roberts, A.P., Pope, K.A. and Fay, D., 2022. The quest for the ring: a case study of a new submarine control room configuration. *Ergonomics*, *65*(3), pp.384-406.

Stanton, N.A. and Bessell, K., 2014. How a submarine returns to periscope depth: Analysing complex socio-technical systems using Cognitive Work Analysis. *Applied ergonomics*, *45*(1), pp.110-125.

Stanton, N.A. and Roberts, A.P., 2020. Better together? Investigating new control room configurations and reduced crew size in submarine command and control. *Ergonomics*, *63*(3), pp.307-323.

Stanton, N.A., Roberts, A.P. and Fay, D.T., 2017. Up periscope: understanding submarine command and control teamwork during a simulated return to periscope depth. *Cognition, Technology & Work*, *19*, pp.399-417.

Svedung, I. and Rasmussen, J., 2002. Graphic representation of accidentscenarios: mapping system structure and the causation of accidents. *Safety science*, *40*(5), pp.397-417.

Tabibzadeh, M. and Meshkati, N., 2015, April. Applying the AcciMap methodology to investigate a major accident in offshore drilling: A systematic risk management framework for oil and gas industry. In *SPE Western Regional Meeting*. OnePetro.

UK Development Concepts And Doctrine Centre (2018) *Human-Machine teaming*, *Joint Concept Note 1/18*.

UK MOD (2018) 'Information Advantage: Joint Concept Note 2/18', *The Development, Concepts and Doctrine Centre*, pp. 1–27.

De Visser, E.J., Monfort, S.S., McKendrick, R., Smith, M.A., McKnight, P.E., Krueger, F. and Parasuraman, R., 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), p.331.

De Visser, E.J., Pak, R. and Shaw, T.H., 2018. From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), pp.1409-1427.

de Visser, E.J. and Parasuraman, R., 2007, October. Effects of imperfect automation and task load on human supervision of multiple uninhabited vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 51, No. 18, pp. 1081-1085). Sage CA: Los Angeles, CA:

SAGE Publications.

Walker, G.H., Stanton, N.A., Stewart, R., Jenkins, D., Wells, L., Salmon, P. and Baber, C., 2009. Using an integrated methods approach to analyse the emergent properties of military command and control. *Applied Ergonomics*, *40*(4), pp.636-647.

Walker, G.H., Jenkins, D., Young, M.S., Stewart, R. and Wells, L., 2010. A human factors approach to analysing military command and control.

Wang, Y.H., Ou, Y., Deng, X.D., Zhao, L.R. and Zhang, C.Y., 2019, June. The ship collision accidents based on logistic regression and big data. In *2019 Chinese Control And Decision Conference (CCDC)* (pp. 4438-4440). IEEE.

Washington, D.C. (no date) *National Transportation Safety Board Marine Accident Brief*.

Wickens, C.D. and Dixon, S.R., 2007. The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), pp.201-212.

Woods, D. and Dekker, S., 2000. Anticipating the effects of technological change: a new era of dynamics for human factors. *Theoretical issues in ergonomics science*, *1*(3), pp.272-282.

Woods, D.D., Patterson, E.S. and Roth, E.M., 2002. Can we ever escape from data overload? A cognitive systems diagnosis. *Cognition, Technology & Work*, *4*, pp.22-36.

Yoo, B., Donthu, N. and Lenartowicz, T., 2011. Measuring Hofstede's five dimensions of cultural values at the individual level: Development and validation of CVSCALE. *Journal of international consumer marketing*, *23*(3-4), pp.193-210.

Zhang, A. and Yang, Q., 2022, February. To Be Human-like or Machine-like? An Empirical Research on User Trust in AI Applications in Service Industry. In *2022 8th International Conference on Automation, Robotics and Applications (ICARA)* (pp. 9-15). IEEE.

APPENDICES

Appendix A: CDM Questions

Question	**Question/Probe Type**
Can you think of a specific example where you have been classifying multiple contacts?	Scenario set-up
Pick a specific example, perhaps a training exercise. Where is it?	Scenario set-up
And what about the quantity of other boats?	Scenario set-up
Think about just before that happened. In the time building up to that what are you following, what are your expectations?	Analogue Probe
Does the planning meeting occur when you are already on the [redacted]?	Clarification
The information displays you are using – the SONAR – are they all in the same area, in the same room?	Clarification
Is all that occurring in the same space?	Clarification
So this is onboard a t-boat?	Clarification
You're communicating over a head-set?	Clarification
So the first thing that happens is that you are given your mission objectives, and then you try to spot these contacts on the SONAR display and then classify them?	Scenario set-up
Go back to the plan. Where does that start?	Scenario set-up
What kind of things would you do to prepare?	Knowledge probe

So the OOW has a route that may be taken – but the other information and the decisions being made at the planning meeting – do the specialists get responsibility for their areas?	Clarification
Do you create any resources at this point, like maps for example?	Scenario set-up
From there, when you start the actual mission, what's the first thing that would happen? Looking at a general idea of what's in the area. Checking it matches what you expect?	Scenario set-up
So is that all communicated from the sound room?	Clarification
what's the defining rule for when you're happy to pass that contact information on?	Knowledge probe
Where is the DEMON?	Clarification
Do you have a contingency plan?	Basis probe
What is "in-out"?	Clarification
Classification is a constantly ongoing task?	Goal probe
So each one [contact] is assigned to a channel, and you keep monitoring the channel?	Clarification
The job is specifically to manage the contacts?	Goal probe
Let's talk more about what a track means, and whether everybody is seeing the same visualisation of a track, or if they have different ways of displaying and interpreting that.	Knowledge probe
What does the track screen look like?	Clarification
Is the display the same in every part of the submarine?	Clarification
In the sonar room, you don't see the colours, you just see the tracks?	Clarification
Does the OOW have a copy of your screen?	Clarification

Does the OOW have to choose which screens to look at, or can they display more than one screen?	Clarification
Let's go back to one of the very wide tracks obscuring the others. Let's use that as a specific case. When you are acting as a Sonar Operator.	Scenario set-up
You're sat in the sound room focusing on your screen in front of you. You can see a wide track with the side lobes. What else is around you?	Cues probe
Who is sat next to you?	Cues probe
Who do you report that to?	Scenario set-up
What happens next?	Scenario set-up
Where do you think they [contact] will go?	Basis probe
Are you interested when it isn't visible?	Knowledge probe
Do you have a prediction of where it will go next?	Options probe
Does this make you panic?	Cues probe
What do you do now you've lost the track?	Knowledge probe
Can you see the trajectory you think it will follow?	Clarification
What guidelines do you have on the screen for where you think it will come back out?	Knowledge probe
You are holding that information just in your head?	Options probe
How do you know where to look in order to re-gain the track?	Cues probe
Is this from training? From experience?	Experience probe
Your experience tells you where it will re-emerge?	Experience probe

What about the fishing vessel?	Clarification
You're hearing all of this on the sonar?	Clarification
What standard scenario do you have for these large areas of uncertainty?	Knowledge probe
Do you usually find all of your tracks again quickly?	Analogue Probe
Is it rare to lose the tracks?	Clarification
Why and how do you know that (where it will emerge)?	Knowledge probe
What extra support, if any, do you have in that situation?	Aiding probe
Think back to when you're struggling to classify something. The Operator would go to the Controller?	Options probe
If the track disappears, and then you acquire a new track, is there a possibility that the new one is actually the old one?	Basis probe
What is your main objective in this scenario?	Goal probe
Do you classify everything?	Knowledge probe
Are you talking on the head-set at the same time?	Clarification
What set procedures does the OOW have for suddenly changing direction?	Goal probe
Is there a standard procedure for when you cannot re-track a contact?	Knowledge probe
This is decided by the OOW?	Clarification
Where does the recording go from the sound room?	Clarification
Track 03 at close quarters, when it moves into the broad tracks, the manoeuvring was deceptive – what would you be able to do if you had lost the track, but still anticipated what could happen?	Hypothetical

They're sitting on the same bearing, so what do you do now?	Situation assessment
What specifically is concerning you?	Situation assessment
What is the SMCS screen telling you about spatial arrangement that you're not getting from the sonar screen?	Knowledge probe
How much time pressure is there?	Time pressure probe
At what frequency do you look again?	Cues probe
Which is worse; having lots of tracks you know on the screen, so you've got the workload of dealing with lots of things, or dealing with a few, but ambiguous, or unknown tracks?	Hypothetical
What strategies can you use to get rid of some of the noise?	Knowledge probe
What can you do to filter if there is masking occurring?	Knowledge probe
Let's say T0 is when you identified the track. Focusing on something ambiguous. It's new. You decide it's definitely a fishing vessel. Let's call that T0. So, what's happening before that?	Situation assessment
T minus-one is when the track appears?	Scenario set-up
T minus-two, before the track appearing?	Scenario set-up
All of these things, about changing direction, changing depth, or the target moving behind a land mass, or coming into range, these are all things that are happening prior to the track appearing. So, I think listing those would be really good. Can you list them in order?	Hypothetical

In the planning meeting, do you come up with lists of potential contacts?	Clarification
What would be the first thing you do in that instance [unexpected contact at t-zero)?	Hypothetical
When do you start tracking the revs?	Knowledge probe
What can you hear?	Cues probe
What can you see?	Cues probe
What is the first thing you look at (when classifying)?	Knowledge probe
List me the characteristics?	Knowledge probe
What is the most important thing, revs or blades?	Knowledge probe
Is this when you can possibly classify it?	Clarification
Work out the CPA, the speed, and then you know when you need to change course?	Clarification
If you're a beginner, what are the main things you would look for?	Experience probe
What are you listening for?	Cues probe
Listening using the sonar?	Clarification
Where are you noticing the revs, blades and shafts?	Knowledge probe
Is it through training you can recognise toughly what the ship's characteristics are?	Experience probe
Do you have tables to refer to (rev ranges)?	Clarification

Is it possible to classify just from that, for example, a large merchant vessel, or do you need more information?	Experience probe
So blades and shafts, just if it's ambiguous?	Knowledge probe
Is that the bare amount of information you need?	Experience probe
How do you validate that classification?	Knowledge probe
It isn't confirmed until the OPSO reports it?	Clarification
What if you don't know what it is from the TPK?	Hypothetical
What happens if the OPSO looks and it isn't in the shipping lane?	Hypothetical
Is there a time limit?	Time pressure probe
Are there times you can think of where you've been unable to classify a contact?	Experience probe
Do you monitor a contact differently when you aren't sure what it is?	Options probe
Will the warship try and disguise some information?	Hypothetical
What stuff (signature information) is more difficult for a contact to obscure?	Experience probe
You can only use passive sonar when being covert?	Clarification
Does that make a big difference in being able to classify?	Experience probe
Is the big difference in terms of expertise familiarity?	Experience probe

How do you look after somebody new?	Experience probe
Give me examples of the kind of help you would give them?	Aiding probe
Trying to keep tracks from the edge?	Aiding probe
How long would it take in terms of hours at sea before they would be comfortable working on their own?	Experience probe
Do you specialise in one (type of sonar) or have to learn everything?	Clarification
Then you're assigned?	Clarification
Once you've done your compulsory number of hours, do you pick what to specialise in?	Clarification
Do broadband and narrowband track the same contacts?	Clarification
If you're unsure from the broadband, is gaining information from the narrowband the next step?	Situation assessment
If you're not sure what a contact is, what are the steps you go through in your head to make it clearer?	Situation assessment
You look for that visually?	Knowledge probe
Is it a rare event when something is classified as unknown?	Knowledge probe
Is it a very bad event if something is classified as unknown?	Knowledge probe
What strategy would you use (if a contact is unknown) – raising the periscope?	Goal probe
If the Sonar Controller was not a human being, but a computer, with a holographic display, is that ludicrous?	Hypothetical

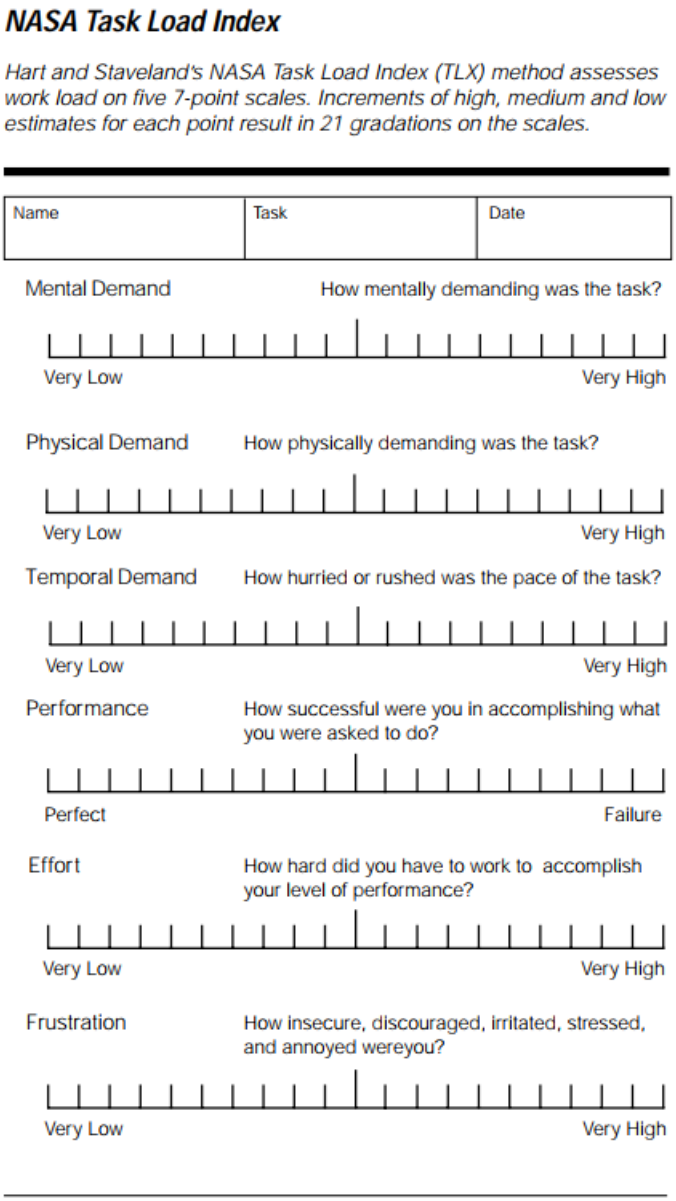
Would you take an order from it?	Hypothetical
What about, in the example of an experienced operator leaning over to a trainee, can you imagine that as an avatar, or some other way of drawing attention to some aspect which may not have been done properly?	Hypothetical
Would this scenario be practical? It may take up screen space. What if it communicated via voice? Or text on the screen?	Hypothetical
Monitoring the sonar, if you had an autonomous system, how would it communicate to the Controller, should it have a voice? A face? Maybe just highlight things on the screen?	Hypothetical
Thinking about how to incorporate feedback from the system on the screen, as part of the display. A little dot on the track, perhaps? Would it be easy to mistake that for tracking?	Hypothetical

Appendix B: Semi-Structured Interview Questions

Questions asked in semi-structured interview

1) Could you tell the difference between the blocks?
2) Which of the blocks do you think was the least reliable?
3) Why?
4) Which of the blocks do you think was the most reliable?
5) Why?
6) Which pieces of information on the screen were most informative?
7) Which pieces of information did you use the most?
8) Was this true for each block?
9) Did you have a strategy for making your decisions?
10) What did you do when the pieces of information did not match?
11) How did you decide whether to accept or reject the suggested classification?
12) Do you think the computer was good or bad at classifying things?
13) Why?

Appendix C: NASA TLX Questionnaire



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Appendix D: Checklist Between People and Automation Questionnaire

Checklist for Trust between People and Automation

Below is a list of statement for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please mark an "x" on each line at the point which best describes your feeling or your impression.

(Note: not at all=1; extremely=7)

1 The system is deceptive

1 2 3 4 5 6 7

2 The system behaves in an underhanded manner

1 2 3 4 5 6 7

3 I am suspicious of the system's intent, action, or outputs

1 2 3 4 5 6 7

4 I am wary of the system

1 2 3 4 5 6 7

5 The system's actions will have a harmful or injurious outcome

1 2 3 4 5 6 7

6 I am confident in the system

1 2 3 4 5 6 7

7 The system provides security

1 2 3 4 5 6 7

8 The system has integrity

1 2 3 4 5 6 7

9 The system is dependable

1 2 3 4 5 6 7

10 The system is reliable

1 2 3 4 5 6 7

11 I can trust the system

1 2 3 4 5 6 7

12 I am familiar with the system

1 2 3 4 5 6 7

Appendix E: Participant Consent Form

Consent Form

PhD Research Experiment: Evaluating trust in multi-sensor systems

	Please tick to confirm:
I am over 18 years old	
I have read and understood the information sheet	
I have been given the opportunity to ask questions about the study	
I understand that I am able to withdraw myself from the experiment at any time during, or afterwards by emailing [redacted]	
I understand that all data recorded shall be anonymised, and will not be linked to me **I agree to the anonymous data collected from me being used for research purposes**	

Name: _____

Signature: _____

Date: _____

Appendix F: Participant Information Sheet

Experiment Information Sheet

PhD Research Experiment – Using Autonomous Aids to Classify Aural Information

Thank you for agreeing to take part in this study!

Overview: This research is concerned with how classification decisions can be made with the help of an autonomous classifier. This experiment is testing a new autonomous AI that classifies sounds. These sounds are hydrophone recordings of ships and marine life, and the computer will try and classify these recordings. It will offer you a suggested classification, which you can accept or reject.

To help with this, you can listen to the hydrophone recording of the contact, see a spectrogram (frequency information) of the hydrophone recording, see suggested classifications provided by the autonomous classifier, and are also provided with a unique type of graphical display known as a VINAS. The VINAS is a colourful grid. Each type of contact makes a different coloured pattern on the grid. A key to understand the VINAS is provided below, and you will also see this key on your screen. Some of the information for each contact may be incorrect – it is up to you to identify which information sources you can trust.

Before you begin, you will be asked to fill in a questionnaire about how trusting you tend to be.

You will then be asked to classify five contacts at a time. After each classification, you will be asked to rate your confidence in your decision. After classifying five contacts, you will be asked to fill in a two questionnaires about workload and trust, and can have a short break. Once all contacts have been classified, we will have a short discussion about your experience.

During the task, some performance measures will be collected, including information about time taken to complete the task, whether the classification decision was correct, and the self-confidence ratings, as well as the questionnaire data.

Time: This experiment will take around half an hour to complete. During this time, you can take a break from the screen when desired.

Data Collection: All of the data collected will be anonymised and you will not be identified in any processing of this data. Data will be stored on an encrypted hard drive which is kept in a secure location. If you choose to withdraw from the study, all of your data will be destroyed. You can withdraw from the study up to one month after performing the experiment by contacting me using the details provided below. Data will be used solely for the purpose of this PhD research experiment and may be published in academic journals or conferences. Further details and copies of published results will be provided upon request.

Risks: The experiment involves using a screen and headphones for an extended period of time, which may result in eye strain or fatigue. If you would like to take a break at any point, please ask. There will be a break provided before completing the end of experiment questionnaires.

Withdrawal: If you wish to withdraw from the study, or would like further details, please send an email to: [redacted]

Contact Details: Faye McCabe, LG07, UKRRIN

[redacted]