Management of multiple heterogeneous unmanned aerial vehicules through transparency capability

Ting Chen

UNIVERSITE
BRETAGNE
LOIRE

THÈSE / Télécom Bretagne
sous le sceau de l'Université Bretagne Loire

pour obtenir le grade de Docteur de Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Sicma
et en cotutelle avec Queensland University of Technology (Australie)
Mention : Sciences et Technologies de l'Information et de la Communication

présentée par

Ting (Brendan) CHEN
préparée dans le département Logique des usages, sciences sociales et de l'Information
Laboratoire Labsticc

Management of Multiple Heterogeneous Unmanned Aerial Vehicles Through Transparency Capability

Thèse soutenue le 26 février 2016
Devant le jury composé de :

Colin Fidge
Professeur, Queensland University of Technology / président

Jill Drury
Docteur - Chercheur, The Mitre Corporation - USA / rapporteur

Stacey Scott
Associate professor, University of Waterloo, Canada / rapporteur

Jessie Chen
Docteur - Chercheur, US Army, USA / rapporteur

Felipe Gonzalez
Assistant professor, Queensland University of Technology / examinateur

Thierry Duval
Professeur, Télécom Bretagne / examinateur

Duncan Campbell
Professeur, Queensland University of Technology / co-directeur de thèse

Gilles Coppin
Professeur, Télécom Bretagne / directeur de thèse

N° d'ordre : 2016telb0423

Sous le sceau de l'Université Bretagne Loire

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Sicma

Co-tutelle avec Queensland University of Technology (Australie)

Management of Multiple Heterogeneous Unmanned Aerial Vehicles Through Transparency Capability

---

Thèse de Doctorat

Mention : Sciences et Technologies de l'Information et de la Communication

Présentée par **Ting (Brendan) Chen**

Département : Logique des usages, sciences sociales et de l'Information (LUSSI)

Laboratoire : Lab-STICC

Directeur de thèse : Gille Coppin

Soutenue le 26 février 2016

Jury :

M. Colin Fidge, Queensland University of Technology (Président)
Mme Jill Drury, Docteur - Chercheur, The Mitre Corporation, USA (Rapporteur)
Mme Stacey Scott, Associate professor, University of Waterloo, Canada (Rapporteur)
Mme Jessie Chen, Docteur - Chercheur, US Army, USA (Rapporteur)
M. Felipe Gonzalez, Associate professor, Queensland University of Technology (Examinateur)
M. Thierry Duval, Professeur, Télécom Bretagne (Examinateur)
M. Duncan Campbell, Professeur, Queensland University of Technology (Co-directeur de thèse)
M. Gilles Coppin, Professeur, Télécom Bretagne (Directeur de thèse)

Management of Multiple Heterogeneous Unmanned Aerial Vehicles Through Capability Transparency

by

Ting (Brendan) Chen

Bachelor of Engineering (Aerospace Avionics) (Hons)

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Electrical Engineering and Computer Science, Départemente de logiques des usages, sciences sociales et sciences l'information

Queensland University of Technology, Télécom Bretagne

2016

"Research is to see what everybody else has seen and to think what nobody else has thought."

— Albert Szent-Györgyi, 1988

Keywords

Human Machine Interaction, Human Machine Interface, User Interface, Autonomy Transparency, Capability Transparency, Functional Capability, Hybrid System, Unmanned Aerial Vehicle, Heterogeneous UAVs, Multiple UAVs, Autonomy Management, Model-View-Controller, Presentation-Abstraction-Control, Ground Control Station, Cognitive Performance, Situation Awareness, Cognitive Workload, Automation Trust, Objective Performance.

Abstract

The benefits of using Unmanned Aerial Vehicles (UAVs) to perform tasks such as search and rescue, biosecurity, air quality sampling and wildlife monitoring are now well recognised. One advantage in particular is the ability to remove humans from operations that are life-threatening, costly and arduous.

The deployment of multiple heterogeneous UAVs to perform a mission has multiple benefits. Firstly, despite rapid advancements in UAV autonomy and payload capabilities, a single UAV can only give one point-of-view in any one instance. This vastly reduces the human operator's capacity to acquire situational information about the environment from multiple perspectives, adding to the reduction in his/her situational awareness which resulted from removing them from the mission scene. Secondly, as UAV technology becomes more complex, fault modalities increase within systems. Consequently, deploying multiple UAVs increases redundancy and reduces the possible impact to the overall mission performance.

Currently, two or more human operators are required to operate a single UAV for complex missions such as battlefield reconnaissance or search and rescue. This creates a challenging scenario for the mission commander and the operator(s) to work collaboratively when multiple heterogeneous UAVs are deployed, as this creates a disproportionate ratio of Operator to UAV management (*n:1*). Effective management of multiple heterogeneous UAVs can be achieved by inverting the management ratio from multiple operators managing one UAV, to one operator managing multiple UAVs (*1:n*). However, numerous hardware/physical and human cognitive challenges may arise as a result. These challenges are being addressed through research which is aimed at increasing the UAV's autonomous capability and improve Ground Control Station (GCS) designs.

This dissertation addresses the three main aspects of the cognitive challenges associated with one operator managing multiple UAVs (*1:n*): Operator Cognitive Workload (CW), Situation Awareness (SA), and Automation Trust.

Novel contributions made in this dissertation are: A framework and its derivation procedure to encapsulate a UAV's subsystem functional capability; an approach to support the operator's understanding of UAV autonomy capability; a software approach which implements the new UI design methodology; and validating experimental designs, procedures and operator performance analysis.

These contributions are accomplished through increasing the UAV's functional subsystem and autonomy capability transparency by leveraging the traditional *Presentation-Abstraction-Control* (*PAC*) UI design paradigm, and proposing an improved implementation to incorporate the UAV's functional subsystem and autonomy capability status. This

results in an increase in the systems' autonomy transparency. A Functional Capability Framework (FCF) is proposed to categorise a UAV's functional subsystem capabilities, and increase the UAV's autonomy capability transparency. These are tested by communicating the UAV's autonomy status to the human operator.

The FCF organises the UAV's functional subsystems into two dimensions; its nature of functionality and it's level of information aggregation abstraction. The framework presents the UAV subsystem information in a categorical (nature of functionality) and hierarchical (information aggregation) manner such that the cognitive performance of the human operator is increased when managing multiple UAVs, compared to unstructured data directly from the UAVs.

Knowledge, understanding and communication of the UAV's autonomy capability information is crucial for the human operator to establish an internal status model of the UAVs improving their cognitive performance when managing multiple heterogeneous UAVs. This was achieved by creating a graphical and natural language (textual) representation of the information, where a textual message box with messages stating the sender, its decisions and intentions is displayed.

The improved *PAC* approach, which consists of the FCF and autonomy transparency, was verified with three experiments: 1) Evaluation of the operator's cognitive performance when using the FCF, 2) Evaluation of the operator's cognitive performance through graphical representation of the UAV's autonomy, and 3) Evaluation of the operator's cognitive performance through natural language information exchange. Results collected from these experiments yielded an improvement to the operators' cognitive performance when using the improved *PAC* model UI design configuration with UAV functional subsystem and autonomy capability status, compared to the traditional *PAC* approach. This suggests that by increasing a UAV's capability (functional and autonomy) transparency in a multiple heterogeneous UAV system, the human operator demonstrated a reduced CW, an improved SA, and an improved reaction time.

Acknowledgments

Before I go into the main portion of this acknowledgement, I would like to first express my utmost admiration for Professor Rodney Walker (1969 - 2011), Founder of the Australian Research Centre of Aerospace Automation (ARCAA), Queensland University of Technology (QUT), and my lecturer in my undergraduate course. Rod was a pioneer in the Australian Unmanned Aircraft Systems (UASs) industry, and one of the significant motivators for me entering into a PhD candidature. The quote presented in the earlier pages of this dissertation is written in honour of Rod, as this was his favour PhD dissertation quote, and a quote that was introduced to me by Rod.

Firstly, I would like to thank my supervisory team, Professor Duncan Campbell, Professor Gilles Coppin, and Associate Professor Felipe Gonzalez for the endless hours of support and encouragement they have put toward my Joint-PhD Cotutelle[1] candidature. I would like to acknowledge the opportunity to be able to undertake this Cotutelle Candidature with Télécom Bretagne and QUT, and the opportunity to be able to spend fourteen (14) months under the direct supervision of Prof. Gilles Coppin of Lab-STICC in Télécom Bretagne, France. A number of research students I would also like to acknowledge: Mr Yoann Torres, Mr Julien Gaye, and Mr Botu Sun, from Télécom Bretagne, as well as my best friend and best man, Mr Benjamin Rudder from University of Stuttgart, Germany. They had tirelessly assisted me in the software development phase of this candidature whilst in Australia and France.

Secondly, I would like to acknowledge and thank the industry partners who have contributed intellectually and financially to this research and this candidature. Thales Netherlands and Télécom Bretagne had financially supported my exchange trip to France and the Netherlands, and gracefully hosted my stay to study in Lab-STICC (Télécom Bretagne, France) and D-CIS Labs (Delft, The Netherlands). Associate Professor Ben Upcroft from Robotics and Autonomous Systems (RAS), QUT had provided financial assistance. Furthermore, Dr Martijn Mooij and Dr Stas Krupenia had made significant intellectual contributions in the early phases of my candidature. They have helped me to gain insightful understandings to the practical aspects of working in the field of human machine interaction.

Thirdly, I would like to thank my host universities (Télécom Bretagne, France and QUT,

[1]The French-Australian Cotutelle is a Joint-PhD candidature programme where the candidate spends time in two universities, the Australian University being the main institution, and the French University being the host institution. The candidate must spend one third of the candidature time in the host university (Télécom Bretagne, France) to conduct research activities, and the candidature is to be assessed by both universities in accordance with the institution specific policy. Upon graduation, the candidate receives two PhD certificates, one from each university.

Australia) and the specific research laboratories: ARCAA, Brisbane, Australia and Lab-STICC, Brest, France for providing me with valuable facilities and resources to conduct research and perform data collection.

Finally, I would like to thank my parents Professor Shaofeng Chen and Ms Yan Nan for providing not only endless emotional support, but also many years of financial support. Although you cannot provide me with any intellectual support, but the long duration and the hardship form a large portion of my experience as a PhD candidate. Without your emotional and financial support, there will be no conclusion to this candidature. Last but not least, my wife, Mrs Shiwenyu Zhu, who entered my life in the later stages of my candidature. She has had a significant influence on me and had bumped me into an unimaginable stage of my life. Without her understanding and her lead-by-example attitude, this candidature also would not conclude.

Ting (Brendan) Chen

Queensland University of Technology, Télécom Bretagne

November 2016

Contents

List of Figures

List of Tables

List of Abbreviations

3Ps	Purpose, Process, and Performance
ACL	Autonomy Control Level
ADT	Air Data Terminal
ADU	Air Data Unit
AH	Abstraction Hierarchy
ALFUS	Autonomous Levels For Unmanned Systems
AUD	Automation Usage Decision
AV	Air Vehicle
AVC	Avionics Computer
BDI	Belief-Desire-Intention
C2	Command and Control
CAC	Contextual Autonomous Capability
COUNTER	Cooperative Operations in Urban Terrain
CSE	Cognitive Systems Engineering
CW	Cognitive Workload
DCPA	Digital Central Processing Assembly
DoD	Department of Defense
EC	Environmental Complexity
EID	Ecological Interface Design
ERT	Event Reaction Time
ETS	Engine Temperature Sensor
F-LOA	Functional LOA
FCF	Functional Capability Framework
FLS	Fuel Level System
FVU	Flux Valve Unit
GCS	Ground Control Station
GDT	Ground Data Terminal
GPS	Global Position Sensor
GPWS	Ground Proximity Warning System
GSE	Ground Support Equipment

GUI	Graphical User Interface
HACT	Human-Automation Collaboration Taxonomy
HI	Human Independence
IFF	Idenitication of Friend or Foe
IMU	Inertial Measuring Unit
IRT	Initial Reaction Time
LOA	Level Of Autonomy
MC	Mission Complexity
MISA	Mixed-Initiative Sliding Autonomy
MIT	Massachusetts Institute of Technology
MVC	Model-View-Controller
NASA	National Aeronautics and Space Administration
NASA-TLX	NASA Task Load Index
NIST	National Institute of Standards and Technology
NWS	Nose Wheel Steering
OODA	Observe-Orient-Decide-Act
PAC	Presentation-Abstraction-Control
PRCS	Portable Ruggedized Control Station
RPAV	Remotely Piloted Aerial Vehicle
SA	Situation Awareness
SAGAT	SA Global Assessment Technique
SART	Situation Awareness Rating Technique
SAT	SA-based Agent Transparency
SISA	System-Initiative Sliding Autonomy
SKR	Skills, Rules, Knowledge
SS	Scale Score
SV	Sheridan and Verplanck
SWAT	Subjective Workload Assessment Technique
TLX	Task Load Index
U.S.	United States
UAS	Unmanned Aircraft System
UAV	Unmanned Aerial Vehicle
UMS	Unmanned System
VGU	Vertical Gyro Unit
VSCS	Vigilant Spirit Control Station
WP	Workload Prole

Statement of Original Authorship

The work contained in this Joint PhD thesis undertaken between QUT and Télécom Bretagne has not been previously submitted to meet requirements for an award at these or any other higher education institutions. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: 

Date: 31 October 2016

CHAPTER 1

Introduction

Robots are motorised craft, designed to perform certain tasks, they can be aquacraft (water based), terrestrial craft (land based), or aircraft (airborne). An aerial robot can be a Remotely Piloted Aerial Vehicle (RPAV), which enables teleoperation of the robot, or Unmanned Aerial Vehicle (UAV). The difference between a UAV and an RPAV is that a UAV contains certain levels of intelligence that allows it to perform tasks autonomously[1], whereas a teleoperated aerial robot is only capable of being operated via remote control.

Traditionally, robots have been introduced to remove human presence from operations that are arduous/costly, and potentially life-threatening, including (but not limited to) casualty extraction, explosive detection and disposals, battlefield reconnaissance and surveillance, rural/urban search and rescue, emergency response, large-scale firefighting, and crop inspection. [41, 90, 214, 178, 179, 60]. Further benefits have been observed, for example, long hours of conventional vehicle operation can introduce significant mental fatigue to the human operator[2], while robotic and teleoperated craft are less mentally stressful to the human operator. Another example is that the financial requirements are less demanding to produce platforms which are not required to accommodate human presence.

Robotic platforms are commonly not equipped with the capability to perform every type of task. Tasks such as large-scale firefighting, rural search and rescue, agricultural inspections, and battlefield reconnaissances will require airborne solutions. The Unmanned Aircraft System (UAS) sector is at the forefront in terms of dynamic growth, in the aerospace industry [4].

1.1 Multiple Heterogeneous UAVs

In the recent decade, significant advancements in UAV capabilities [94] have been seen through the development of bigger and more robust platforms [1, 131], the integration of more capable payloads [7, 91, 92, 228], and the increments of onboard intelligence [227, 142]. This sparked significant interest in co-locating multiple UAVs in a close proximity to

[1]The word *Autonomy* stems from Greek, where *autos* denotes "self" and *nomos* denotes "law" [13].

[2]The *human operator* is the person who remotely pilots the UAV in the traditional sense, while in the situations where one operator is to manage multiple UAVs, he/she then becomes the mission commander/supervisor.

perform more complex missions.

1.1.1 Benefits & Potentials

The fundamental advantage for deploying multiple UAVs with heterogeneous capabilities in close proximity is to harness the strengths of each (UAV) agent. Since each UAV pertains a set of specialised capabilities, a dynamic team of hybrid[3] agents is formed (e.g. [145]). Through this, two clear benefits have been realised: mission capability, and fault tolerance/flexibility [160].

Mission Capability

Despite of the large array of available payloads, platform performances, and onboard computational capacities, a single highly specialised UAV has the capability to only complete a single mission - the mission that the UAV was deployed to perform. Furthermore, it can only provide a perspective from one point of view. On the contrary, a team of multiple heterogeneous UAVs working in close proximity extends the capability of a single UAV. This extension enables information collection, aggregation, and exploitation from various different locations simultaneously. As a result, human operators can acquire greater awareness of the situation and the mission, subsequently promoting more effective decision-making [160].

Fault Tolerance and Flexibility

Although advancements in Unmanned Aircraft System (UAS) technology mean greater reliability and integrity in the system, system complexity and versatility inherently contribute to faults and inflexibilities in the UAV platforms. A mission which involves multiple heterogeneous UAVs can promote fault tolerance from a mission perspective simply through the redundancy of UAVs in the team, and improve flexibility through the heterogeneous capabilities of the different UAVs [160].

To this end, the research to utilise UAVs in teams is primarily driven by the military domain [39, 145, 58, 128]. For example, the United States military, realising these benefits and potentials to their operations, has published its plans in the *Roadmap for UAS 2010-2035* to invest in the future of multiple UAS research and development; their aim is to support multiple UASs from a common ground station unit [2]. However, the civilian sector has also realised the benefits in deploying multiple UAVs for a wide range of missions, such as aerial mapping [200, 199], extended communication [217, 67], forest fire monitoring [31, 32], and rural search and rescue [93, 3].

However, many of these applications are only either in the early stage of research, or experimenting through modeling and simulation. For example, a multi-vehicle experimental platform for distributed coordination and control was developed at Massachusetts

[3] Refers to heterogeneous functionality and/or autonomy.

Institute of Technology (MIT) [102]. The project aimed to develop a platform that evaluates the coordination and control methodologies of managing multiple robots simultaneously, and was funded through an ongoing initiative by the United States Department of Defense (DoD). Another example is the project *Interacting with Multi-Agent Systems/UAV Swarms* conducted in Télécom Bretagne, France [52]. The project aims to contribute a solution that provides the capability for small groups of rotory wing UAVs, organised in swarms, to survey a strategic Air Force base.

With many institutions, organisations, and government bodies slowly becoming more aware and realising the potentials of deploying teams of multiple heterogeneous UAVs to perform missions, the industry is shifting its attention to better support the effective use of multiple heterogeneous UAVs.

1.1.2 Issues & Challenges

The aim of effectively employing multiple heterogeneous UAVs to be deployed in close proximity proposes the major challenge of operating ratio, and this challenge is preventing the wide implementation of UAV teams in the real world. Currently, at least two or more (in some cases, as many as four) human operators are required to ensure the safe operation of one UAV (traditionally known as the *many-to-one* relationship/ratio; where many operators are required to operate one UAV) [40, 51, 156, 34]. The issues associated with the current operating ratio in the context of a hybrid system[4] are twofold: effective operator collaboration [83, 27] and commercial viability.

The *many-to-one* operating ratio, which is currently accepted as a minimum operating requirement, proposes the challenge of effectively coordinating teams of UAV operators. The number of operators that are required for a multi-agent mission is a product of the number of operators required to operate each platform, by the number of platforms. In this scenario, the number of operators to UAV agents is highly disproportionate, creating a highly complex and challenging Command and Control (C2) task for the mission commander and his/her subordinate operator teams. Furthermore, although the UAV operators are remotely operating the machines, they may not be necessarily co-located, whereby remote operation may be achieved over long distances. This also proposes further challenges for the mission commander and other distributed operator teams to collaborate effectively.

Although the issue of commercial viability associated with the current UAV operating ratio is often less mentioned, it should not be disregarded. Not only is it financially and time consumingly cost intensive to provide a complete suite of training for any UAV operators in both the military and civilian domain, it is also expensive to develop hardware equipment for the Ground Control Stations (GCS) of the UAS. Therefore, the benefits and potentials of deploying multiple heterogeneous UAVs to perform complex mission can not

[4] A hybrid system refers to a system that includes both the human operator(s) and more than one machine (heterogeneous UAV) agents.

be maximised.

The advantages of overcoming these issues would mean less cost and resource intensive efforts are required to be effective in managing multiple heterogeneous UAVs. Moreover, the flexibility to manage these agents would be enhanced because mission commanders/-operators would not be required to moderate their assets through another layer of human interpretation (the current layer of UAV operators). This would eliminate the potential of the geographically distributed UAV operators inherited in the current operating ratio.

To enable the most effective deployment of multiple heterogeneous UAVs, the operating ratio must be inverted, forming a *one-to-many* (one operator to many UAVs) relationship [45, 51, 83, 113]. However, a number of challenges have been identified in the process of this inversion of operating ratio.The fundamental challenge of operating ratio inversion is the cognitive overhead the human operator experiences when a beyond-average volume of work is introduced. There are three cognitive attributes that are directly associated: Cognitive Workload (CW), Situation Awareness (SA), and Automation Trust [174].

CW refers to the physiological stress that a human operator experiences when any form of work is placed upon him/her. This attribute is one of the major human factor constructs that is limiting successful and reliable human machine interaction. CW is associated with an operator's SA [174, 154]. SA refers to the knowledge that a human operator has of his/her agents and the mission environment. An operator's SA dramatically decreases when he/she is removed from the situation scene, such as that of a traditional vehicle (eg. fighter aircraft, commercial airliner) to remotely controlling an RPAV or UAV, and then finally, mission supervising/commanding multiple heterogeneous UAVs.Automation Trust refers to the degree of trust that a human operator has of his/her autonomous agent. Due to the complexity of autonomy inherent in many modern day UASs, the system is a "black box", where the inner-workings of the UAV's autonomy capabilities are non-transparent to the human operator. An operator without an adequate level of knowledge about his/her UAV(s) will not be able to effectively manage a group of multiple heterogeneous agents during operation [172, 125].

Research to address the human cognitive challenge emergent in the transition from the *many-to-one* to *one-to-many* operating ratio has been conducted in two directions: from a human-decentralised and a human-centred approach. Firstly, the human-decentralised approach is aimed to modify the UAV's autonomy behaviour, such as proposing novel ways to classify the UAV autonomy capabilities [28, 48, 78, 105, 127, 173, 195, 139], or novel methods to enable autonomy to adapt to the mission situation as necessary [148, 117, 20]. Secondly, the human-centred approach is aimed at proposing a novel way to support interactions between the human operator and his/her UAVs [51, 128, 225, 89, 5, 20]; this includes physical interaction (i.e. haptic or vocal controls), management protocols (i.e. delegation of tasks/functions between the human and the machines [149, 151, 169]), and visualisation (i.e. Graphical User Interface (GUI) [224, 151, 38]).

Successful human-machine collaboration is only achieved when the human operator possesses a balanced CW, SA, and calibrated trust in automation [174, 126, 72, 165, 226,

172, 125, 37, 39, 44, 43]. Therefore, in order to support effective inversion of the operating ratio such that multiple heterogeneous UAVs can be effectively managed by a single human operator, both directions (human-centralised and decentralised) of research must be combined.

1.1.3 Capability Transparency

Techniques, methods, and algorithms to improve autonomy have been proposed to alleviate human operator's cognitive resource overhead to achieve effective multiple heterogeneous UAV management. However, this capability alone is insufficient to support effective human-machine interaction as these advancements in autonomy can be mishandled by the human operator [167]. Without the appropriate transparency[5], or information feedback about the UAVs' capabilities and intentions, the human operator can become confused, and over or under reliant on the autonomy due to a reduced level of SA. Consequently, this negatively impacts his/her CW, and the operator's trust in the system is jeopardised [126].

Transparency can also be interpreted as an operator's understanding of a machine's inner workings

The existing control interfaces that are designed to support Multiple-UAV operations with the traditional object oriented *Model-View-Controller* (MVC) GUI design paradigm, have not yet shown evidence of including the UAVs' capability information [157, 195]. Only within the past two years has research beyond this thesis emerged to evaluate the effects of capability information feedback on human cognitive performance [145], using an SA-centred framework proposed in 2014 [39]. This confirms the importance of capability transparency on human cognitive performance in multiple heterogeneous UAV management.

In the traditional *Presentation-Abstraction-Control* (*PAC*) model (illustrated in Figure 1.1), the *(P)resentation* component contains the visualisations of all the model data such as low level UAV state information and direct video feeds, the data contained in the model is not organised in any way to reflect directly the capabilities and its implications on the UAVs: the *(A)bstraction* component contains all the data about each UAV, the mission and global states, and their capabilities at a systems and subsystems level; and the *(C)ontrol* object contains the necessary exchange between the information stored in the *model* object, and displayed in the *view* object [54, 71].

[5]Transparency can also be interpreted as the ability to see through a system's internal workings which causes certain higher-level functions. However, this thesis focuses on the human operator's ability to understand a system's capabilities and intentions.

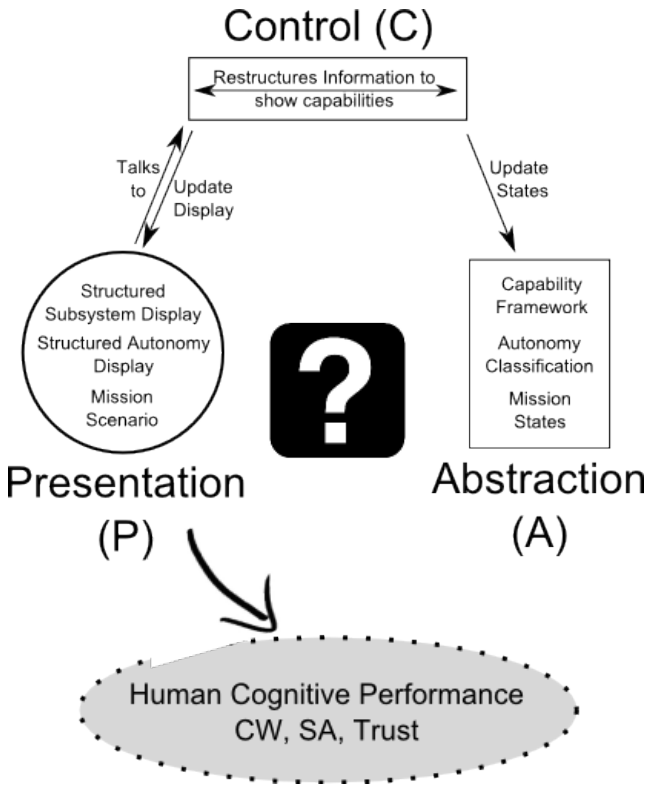


Figure 1.1: An example of the current GCS interfaces developed with the traditional *PAC* design model [219] (the image in the centre of this illustration is an example of the current GCS).

Capability Transparency is divided into two concepts: The *capability* of the UAV, and the *transparency* (availability) of this information to the human operator. The UAV capability is further divided into two types: *Functional Capability,* and *Autonomy Capability*. The *Functional Capability* encompasses the UAV's subsystem status and operational states. The *Autonomy Capability* encompasses the UAV's ability to perform tasks autonomously. This information is needed to be visually represented on the GUI used by the human operator to manage the agents and the mission. *Transparency* can be established in several ways, however, a key requirement is that the human operator must be able to easily comprehend the information. For this reason, information presented graphically or textually is considered.

In order to design an interface that supports capability transparency, a novel implementation of the existing *PAC model* is proposed. However, certain hurdles must be overcome prior to arriving at an interface solution; such as the representation of a UAV's functional capability transparency, and the autonomy capability transparency [44, 43, 42].

1.2 Research Program

The Motivation of the research presented in this thesis are threefold: The benefits and potential of employing multiple heterogeneous UAVs to perform missions (section 1.1.1), the

inherent cognitive issues associated with operator ratio inversion (1.1.2), and particularly the pending challenges to establish *Capability Transparency* in a hybrid system (section 1.1.3).

In search of a generic approach, efforts have been directed at extending the traditional *PAC* model of GUI design to incorporate elements associated with the functional (subsystem) and autonomy capabilities of a UAV in a hybrid system setting. This includes a **framework** to structure the UAV's subsystem functionalities and autonomy (*A*), a **visualisation** of this framework and the autonomy capabilities (*P*), an ability to link these two components together (*C*), and its impact on a human operator's cognitive performance as depicted in Figure 1.2.

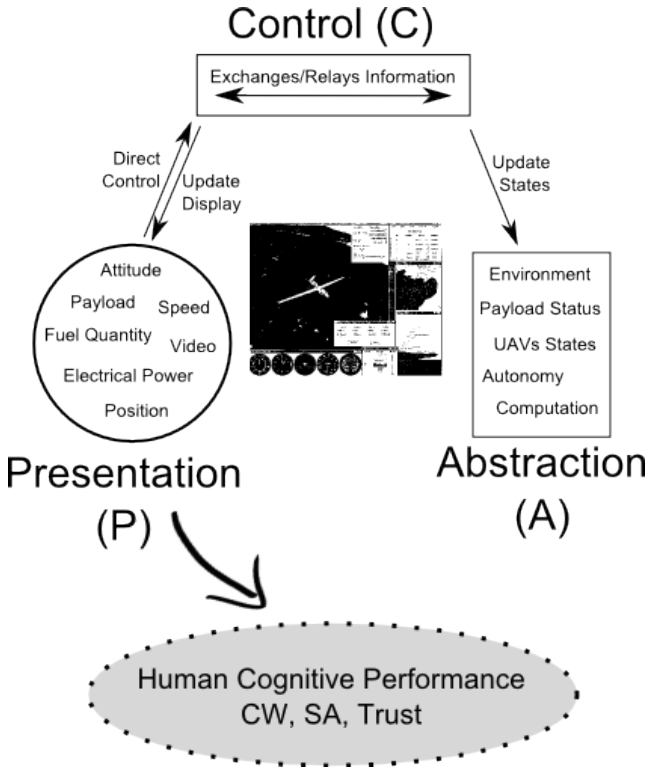


Figure 1.2: The proposed GCS interface developed with the instantiate the PAC GUI design model to support capability transparency.

1.2.1 Scope

Given the complexity of the chosen approach to improve multiple heterogeneous UAV management, implementing the existing *PAC* model to incorporate capability transparency, and the importance of evaluating its impact on human cognitive performance, a few scoping elements must be defined:

- Human cognitive functions is a broad research area which incorporates numerous psychophysical factors and constructs. It is not the aim of this research to encompass all aspects of human cognitive functions, but to include only CW, SA and Trust.

These factors are the most indicative of the human cognitive performance and have been accepted as the standard construct to evaluate the effectiveness of novel methods in human-machine interaction [44, 168, 175].

- Although the significance of Automation Trust is realised and acknowledged, heavy emphasis is not the intention of this research. This is because this research aims to demonstrate the value in visually representing functional and autonomy capability, and not explicitly the effects of the autonomy policy (i.e. adaptive autonomy). Hence, the experiments are not designed to be deceitful, which influences on human trust in the system.

The following constraints were applicable for this research:

- All software interface designs were developed purely for the purpose of this research, no pre-existing interfaces were available to be modified.
- All software interfaces were designed for experimental use only, and they were not certified for real world applications.
- All experiments were performed in a laboratory, no real world experimentations were conducted due to numerous existing regulatory and safety restrictions.
- All participants are volunteers with no cash incentives.
- No physiological instruments (e.g. eye trackers) were used for this research. Hence, all cognitive measurements were conducted based on self-reporting (Chapter 2 Section 2.3).

A major application for the use of multiple heterogeneous UAVs in the civilian sector is large scale search and rescue operations (e.g. the search for Malaysian Airlines MH370). To date there are limited existing real world missions in the civilian domain which deploy multiple UAVs to carry out the such tasks, several experiment scenario design considerations are defined:

- The context of multiple agents does not extend to swarm technology. Swarm technology considers the science of inter-agent collaboration, such as self-organisation capabilities [166] and pheromone behaviours [128]. This research aimed to study the agents as individual entities and teammates of the human operator. For this reason, swarm technology has very little overlap with the domain of this research.
- The scenarios designed for the experiments were aimed to assert certain levels of cognitive challenge for the participant. However, these scenarios are hypothetical only.
- All the components visually displayed in the scenarios had been explicitly placed, and the performances of the UAV agents had been implemented to work only within the predefined experimental arena.

1.2.2 Research Objective and Questions

The principal research objective of this thesis is:

Research Objective *To achieve functional and autonomy capability transparency, and to understand the effects of this transparency on human Cognitive Workload, Situation Awareness, Automation Trust and operator's objective performance in a multiple heterogeneous UAV management setting*

This objective is divided into three research questions which framed the agenda of this research. Each research question was addressed with a theoretical proposal, an experiment design, and an evaluation experiment involving human participants to validate the significance of the proposal.

Question 1 What are the effects of UAV functional transparency on human operator's CW and SA in heterogeneous multiple UAV management?

The first component of this research is aimed at addressing the needs for a UAV's **functional capability visualisation**. This is done by establishing a fundamental approach to structure these capabilities. It also provides support through an effective functional information exchange feature that enables the human operator to acquire information about the mission and agents. Firstly, to answer this question, an understanding of a generic UAV's subsystem elements and functions was established, this also includes the existing formats/methods of representation of such on a traditional GCS GUI. Secondly, human information processing and ecological interface design principles was investigated to establish an understanding of the human's ergonomic[6] preferences (Chapter 3 Section 3.2). Thirdly, these design principles was adopted and applied to structure the UAV's functional capabilities, proposing a novel multi-dimensional, multi-layered framework that encapsulates the UAV's subsystem functions, in an ergonomically sound way. Finally, the positive effects of using the proposed framework for multiple heterogeneous UAV management missions on the human operator's CW and SA was observed and validated in a live experiment involving no less than 23 human participants[7].

Question 2 What are the effects of UAV autonomy transparency on human operator's CW and SA when partial capability autonomy is represented visually in heterogeneous multiple UAV management?

The second component of this research is aimed to address the effects of a UAV's **partial autonomy capability transparency** on a human operator's CW and SA. It also provides

[6]The term *ergonomics* used in this context is to denote a human's natural process of information perception. It is not to be confused with the area of study.

[7]Minimum number of effective samples required for the results acquired from a pilot study to be of any statistical significance.

evidence to demonstrate the effectiveness of graphical representation of autonomy information on human operators. Firstly, to answer this question, a review of the existing methodology to classify autonomy was conducted to establish an understanding of its fundamental theory and applications. Secondly, a method was introduced to define the degree of autonomy applied to each of the UAV's subsystem functional capabilities, its meaning, implications, and instantiations. Thirdly, the graphical representation of a UAV's specific functional autonomy to increase autonomy transparency was considered and implemented in the experiment software prototype. Finally, the positive effects of increasing a single functional subsystem's autonomy transparency through visual representation on the human operator's CW and SA must be confirmed in a live experiment involving no less than 23 human participants.

> **Question 3** What are the effects of UAV autonomy transparency on human operator's CW, SA, Trust and operator's objective performance when the UAV's autonomy is represented visually in heterogeneous multiple UAV management?

The final component of this research is aimed to address the effects of a UAV's **complete autonomy capability transparency** on a human operator's CW, SA, and Automation Trust. The evidence collected in the process of answering this question aims to provide justification of the significance of information transparency on effective human-machine interaction. Firstly, to answer this question, investigation must be conducted on the most effective taxonomy to classify autonomy capabilities in a mission environment. A suitable taxonomy must incorporate elements of human information processing and levels of automation, and its successful application in recent research involving management of multiple agents must also have been demonstrated. Secondly, a review of existing modalities of visual information communication between a human operator and his/her computer simulated agents was conducted. The visualisation must have the capacity to communicate a variety of information about the UAV's autonomy to the human operator in a direct but non-intrusive manner. It must also have the capacity to enable the human operator to exercise acknowledgement of the communicated information for experiment verification purposes. Finally, the positive effects of increasing the UAVs' complete functional autonomy transparency through the visual representation on the human operator's CW, SA and Trust was confirmed in a live experiment involving no less than 23 human participants. Positive effects must also be observed on the participant's objective performances.

1.2.3 Contributions & Significance

The main contribution of this research is a framework of interface design that supports UAV capability transparency for hybrid systems. As a result, five significant theoretical, practical, and physical contributions toward the inversion of human-to-machine operating ratio to support effective management of multiple heterogeneous UAVs are presented.

- ***Contribution 1*** *A novel Functional Capability Framework (FCF) to encapsulate agent subsystem functional capability and its derivation procedure.* The FCF is proposed with the aim of not only encapsulating the subsystem components of a UAV for use in a hybrid system, but also structure it with from a human-centred perspective.
- ***Contribution 2*** *The concept of Functional-LOA to enable a system's autonomy to be classified in its separate capabilities.* F-LOA was proposed as a part of answering Research Question 2 and further used to answer Research Question 3. It aimed to describe a UAV's autonomy levels based on it's functional capabilities, which had not been seen in previous research.
- ***Contribution 3*** *A novel approach to support autonomy transparency.* The natural language communication of the autonomy capability information in a hybrid system is proposed as the method to allow autonomy status to be presented to the human operator, thus increasing the UAV's autonomy transparency to the operator.
- ***Contribution 4*** *An experimental software prototype that is customisable, and incorporates all elements proposed in contributions 1 to 3.* This is a physical set of software source code and compilable binaries that can be reconfigured and modified to satisfy future research needs.
- ***Contribution 5*** *A complete set of experimental scenario designs, procedures, and results for future research.* This includes three sets of experimental procedures and supporting materials that can be used to re-perform the experiments by a third-party, as well as its complete set of analysed results.

The significance of this research and the future impacts to the operator ratio inversion are evident through the consideration of all the contributions, these encompass: Theoretical contributions, practical guides and results, and physical systems. This research provided an approach to the investigation of effective human-machine interaction in hybrid systems.

From a theoretical perspective, Contributions 1, 2, and 3 provided a novel and empirically verified approach to the designs of future human-machine interfaces to support effective management of multiple heterogeneous agents. This opens up many future research opportunities to further investigate other avenues of human-machine interaction.

From a practical perspective, Contribution 4 provides a foundation platform which was established to encourage further refinement of that software prototype to enable a more robust experimentation platform. Although the scope of this research has limited the viability of the tools to only be tested within the confines of a laboratory, positive results that were observed and gathered during the experimental stages provide positive evidence that these contributions will become an integral part of future research and commercialisation.

Finally, from the physical contribution's perspective, Contribution 5 provides a complete set of documented experimental design and procedure with its supporting materials,

as well as the raw and analysed results, where statistical analysis had been conducted on the raw results collected during the experiments, provide a foundation to support future comparative studies.

The human operator's cognitive hurdles continue to be a considerable bottleneck of achieving the *one-to-many* (one human to many machines) operation ratio. In the absence of an effective model of capability transparency, the foreseeable potentials of deploying multiple UAVs in close proximity to perform a mission remain illusive. The contributions of this research demonstrate significance in achieving the *one-to-many* operation ratio, hence, another step closer to deploying multiple heterogeneous UAVs into the field.

1.3 Research Publications

Publications stemming from this research are listed in chronological order below:

T. (B.) Chen, D. Campbell, G. Coppin, M. Mooij, and F. Gonzalez, "A Capability Framework Visualisation for Multiple Heterogeneous UAVs: From a Mission Commander's Perspective" in *28th International Congress of the Aeronautical Sciences*, Brisbane, Australia, 2012.

T. (B.) Chen, D. Campbell, G. Coppin, and F. Gonzalez, "Management of Heterogeneous UAVs Through a Capability Framework of UAV's Functional Autonomy", in *15th Australian International Aerospace Congress*, Melbourne, Australia, 2013.

T. (B.) Chen, F. Gonzalez, D. Campbell, and G. Coppin, "Management of Multiple Heterogeneous UAVs using Capability and Autonomy Visualisation: Theory, Experiment and Result" in *The 2014 International Conference on Unmanned Aircraft Systems*, Orlando, Florida, 2014.

T. (B.) Chen, D. Campbell, F. Gonzalez, and G. Coppin, "The Effect of Autonomy Transparency in Human-Robot Interactions: A Preliminary Study on Operator Cognitive Workload and Situation Awareness in Multiple Heterogeneous UAV Management", in *The 2014 Australasian Conference on Robotics and Automation*, Melbourne, Australia, 2014.

T. (B.) Chen, D. Campbell, F. Gonzalez, and G. Coppin, "Increasing Autonomy Transparency through Capability Communication in Multiple Heterogeneous UAV Management", in *The 2015 International Conference on Intelligent Robots and Systems*, Hamburg, Germany, 2015.

T. (B.) Chen, D. Campbell, G. Coppin, and F. Gonzalez,"A Framework to Support Functional Capability Transparency in Multiple Heterogeneous UAV Management", in *IEEE Transactions of Human Machine Systems*, 2016 [submitted].

1.4 Thesis Structure

Chapter 1 presents an introduction to this research from the definitions of the terms to be used throughout this thesis, to various real world applications, benefits and challenges.

A research program outlining the specific research questions, scope, contributions and significance has also been presented.

Chapter 2 presents an in depth review of the literature that supports the formulation of the proposed contributions.

Chapter 4 presents a brief background overview of the research, while describing the process of formulating the proposed contributions.

Chapters 5, 6, and **7** describe the design of the validation Experiment 1, 2, and 3, their components visualisation, their interaction strategy, their experimental methodology, and their result analysis.

Chapter 8 provides a summary of all the contents discussed in the previous chapters, its relationships with the research contributions, and concludes the thesis with recommendations of directions for future research.

CHAPTER 2

Literature Review

This chapter presents the literature review for this research, organised in three components: Systems and Automation (Section 2.1), Interfaces and Interaction (Section 2.2), and Cognitive Constructs (Section 2.3).

Section 2.1, *Systems and Automation*, focuses on the building blocks which defined the descriptive framework of autonomous agents; in the context of this research, multiple Unmanned Aerial Vehicles (UAVs). This section includes an in depth review and evaluation of many forms of autonomy classification taxonomy. These include: The Ten Levels of Automation, or the Sheridan and Verplanck (SV) Scale (Section 2.1.1), the four-stage model of human automation interaction (Section 2.1.2), the autonomy spectrum (Section 2.1.3), the Autonomy Control Level (ACL) (Section 2.1.5), Endsley and Kaber's Levels of Automation (LOA) (Section 2.1.6), the (ALFUS) (Section 2.1.7), 3D Intelligent Space (Section 2.1.4), and Section 2.1 concludes with the Human-Automation Collaboration Taxonomy (HACT) (Section 2.1.8).

Section 2.2, *Interfaces and Interactions*, highlights the existing concept of human-computer interface design principles (the Model-View-Controller (*MVC*) paradigm and the Presentation-Abstraction-Control (*PAC*) model in Section 2.2.1), ergonomic interface design methods (the Ecological Interface Design principles in Section 2.2.2), modes of interaction (dialogues, Belief-Desire-Intention model, authority sharing, and adaptive automation in Section 2.2.3 to 2.2.6), and concludes with the review of current progress, challenges and research trends in the area of autonomy, systems and agent transparency (Section 2.2.7 and 2.2.8), as well as its significance to the human operator's cognitive performance.

Section 2.3, *Cognitive Constructs*, reviews and evaluates three cognitive constructs that had been well established as most significant in the field of human-automation interaction: Cognitive Workload (CW, in Section 2.3.1), Situation Awareness (SA, in Section 2.3.2), and Automation Trust (in Section 2.3.3). This section presents the foundation theory of each of these constructs, their significance to this research, and an evaluation of the main methods to measure them. The conclusion addresses the relevance of each topic presented in this chapter and the methodology adopted in this research.

Table 2.1: Ten Levels of Autonomy (Adapted from [211])

LOA	Description
1	Human does it all
2	Computer offers alternatives
3	Computer narrows alternatives down to a few
4	Computer suggests a recommended alternative
5	Computer executes alternative if human approves
6	Computer executes alternative; human can veto
7	Computer executes alternative & informs human
8	Computer executes selected alternative & informs human only if asked
9	Computer executes selected alternative & informs humanonly if it decides to
10	Computer acts entirely autonomously

2.1 Systems and Automation

Systems, or machine agents in the context of this research, include both a physical platform and an autonomous agent, and the need to effectively classify the machine agent's autonomy capability had been a key motivation in many efforts to develop taxonomies [211, 28, 48, 173, 51, 104, 78].

2.1.1 Ten Levels of Automation (SV Scale)

One of the most widely known LOA taxonomy is the Ten Levels of Automation (also known as the *SV Scale*), originally proposed by Sheridan and Verplanck [211]. This taxonomy contains ten levels; Level 1 represents the system has no portion of autonomous control and it is entirely manually operated, while Level 10 represents the system has no portion of manual control and it is entirely automated. Figure 2.1 illustrates the details of each of the levels in this ten-level taxonomy.

This taxonomy is set on a linear scale and the attributes used to classify the levels are limited. The human machine interaction process is classified from a linear scale of 1 to 10 (2.1) with only descriptions relating to each level available. A lack of the inclusion of any information processing flow attributes was stated [28, 173, 53, 48].

For this reason, Parasuraman, Sheridan and Wickens have revised the taxonomy to incorporate a simple four-stage human information processing model [173] (Section 2.1.2).

2.1.2 Model for Human-Automation Interaction

Parasuraman, Sheridan and Wickens agreed and acknowledged the need to classify autonomy beyond a linear scale, rather, automation can vary across a continuum of levels. However, the purpose of automation is to alleviate, assist, or sometimes supplement human effort to conduct work particularly in situations where repetitive and fast-paced tasks must be done [168, 132]. Hence, Parasuraman, Sheridan and Wickens acknowledged the need to incorporate a simple four-stage model of information processing into the ten

levels of automation [173].

The four stages of human information processing include: *Sensory Processing*, *Perception/Working Memory*, *Decision Making*, and *Response Selection* [173]. These stages were simplified as Information Acquisition, Information Analysis, Decision Selection and Action Implementation (Figure 2.1). The concepts associated with this process equate closely to the renowned Observe-Orient-Decide-Act (OODA) model of human decision making, proposed by Boyd [19].

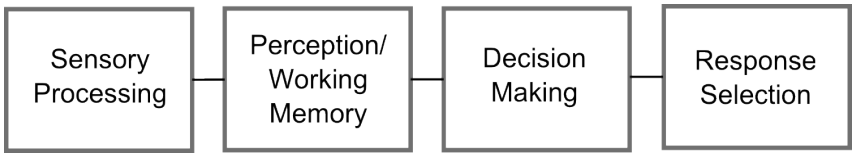


Figure 2.1: Simplified four-stage model of human information processing (Adapted from [173])

The ten levels of automation are then applied to each stage of information processing which defines the amount of responsibility of work of the human operator and the automation.

The first stage, information acquisition, or the *Observe* stage in the OODA model, considers the process of acquiring and gaining information of various forms from various sources. This stage may be performed at any level of the ten LOAs. For example, at the lowest level (LOA 1), the human operator is responsible for acquiring all the possible information regarding the machine agents and the environment from the available sensory nodes. Depending on the complexity of the system, the operator's CW may be saturated, hence, negatively impacting the SA as well [74].

The second stage, information analysis, or *Orient*, considers the process of invoking working memory to deciphering and decoding the information acquired in the first stage. This process then constructs a "mental picture" of the situation. For example, in the aviation domain, information acquired on one aircraft's sensory systems (such as radars) about its neighbouring aircraft's position and state information, can be aggregated, and graphically presented to the pilot with a prediction of the neighbouring aircraft's trajectory [153, 98]. The LOA of this stage is higher up in the scale for this example as all the information aggregation was carried out by automation with no human intervention.

The third stage, decision selection, or *Decide*, considers the act of generating decisions based on the information acquired in the information stages (stages one and two), and selecting the most appropriate action as a response. Both steps in this stage can be done by humans or automation, and an action selection may not necessarily require a physical input to be initiated. The LOA which applies to this stage governs the human/automation interaction, that is, at a high LOA, the automation makes the decisions based on the situation, at low LOA, the human operator assumes responsibility. An example is the current Ground Proximity Warning System (GPWS) installed on most of the modern commercial aircraft, which has an LOA of 4. This system processes the data related to the distance the

aircraft is from the terrain/ground, and provides advisories (warnings) when the aircraft is close to the ground but not properly configured to land. The human operator (pilot) has the decision to ignore the warning, or accept it and perform the next stage of information processing – *action implementation*.

The fourth and final stage, action implementation, or *Act*, refers to the execution of the action selected in the previous stage. Similar to the previous stages, the agent (human or autonomous) responsible to carry out the decision is governed by the LOA. An example of which is in the commercial aviation sector where the flight plans specific to a particular flight are pre-planned in the briefing/preparation room, and the flight plan is automatically uploaded to the aircraft when the flight details are finalised. This system incorporates a system that has a high LOA where flight details are not required to be manually inserted by the human operator (the pilot) [161, 135]. Figure 2.2 illustrates an example provided by Parasuraman, Sheridan and Wickens of the proposed future Air Traffic Control (ATC) systems with their model of types and levels of automation.

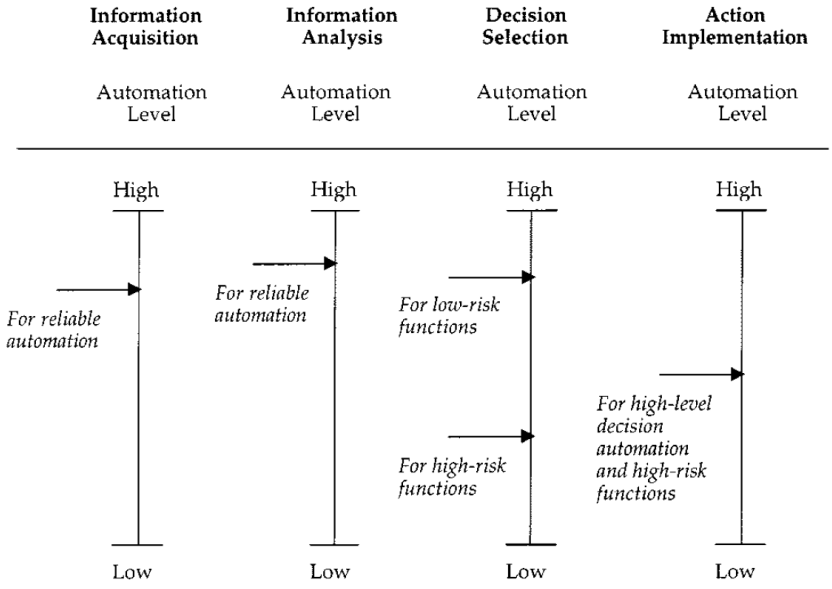


Figure 2.2: Example of Levels of Automation applied to the four stages of information processing (Published in [173])

The information automation (the first two stages) is high, as other studies had shown positive affect of high levels of information automation can have on a human operator's cognitive workload. The decision selection proposes a high LOA for low-risk functions such as the automatic sorting of electronic flight strips, and a low LOA for high-risk functions such as the initiation of the handoff protocol. Action implementation is set at a moderate LOA such as changing of the flight plan details, which the human operator and autonomy have similar levels of responsibility to execute a successful flight plan detail change.

Although Parasuraman, Sheridan and Wickens acknowledge and agree that this four-step model is a simplification of human information processing compared to ones dis-

cussed cognitive psychology [26], it encapsulated the core concepts of both automation and humans, enabling a true human-automation taxonomy to classify interaction.

Legras, Coppin and the team have further adopted this approach and proposed the concept of autonomy spectrum discussed in Section 2.1.3.

2.1.3 Autonomy Spectrum

The Autonomy spectrum, proposed by Legras and Coppin [51], is a graphical representation of the different configurations of operating modes. It has adopted two concepts in their proposal of the Autonomy Spectrum, which was employed in SMAART Project, and is now adopted as the taxonomy to classify autonomy.

The concepts included the implementations of multiple configuration patterns for each mode of operation [43, 127], and Sheridan and Verplanck's ten levels of automation (SV Scale) [211] applied to Parasuraman, Sheridan and Wickens' four stage model of human automation interaction [173]. Coppin and Legras have agreed with the significance of human-machine operating ratio inversion (from multiple human operators to control a single machine, to a single human operator managing multiple machines) [83], a human-centred approach to classify human-autonomy interaction was required. And by extending Parasuraman, Sheridan and Wickens' model of human automation interaction (which was effective to describe a static automation system [173, 51]), autonomy transparency has the capacity to incorporate adaptive systems deployed in a highly dynamic environment.

The autonomy spectrum was dedicated to highlight the UAV autonomy configuration patterns, which describes the autonomy of each mode-of-operation in two dimensions by projecting the SV Scale onto the four-stages of the simplified human decision making process. The autonomy of each mode-of-operation in the stages is divided into the ten LOAs, while the mode-of-operation is described as a specific task/aim a UAV agent (human and/or machine) must complete.

An example of the autonomy spectrum is illustrated in Figure 2.3. This figure presents a reference spectrum for a mode-of-operation on the left, and its four corresponding single-mode spectra on the right. The purpose for the reference spectrum on the left is to enable a global "snapshot" of the adaptive system's autonomy capabilities. The four spectra on the right denote the possible autonomy configurations of the same mode-of-operation [51].

The example spectrum presented in Figure 2.3 is interpreted to possess an LOA 8 capability for information acquisition, LOA 3 or 9 for information analysis, LOA 1 or 6 for decision selection, and LOA 8 for action implementation [51].

- LOA 8 for information acquisition denotes that the machine agent has been allocated greater responsibility in the task of collecting sensory information, however, the operator is able to gain access to the raw data if he/she so chooses;
- LOA 3 or 9 for information analysis denotes that the information collected previously can either be interpreted autonomously by the machine, or deciphered rather manually by the human operator;

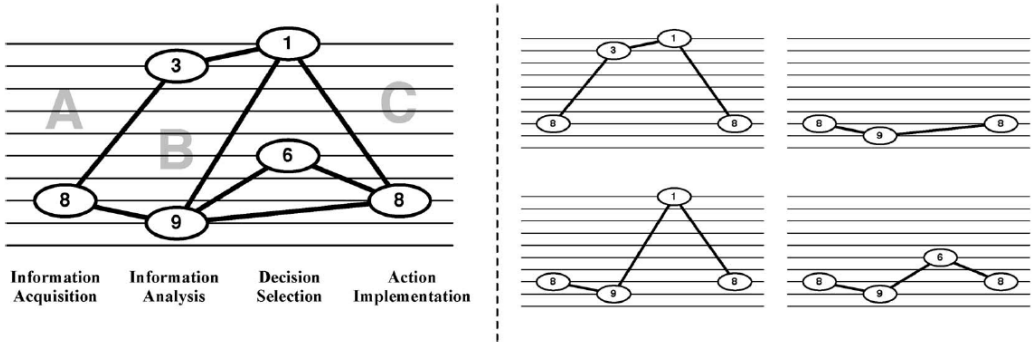


Figure 2.3: An example autonomy spectrum for a mode-of-operation (Published in [51])

- LOA 1 or 6 for decision selection denotes that in the responsibility of making the final decision can either be done completely manually, or semi-autonomously, that is, with some decision-aiding capabilities for example;
- LOA 8 for action implementation denotes that the machine agent has been allocated greater responsibility to carry out the execution of the selected decision from the previous stage;

Also note that the Figure 2.3 contains path lines, which defines the autonomy configuration of the mode-of-operation. The four right figures illustrate all the possible autonomy configurations for this mode; from a high LOA system (8-9-6-9), to a mixed LOA system (8-3-1-8). However, one example illustrated the configuration of 8-9—8, this denotes that no decision making is required under this specific autonomy configuration [51]. Each scenario include multiple adaptive systems, which propose multiple modes-of-operations, and autonomy spectrum is designed to enable system designers or analysts to determine the autonomy capabilities of their systems (or in the context of this research, UAVs) in an operational manner [43].

2.1.4 3D Intelligent Space

Charles Stark Draper Laboratory developed a set of metrics that measure the performance of their systems carrying out various tasks. 3D Intelligence Space is one of these options for measuring the intelligence and autonomy of a system [48, 47]. The proposed taxonomy expands into multiple dimensions as illustrated in Table 2.2. There are only four levels separated into three different spaces in this metric;

- Mobility Control refers the ability to move in the given environment
- Task Planning refers to the ability to reactively plan unforeseen tasks according to the surrounding environment
- Situation awareness refers to the level of sensory feedback that is ported to the system

Table 2.2: 3D intelligence space (Adapted from [48])

Levels	Mobility Control	Task Planning	Situation Awareness
1	None, RPA Only	None, RPA Only	None, RPA Only, or sensor as conduit
2	Operator Assisted	Waypoint or feature oriented	Low-level sensor processing, e.g. visual servoing (template tracking)
3	Get to waypoint, do one feature-based command	Interpret goals into action	Single-Sensor model matching
4	Integrate multiple actions	Multi-Agent Collaboration and C2	Integrated, multi-sensor fusion

The limitations of this metric are;

- The "Task Planning" dimension may need to be re-cast, as many successful autonomous systems are built on reactive behaviours, and task planning is not a compulsory element of autonomy [48].
- SA can apply to one, or both of the entities; the human operator, who can be more aware of the surrounding environment to perform more efficiently in decision-making and also the machine agent, which can generate and react more accurately according to the level of SA it experiences. It does not imply whether the system understands the details of the actual surrounding factors [48].

Given these limitations, efforts to develop other autonomy classification (i.e. the development of ACL) were initiated by the United States (U.S.) Air Force Research Laboratory's Air Vehicles Directorate [48]. Other efforts to classify autonomy from the a multidimensional perspective include: Clough's Autonomous Control Levels (ACL) [48], and Endsley and Kaber's eight levels of autonomy, from the perspective of human cognition [78].

2.1.5 Autonomous Control Levels (ACL)

The fundamental questions of how autonomous a system is and how this autonomy can be classified have also been explored by a number of researchers at the U.S. Air Force Research Laboratory and their publications on ACL [48].

Capturing the idea from both, but not limited to the SV scale [211] and the 3D intelligence space [47], ACL has been developed. The resolution of the scale has increased to eleven levels as illustrated in Table 2.3, while retaining the multi-dimensional facet presented in the 3D intelligence space [48].

This taxonomy also incorporates the four-stage human information processing model [48] as reviewed in Section 2.1.2.

- Perception/Situational Awareness: The robustness of acquiring live information from the surroundings;

Table 2.3: Final ACL Chart as published in (Published in [48])

Level	*Level Descriptor*	*Observe*	*Orient*	*Decide*	*Act*
		Perception/Situational Awareness	Analysis/Coordination	Decision Making	Capability
10	*Fully Autonomous*	Cognizant of all within Battlespace	Coordinates as necessary	Capable of total indepenance	Requires little guidance to do job
		Battlespace inference - Intent of self and others	Strategic group goals assigned	Distributed tactical group planning	Group accomplishment of strategic goal with
9	*Battlespace*	(allies and foes).		Individual determination of tactical goal	no supervisory assistance
	Swarm	Complex/Intense environment - on-board tracking	Enemy strategy inferred	Individual task planning/execution	
	Cognizance			Choose tactical targets	
8		Proximity inference - Intent of self and others	Strategic group goals assigned	Coordinated tactical group planning	Group accomplishment of strategic goal with
	Battlespace	(allies and foes)		Individual task planning/execution	minimal supervisory assistance
	Cognizance	Reduced dependance upon off-board data	Enemy tactics inferred	Choose targets of opportunity	(example: go SCUD hunting)
			ATR		
7		Short track awareness - History and predictive battlesp	Tactical group goals assigned	Individual task planning/execution to meet goals	Group accomplishment of tactical goal with
	Battlespace	data in limited range, timeframe, and numbers	Enemy trajectory estimated		minimal supervisory assistance
	Knowledge	Limited inference supplemented by off-board data			
6	*Real Time*		Tactical group goals assigned	Coordinated trajectory planning and execution to meet	Group accomplishment of tactical goal with
	Multi-Vehicle	Ranged awareness - on-board sensing for long range,	Enemy location sensed/estimated	goals - group optimization	minimal supervisory assistance
	Cooperation	supplemented by off-board data			
					Possible close air space separation (1-100 yds)
5	*Real Time*	Sensed awareness - Local sensors to detect others,	Tactical group plan assigned	On-board trajectory replanning - optimizes for	Group accomplishment of tactical plan as externally
	Multi-Vehicle	Fused with off-board data	RT Health Diagnosis; Ability to compensate for most	current and predictive conditions	assigned
	Coordination		failures and flight conditions; Ability to predict onset of	Collision avoidance	Air collision avoidance
			failures (e.g. Prognostic Health Mgmt)		Possible close air space separation (1-100 yds) for
			Group diagnosis and resource management		AAR, formation in non-threat conditions
4		Deliberate awareness - allies communicate data	Tactical plan assigned	On-board trajectory replanning - event driven	Self accomplishment of tactical plan as externally
	Fault/Event		Assigned Rules of Engagement	Self resource management	assigned
	Adaptive		RT Health Diagnosis; Ability to compensate for most	Deconfliction	
	Vehicle		failures and flight conditions - inner loop changes		
			reflected in outer loop performance		Medium vehicle airspace separation (100's of yds)
3	*Robust Response*	Health/status history & models	Tactical plan assigned	Evaluate status vs required mission capabilities	Self accomplishment of tactical plan as externally
	to Real Time		RT Health Diag (What is the extent of the problems?)	Abort/RTB if insufficient	assigned
	Faults/Events		Ability to compensate for most control failures and		
			flight conditions (i.e. adaptive inner-loop control)		
2		Health/status sensors	RT Health diagnosis (Do I have problems?)	Execute preprogrammed or uploaded plans	Self accomplishment of tactical plan as externally
	Changeable		Off-board replan (as required)	in response to mission and health conditions	assigned
	Mission				
1	*Execute*	Preloaded mission data			
	Preplanned	Flight Control and Navigation Sensing	Pre/Post Flight BIT	Preprogrammed mission and abort plans	Wide airspace separation requirements (miles)
	Mission		Report status		
0	*Remotely*	Flight Control (attitude, rates) sensing	Telemetered data	N/A	Control by remote pilot
	Piloted	Nose camera	Remote pilot commands		
	Vehicle				

- Analysis/Coordination: The ability to adapt and coordinate with the remaining of the team from the acquired live information and health of the system;
- Decision Making: The ability to make appropriate decisions based on the available data;
- Capability: The ability to carry out tasks autonomously as required by the scenario based on the decision (autonomously or manually);

Although this taxonomy is very extensive and has considered information processing at each level of automation, the ability to classify a heterogeneous multi-agent autonomous system was not clearly captured.

2.1.6 Endsley and Kaber's Level of Automation

Traditionally, the design decision of automation is made to maximise the capability of the technology, this results in a reduction in cost and human errors [78]. The aim of automation is to minimise human involvement. Unfortunately, the cognitive capability of a human operator to perform monitoring or pure supervisory tasks is very limited as the human operator's sense for SA decreases with time [78]. To this end, a system that is fully autonomous (with human operators being completely excluded from the process) is not the optimal solution to maximise task efficiency. Instead, a balance of human and machine involvement has a more positive effect. To address this from the view of the performance, SA and CW in a dynamic control task scenario, Endsley and Kaber had proposed a method of classifying levels of automation [78].

The *Endsley and Kaber* scale consists of ten autonomy levels and four roles. The roles are presented as processes starting from information collection to performing selected decisions. Table 2.4 illustrates a hierarchy of levels of automation presented by Endsley and Kaber.

Each level of automation states the responsibility of the roles (either human, machine or both):

1. Manual Control: In this level, human agent assumes responsibility for all roles;.
2. Action Support: In this level, the majority of the roles are the human agent's responsibility, with the computer agent aiding in the information observation, and implementation processes.
3. Batch Processing: It is the human agent's responsibility to generate and make the appropriate decisions, and then it is up to the machine agent to carry out the selected action.
4. Shared Control: The process of generating options is carried out by the machine and human agents, although the human agent still retains full control over the selection and implementation of the selected action.

Table 2.4: Hierarchy of LOA Applicable to Dynamic-Cognitive and Psychomotor Control Task Performance, where H denotes Human, and C denotes Computer (Adapted from [78])

		Roles			
Level	**Descriptor**	**Mon.**	**Gen.**	**Sel.**	**Impl.**
1	Manual Control (MC)	H	H	H	H
2	Action Support (AS)	H/C	H	H	H/C
3	Batch Processing (BP)	H/C	H	H	C
4	Shared control (SHC)	H/C	H/C	H	H/C
5	Decision Support (DS)	H/C	H/C	H	C
6	Blended Decision Making (BDM)	H/C	H/C	H/C	C
7	Rigid System (RS)	H/C	C	H	C
8	Automated Decision Making (ADM)	H/C	H/C	C	C
9	Supervisory Control (SC)	H/C	C	C	C
10	Full Automation (FA)	C	C	C	C

5. Decision Support: The machine agent will automatically perform solution generation. The human agent may veto and perform his/her own decision-making process. The selection of the solution is the responsibility of the human agent, and then automation will initiate the implementation of the selected decision.

6. Blended Decision Making: Automation is employed through most of the solution generation and selection process with the human agent's consent to carry forth the action. The machine agent also performs the implementation;

7. Rigid System: This level implies that human agent cannot manually generate new solutions, he/she must select a solution from the list of solutions generated by the machine agent automatically.

8. Automated Decision Making: At this stage, the system will autonomously generate and selects the most ideal solution for the particular situation, human agents may suggest alternative solutions, and machine agents will incorporate the new suggestions and carryout the decision making processes.

9. Supervisory Control: The role of the human agent is to supervise a system from a very high level perspective. The machine agent carries out all tasks from solution generation, to selection implementation. The human agent may choose to veto when it is necessary.

10. Full Automation: Machine agent has the responsibility to maintain a system at a fully autonomous functional level. That is, this level implies human out-of-loop scenario.

Although this metric presents a method to classify the levels of automation in terms of a workflow model (from solution generation to implementation), there are possible is-

sues associated. For example, one possible issue is that this taxonomy seems not to be able to support multiple autonomous agent scenarios, only one human operator and one autonomous agent was addressed in this literature.

Beyond the classification of human and automation interaction, Huang *et al.* had proposed a "tool" designed for military end users [51].

2.1.7 Autonomous Levels For Unmanned Systems

The Autonomous Levels For Unmanned Systems (ALFUS), initially proposed by the National Institute of Standards and Technology (NIST), addresses the issues of a linear automation classification, and lack of information processing flow attributes, which was visible from Sheridan and Verplanck's Ten Levels of Automation.

In this taxonomy, an Unmanned System (UMS) is defined as *a powered physical system, with no human operator aboard the principal components, acts on physical world for the purpose of achieving assigned tasks. May be mobile or stationary. May include any and all associated supporting components. Examples include unmanned group vehicles (UGV), unmanned aerial vehicles (UAV), unmanned underwater vehicles (UUV), unmanned water surface borne vehicles (USV), unattended munitions (UM), and unattended ground sensors (UGS). Missiles, rockets, and their submunitions, and artillery are not considered UMSs [103].*

ALFUS describes autonomy in three tiers: Subsystems, systems, and system of systems. Each tier is then broken into five LOAs as illustrated in Table 2.5. These levels are defined according to three aspects of Contextual Autonomous Capability (CAC), Mission Complexity (MC), Environmental Complexity (EC), and Human Independence (HI).

ALFUS uses the CAC model (presented in Figure 2.4) to define UMSs. This model consists of three aspects or axes: MC, EC, and HI. MC is concerned about the various details of the mission task which contributes to a more difficult mission. Examples of these tasks are mission time constraints, resource availability, asset availability, intelligence gathering, task planning, and mission planning etc. EC is concerned about the environment which the mission is surrounded by. Environmental variables can have great impact on the success of the mission. These variables can range from natural causes, atmospheric disturbances, meteorological conditions, geological limitations, and man-made obstacles etc. HI is concerned about the UMS's independence from human inputs. This is also known as LOA.

These aspects can be rated on a scale from 0 to 10:

- At 0, the simplest task in relation to MC, static and simple environmental impacts in relation to EC, and least independence (remote control) of human involvement in relation to HI,
- At 10, the highest level of capability of the UMS, where it is able to adapt to the most difficult mission in relation to MC, the most difficult and dynamically changing environment perceived by the UMS in terms of EC, and UMSs with maximum autonomy, or with complete human independence in terms of HI.

Table 2.5: Levels of Automation within each Tier of ALFUS (Adapted from [104])

Levels	Explanation	Example
Simple goal attainment [104]	An agent being only able to follow set rules (rule based system), and cannot respond according to the surrounding environment.	A ground robot moves from point A to point B with simple instructions (e.g. speed, direction etc.)
Goals can be achieved in static environment in a single UMSs scenario	According to static surrounding, an individual autonomous agent can respond appropriately to achieve a goal.	A ground robot travels from point A to point B, following a set route and avoiding any object along the way.
Goals can be achieved in a static environment in a group of UMSs	A group of UMS can respond according to the static surrounding to achieve a goal.	A group of ground robots travel from point A to point B, following a set route and avoiding any obstacles along the way.
Single UMS being able to achieve goals in dynamic environment	According to the ever-changing surroundings, with a wide variety of information data, an individual autonomous agent can respond appropriately to achieve a goal.	A UAV travels from point A to point B, avoiding building storms, windshear, other surrounding traffic etc.
Group of UMS being able to achieve goals in dynamic environment	A group of UMS can respond accordingly to the dynamic surrounding, with a wide variety of information data, to achieve a goal.	A squadron of UAVs is to travel from A to B, and is to confront live enemy units.

This taxonomy addresses the automation levels in greater detail and this taxonomy concentrates on the entire problem domain as illustrated in Figure 2.5 (environment, mission, and human) [105]. However, it was designed to classify a single or groups of UMSs, rather than heterogeneous autonomous systems, and it was not designed for system designers or academics. Hence, the applicability of this taxonomy is very limited.

From the perspective of capturing the collaborative relationship between human operators autonomous agents, a more complex taxonomy, HACT, was proposed by Bruni *et al.*. HACT was proposed to focus on the information analysis and decision selection stages of the simplified human information processing model was the result [51, 28].

2.1.8 Human-Automation Collaboration Taxonomy

In 2005, Bruni *et al.*, have encapsulated the collaborative nature in human automation systems through the proposal of the HACT [56, 55, 28].

The attributes proposed in this taxonomy are completely independent from the existing LOA taxonomies with some similarities shared between ACL [48] and the Model of

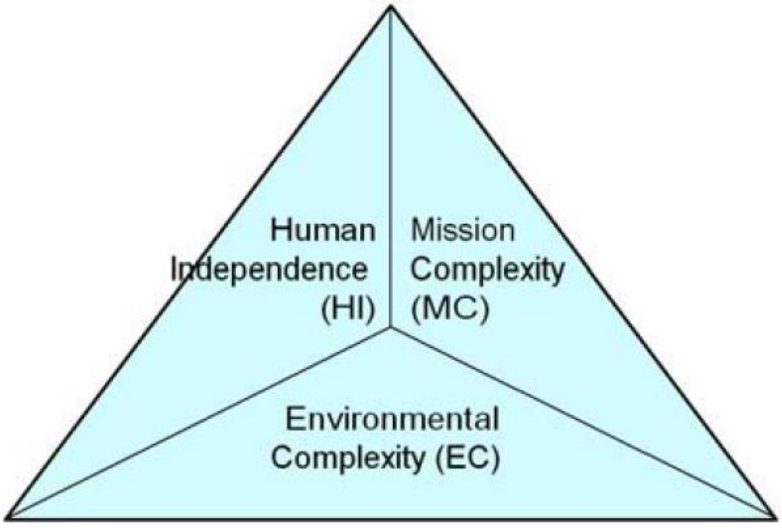


Figure 2.4: Contextual Autonomou Capability (Published in [104])

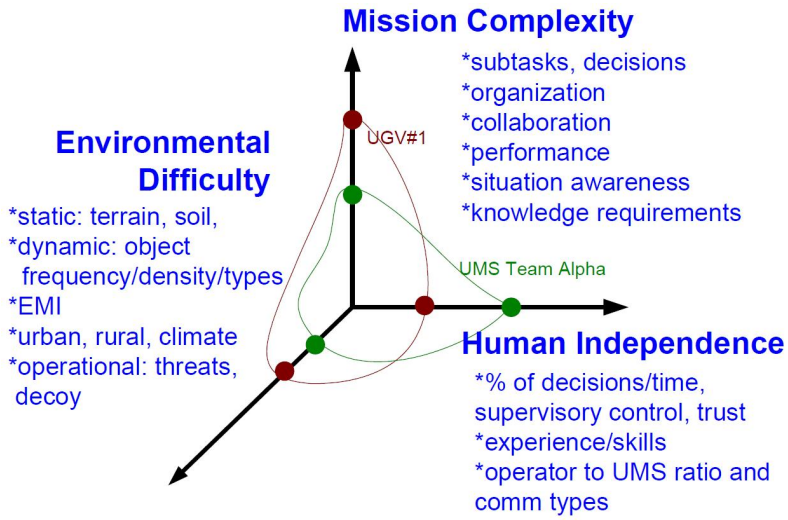


Figure 2.5: A detailed model of ALFUS (Published in [105])

Human-Automation Interaction [173]. A new framework and approach was employed and described in details in the following paragraphs.

HACT has employed Parasuraman, Sheridan and Wickens' information processing flow model [173], and has focused on the Decision-Making Process (DMP) (Figure 2.6). This information process flow model also has similarities to Dynamic Decision making model [48] and OODA model [24] discussed in Section 2.1.2.

Figure 2.6 illustrates the three fundamental roles of HACT: Moderator, Generator, and Decider. The role of the *moderator* is to ensure the entire information processing flow carries forward at necessary rates and standards. The role of *generator* is to analyse data collected during the data acquisition stage and presents possible solutions (if necessary) in response to the evaluated sensory data. The role of *decider* evaluates the proposed solutions and make a selection on the most suitable solution for the following (action) stage

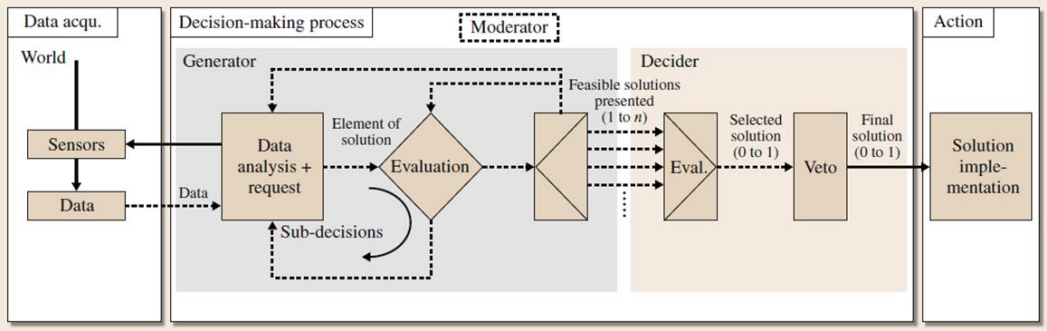


Figure 2.6: The three collaborative decision-making process roles (Published in [56])

for execution. These roles can also be classified into five levels with each level defines the responsible agent (Table 2.6, and the roles of the generator and the decider have five levels (Table 2.6 and 2.7).

Table 2.6: Moderator/Generator scale of automation (Adapted from [28, 55, 56])

Moderator/Generator Level	Who assumes the role?
1	Human
2	Mixed, but more human
3	Hybrid
4	Mixed, but more automation
5	Automation

Table 2.7: Decider scale of automation (Adapted from [55, 56, 28])

Decider Level	Who assumes the role?
1	Human
2	Mixed, but more human
3	Hybrid
4	Mixed, but more automation
5	Automation

In addition to the collaborative decision making, the roles and the levels, three primary characteristics are also included: *Functional Transparency, Information Transparency,* and *Interactivity*. These characteristics may assist in improving the SA of the human operator which have been studied in subsequent publications related to this research [42, 40, 41, 44, 43], as the levels that govern these characteristics indicate the opacity of the automation process, information feedback and the interactivity between the machine and the human agent. A secondary characteristic defined in this taxonomy is the adaptability of each of the basic roles and their primary characteristics completes this taxonomy. The adaptability is defined as the ability to manipulate the levels of a system without the human intervention [28].

Table 2.8: Basic roles versus primary characteristics where Func. denotes Functional, Info. denotes Information, and Trans. denotes Transparency (Adapted from [56])

	Func. Trans.					**Info. Trans.**					**Interactivity**				
Moderator	x	2	3	4	5	x	x	x	x	x	x	2	3	4	x
Generator	x	2	3	4	5	1	2	3	4	5	x	2	3	4	x
Decider	x	2	3	4	5	1	2	3	4	5	x	x	x	x	x

The *Functional Transparency* is classified in three levels [28]: Black, Grey, and White. Black denotes that the automation process has been completely hidden from the human agent; therefore the human operator does not know the actual working process of the machine agent. Grey denotes that partial automation internal process is presented to the human agent. These may include limited abstract information, or the use of metaphors (processed data). White denotes that the complete representation of the internal automation process is presented to the human agent, but this may cause information overwhelming for the human operator. The impact of this concept on human cognitive performance was also evaluated as a part of this research project [42, 43].

The *Information transparency* is also presented in three levels [28]: Raw, Mixed, and Aggregated. Raw denotes the instantaneous sensory information that is being presented to the human agent without any processing or manipulating. Mixed denotes that some information is processed, some remains raw. Aggregated denotes that all sensory measurements are processed into meaningful information, and are presented to the human operator. This concept had influenced the Functional Capability Framework proposed in this research [41, 40, 44].

The *Interactive Scale* is presented in two levels [28]: Command, and Dialogue. Command denotes that the controlling agents issue instructions or decisions to other agents to carry out processes. The other agents may provide confirmation or feedback regarding the commanded outcome. Dialogue denotes that the collaboration between agents is established to achieve a common goal.

Table 2.8 summarises the relationship between the basic roles and primary characteristics; the number indicate the level within each roles, grey text implies that the characteristic does not apply to that particular level. The components that constructed the systems and machine must be managed effectively by the human operator through interfaces and interaction methods that are at the forefront of human-machine interaction. Hence, the following section presents the foundation literature related to Interfaces and Interactions.

2.2 Interfaces and Interactions

Many current UAV Ground Control Stations (GCS) User Interfaces (UIs) are implemented [64, 108, 187, 176, 10, 115, 233, 180, 158, 147, 157] using the traditional UI design patterns/-models (i.e. *MVC* or *PAC*), while attempting to enable the capability of managing multiple

heterogeneous UAVs with human-in-the-loop [64, 108, 176]. Althought this approach is robust in many instances involving less autonomous and less complex UAVs, cannot cope with the increased complexity of the autonomy [57, 59, 149, 172, 168]. Consequently, negatively impacting the human operator's cognitive performance by increasing their CW [86, 163, 45, 53, 68], reducing their SA [85, 69, 206] and Automation Trust [198, 61, 126, 136]. As a result, the human operators experience complacency and skill degradation [127, 173] by automation a combination of misuse, disuse and abuse [172, 82].

This section explores and reviews the UI design models used in the implementation of the traditional GCS interfaces, the foundation of the newer process of interface design (Ecological Interface Design - EID), its process of cognitive and work domain analysis, as well as an alternate way to represent information through the usages of dialogues.

2.2.1 User Interface Design Models

Two UI design paradigms/models that are commonly used with the development of existing GCS UIs: the *MVC* design pattern [122, 88, 186] and the *PAC* model [71, 54].

The *MVC* paradigm is a robust, yet simple approach to human-computer interface development. It is a framework designed to assist software developers to organise and manage the development of a more effective and maintainable user interface.

The formalisation of the *MVC* paradigm was backdated to pre-1980s [186] where version 2.5 of the Smalltalk-80 programming interface (a state-of-the-art graphical software development user interface) was implemented using the *MVC* paradigm in 1980 [122]. Since then, the paradigm has been actively used in various domains of interface development [23, 180, 81, 64] including the development of the current UAS GCS interfaces.

The *MVC* paradigm is constructed of three components (Figure 2.7):

- Model (M): A representation of the system's knowledge, data, and rules
- View (V): A visual representation of the application's model textually, or graphically, and
- Controller (C): The physical interface between the user and the system, where the user's manual input are translated into the format acceptable by the software systems

Many examples of the current UAS GCS can be found; such as the Vigilant Spirit Control Station (VSCS) user interface in the Cooperative Operations in Urban Terrain (COUNTER) program [82], the interface for fixed-wing mini-UAVs proposed by Quigley, Goodrich and Beard [180], and the more recently improved user interface for the multiple gliding UAV experiment proposed by del Arco *et al.* [64]. Although these examples had not discussed the underlying design principles, it could be analysed using the traditional *MVC* design paradigm.

The typical design when the traditional *MVC* paradigm is applied was that the functional information of the UAVs; such as sensory data, mission data, control algorithms, and

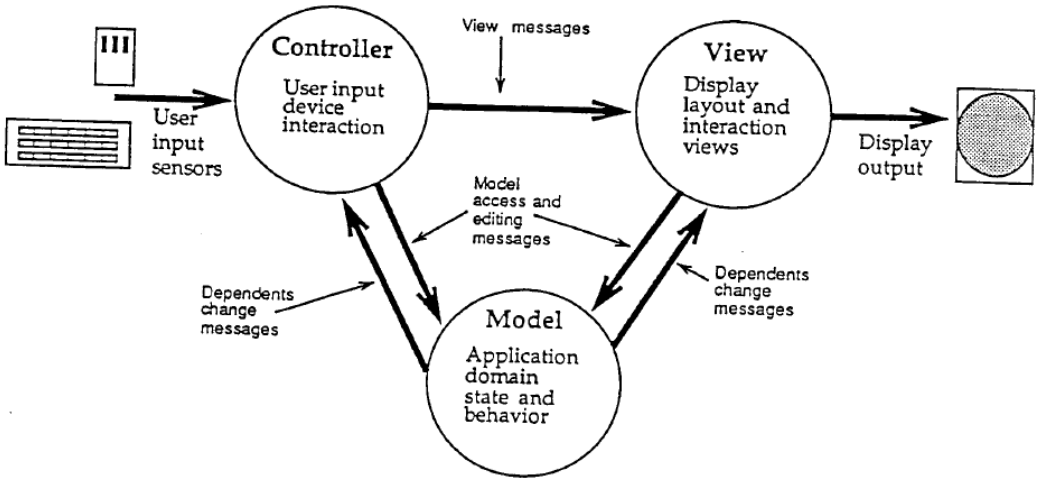


Figure 2.7: The three components of the *MVC* design pattern: Model, View, Controller (Published in [122])

other informational data were encapsulated and formed the *Model* component [157]. The process of updating the model's information is triggered by the sensors [64, 10] moderated by internal algorithms [82] or manipulated by the human operator formed the *Controller* component [115]. The direct visual representation of the model in a humanly comprehensible way; such as the panoramic camera view of the field (Figure 2.8) [187], or the numeric Inertial Measuring Unit (IMU) data [176, 82] formed the *View* component [86, 10, 176].

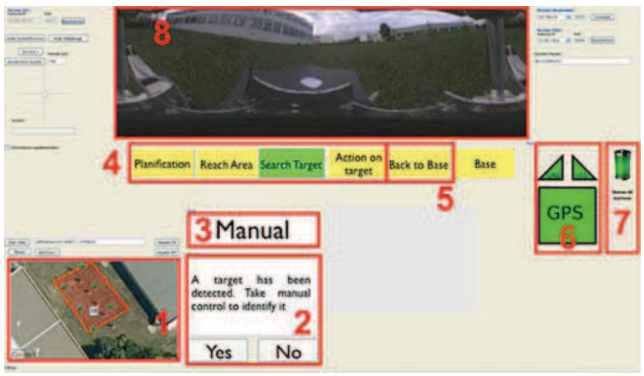


Figure 2.8: Example GCS interface, where (8) illustrates the panoramic view captured by multiple cameras (Published in [187])

The *PAC* model aims to interface the work domain/system's abstract, theoretical representation into a practical UI implementation (Figure 2.9). This model consists of three components:

- Presentation (P): A visual representation of the application's model data to the user textually or graphically, and enables the input behaviour to the application made by the user

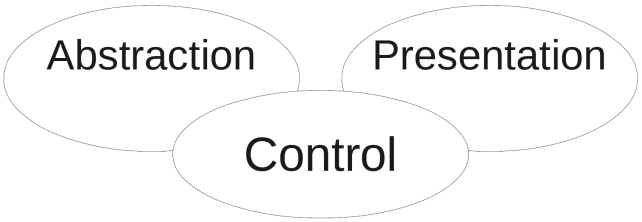


Figure 2.9: The three components of the *PAC* model: Presentation, Abstraction, Control (Published in [54])

- Abstraction (A): The functions which describes the work domain's abstraction of data, and
- Control (C): The bridge between the abstraction of system data in the "Abstraction" component and the input/output behaviour in the "Presentation" component

An example is an implementation of a pipe object as a part of the monitoring simulator of an industrial processing plant illustrated in Figure 2.10 using the *PAC* model [54]. The *Abstraction* component of this example includes all the numerical information regarding the process; such as the pressure, temperature, size of the pipe, and the viscosity level of the fluid. The *Control* component moderates and manages the system's representation in the *Abstraction* component; such as the pressure control valve and warning systems, which are initiated by the simulator user in the *Presentation* component. The *Presentation* component includes an output of information and a user input. The output is responsible for visually representing the abstracted data: such as the pipe's internal pressure, fluid temperature, viscosity etc., in a textual or graphical form. The user input is responsible for accepting physical interactions made by the user; such as mouse actions, keyboard inputs etc. The *Control* component interprets this data and translates the actions into messages recognisable by the software and parsed into the *Abstraction* component to update the data models.

Although the two models are often assumed to be identical; in some cases, the *Presentation* component in the *PAC* model is said to equate to the *View* component in the *MVC* model where the *Abstraction* component is said to equate to the *Model* component; and the *Control* component is said to equate to the *Controller* component, a clear distinction must be drawn. In fact, the two models behave very differently [71]. The *Abstraction* and the *Presentation* components of the *PAC* model do not communicate, as all exchanges of system data are mediated by the *Control* component, whereas the *View* components in the *MVC* model observes changes in the *Model* component (implemented with the *Observer Design Pattern* [84]), and the *Controller* component directly manipulates the data in the *Model* component.

The resulting products from using the traditional *MVC* and *PAC* models to develop GCS UIs for the modern UAS technologies given their advancements [148, 195, 151] and the cooperative nature [37, 33, 195], is disadvantaged. This is due to the fact that a collection of unified models to encapsulate the UAV's functional [8, 41, 44] and autonomy behaviour

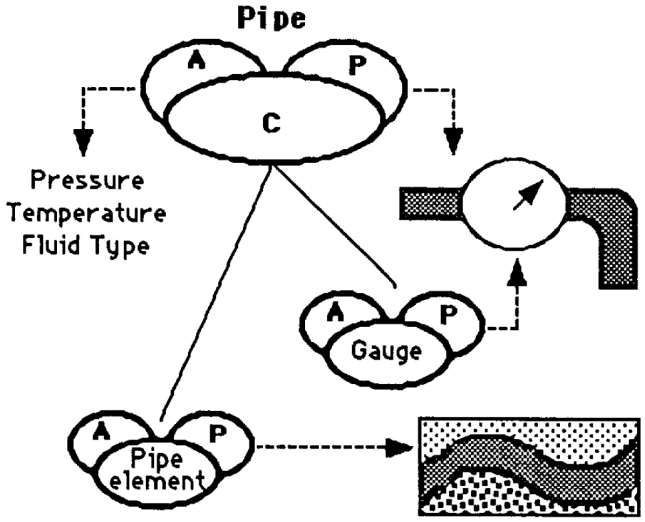


Figure 2.10: An example of a pipe object of an industrial processing plant monitoring simulator, implemented using the *PAC* model (Published in [54])

[162, 43], a novel way to visually represent these models [43, 40, 225], an effective user-model-controller that moderates the shared authority process [89, 128, 197], at the same time, demonstrated positive effects on the human performance [165, 38, 69, 220, 154] is not yet available [39, 37].

Aubert *et al.* [10], Lemon *et al.* [129], Saget *et al.* [197], and many other researchers have identified this shortfall, particularly in the realm of multiple UAV management, and commenced research in these areas. These research include the modelling and classifying of autonomy states and human-machine collaboration [127], proposal and modifying existing graphical representation of UAVs, and investigating multimodal and mixed-initiative interaction methods [130, 129].

2.2.2 Ecological Interface Design

As the complexity of sociotechnical systems increased, operator workload also increased [197, 37]. Subsequently, higher degrees of automation were designed and used to complete repetitive and predictable processes, reducing the operator workload [168, 132]. However, automation could only be used for tasks which were repetitive in nature and deployed in a predictable environment. Knowledge workers were still required to dynamically solve problems which were unpredictable and non-routine [168]. In order to maximise the human operator's cognitive ability to solve problems, Rasmussen and Vicente identified that the method of designing the human-machine interface was an important aspect in increase productivity and reducing cognitive overhead for the knowledge workers. However, minimal methodologies or research was conducted in this area. To address this niche, the Ecological Interface Design (EID) was proposed [224].

EID has been widely implemented in many domains of interface designs; such as nuclear power plant monitoring [95, 29], chemical refinement processing [109, 110], patient

monitoring [138], military command and controls [185], and more recently, interest moves towards the design of the state-of-the-art multiple heterogeneous UAV management interfaces [145, 43]. EID approaches interface design from a top-down perspective by abstracting the task and work domain in a hierarchical order. It focuses on the functional capabilities of the machines from a human-centred perspective rather than the traditional, human-decentralised design, which focused on the machine's low-level subsystem information. Interfaces that were human-decentralised often proved to include significant amount of detail and information about all aspects of the machine, caused operator workload overload, consequently, reducing their quality of SA [42, 44]. In a supervisory control of multiple machine agent situation where the agents possess a high level of autonomy capability; such an in depth agent subsystem information may not be necessary. This causes the operators to remove that information and overtrusting, or over relying on the onboard capabilities [167, 39].

The foundation of the EID design paradigm is Cognitive Systems Engineering (CSE), and Rasmussen and Vicente initially proposes three main models: the *Abstraction Hierarchy (AH)* [183], the *Decision Ladder*, and the *Skills, Rules, Knowledge (SKR)* taxonomy [182].

Abstraction Hierarchy

The AH aims to structure a system's information from a functional to a physical domain. The hierarchy, initially proposed by Rasmussen [183], is organised in five levels: Functional purpose level, abstract function, generalized functions level, physical function level, and physical form level (Table 2.9). These levels are described with the *Means-Ends* and *Part-Whole* relationship.

Table 2.9: A simplified description of the levels of abstraction adopted from Rasmussen (Adapted from [183])

Levels	Property Description
Functional Purpose	The purpose of the system's existence in the environment
Abstract Function	The system's high level representation in a generalised form
Generalised Functions	The high level description of the system functionalities
Physical Functions	The physical represntation of the functional states
Physical Form	The appearance, materials, and configuration of the system

The *Means-Ends* relationship was described by the *why-what-how* relationship (Figure 2.11 [9]) in the levels. The upper level describes the *what*, its upper level describes *why* this 'what' is, while its lower level describes *how* this 'what' could be.

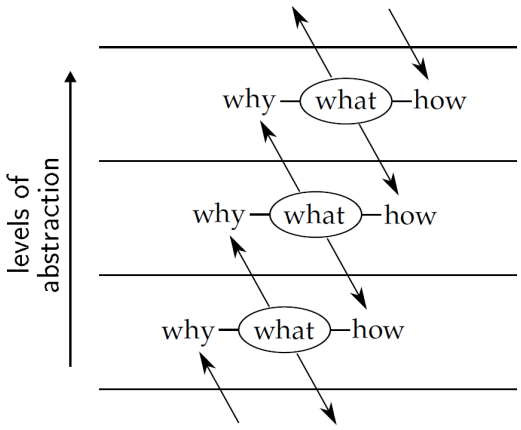


Figure 2.11: The *means-ends* relationship between Abstraction Hierarchical levels (Published in [9])

Decision Ladder

The decision ladder is a process which determines the decision sequences and mental strategies of a human operator in a work domain. There are two forms of decision making: Formal decision-making (skill and knowledge based behaviour) [223] and heuristic decision-making (rule-based behaviour) [184]. This ladder is a recognised structure for human decision-making, which is constructed from a series of nodes including data processing activities (rectangular nodes) and resulting knowledge from the activities (elliptical nodes) as illustrated in Figure 2.12. It is primarily used in to analyse cognitive systems and cognitive work domains, it is also accepted that the process is more significant when analysing work domains of existing systems as oppose to conceptual systems [9].

Skills, Rules, Knowledge Taxonomy

The SRK taxonomy proposed by Rasmussen [182] aimed to classify human information processing where three cognitive levels: Skills-based, rule-based and knowledge-based behaviour were identified.

Skill-based behaviour was referred to a human's instinctive action [229], where environmental variables were detected by the human's sensors, and instinctively and without conscious control, the human's motor systems react to this stimuli [182]. For example, the intention of "picking up a glass" [14] does not require any conscious, step-by-step thought to be completed. The human subject can instinctively perform the task base on the sense of touch and strength.

Rule-based behaviour was referred to a human's reflective action [229], where conscious processing is necessary for situations which were familiar or known in the human's mental model for this world. Conscious effort was used in the selection and the sequencing of control rules or cognitively stored procedures, which were previously established by the human being in similar situations through knowledge-based information processing. For example, to solve a complex engineering problem, human engineer are required to draw

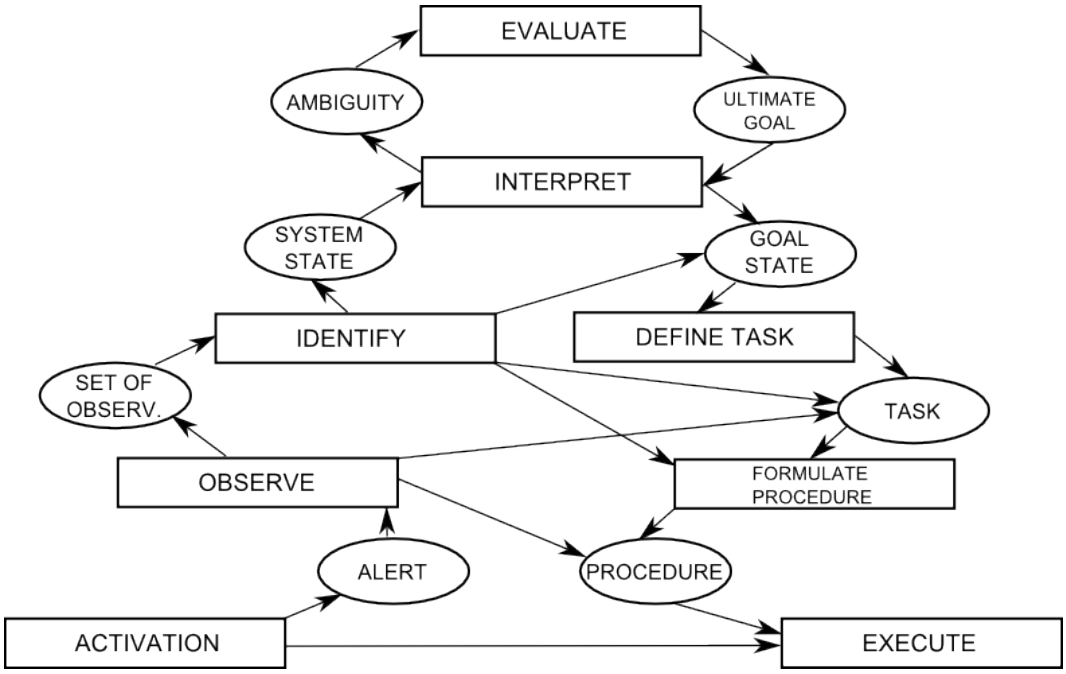


Figure 2.12: A model of the decision ladder, constructed with rectangular nodes representing data processing activities, and elliptical nodes representing resulting knowledge from the activities (Published in [184])

upon his/her knowledge inventory, selecting the appropriate rule sets and procedures, and apply them in a logical fashion.

Knowledge-based behaviour was referred to as the *symbolically conditioned action* [229] where new rule sets were established in situations which were unfamiliar or unseen by the human subject previously. Logical reasoning and other cognitive efforts were required to complete this process, and as a result, high levels of cognitive workload are associated [182]. For example, the process of observing, diagnosing and determining a safe decision to carry out when an aircraft pilot experiences abnormal flight behaviour of the aircraft, and especially when the sensory information is contradictory to any of the situations the pilot may have experienced during the flight training stage.

Decision ladder was to support the process of decision making. As compared to the abstraction hierarchy presented in the section above, it was to support the knowledge-based information processing by promoting the levels of abstract machine information to support human decision making strategies [184].

Transparency in the system information must be achieved between the human operator and the machine agents in both directions in order to effectively support human decision making strategies [197]. This can be achieved through the use of natural language dialogues [43].

2.2.3 Dialogues

Several researchers have identified that in order to enable effective human-machine interaction, a mutual understanding of each other's state is crucial [53, 63, 61, 42, 37]. As such, innovative modalities of human-centred controls to reduce the operator's sense of environmental isolation (created by the remote location of the operator from the mission scenario); such as spoken/written language exchange, haptic displays, gestures etc., were introduced [154]. The most direct and natural way [46] to establish such understanding is through the use of natural dialogue interactions [130, 6, 128, 81].

Natural dialogue interactions can be achieved through speech dialogue [6, 129] and graphical representation [40, 86, 213]. Each of these methods posed its strengths and limitations [129, 213].

Shneiderman [213] had identified that significantly greater cognitive resources (used for problem solving and memory recall) were necessary to perform the act of speech input/output, where other modalities of interaction involving hand-eye coordination require less cognitive resources (Figure 2.13 [213]). This observation was supported by a study conducted at the University of Maryland in 1993 [118] where the experiment subjects experienced greater difficulties in issuing voice commands which required the memorisation of certain mathematical symbols. The action of issuing voice commands drew upon the human subjects' short-term working memory resources, when compared to subjects which used a computer mouse. Furthermore, speech dialogue exchange requires a robust, yet reliable speech recognition system to prevent further cognitive resources expended on error-correction tasks [213].

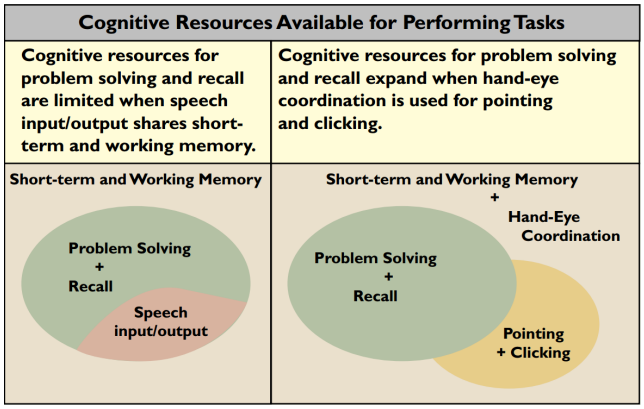


Figure 2.13: A simple resource model illustrating that cognitive resources are limited, and the use of speech consumes a portion of the cognitive resources (Published in [213])

In human-human communication, it is well understood that a voice message consists of three components: 55% of the intention of the message comes from human body language, 38% comes from vocal tonality, while only 7% comes from the actual spoken contents[143, 144]. In order to effectively communicate a message from a human operator to a machine in a human-to-human fashion, all three components must be communicated. However,

even modern day speech recognition system have not yet been able to effectively achieve such capability [106].

Allen *et al.* [6] and Lemon *et al.* [130] had challenged Shneiderman's discussion on the limitations and presented the strengths of natural speech-supported dialogue. Allen *et al.* [6] identified that in interfaces which either have very large GUIs or highly complex mission tasks, graphical presentation of the information becomes infeasible. Operator tasks that require the constant use of their eyes and hands also limits the effectiveness of graphical representation of information as the operator's physical capabilities to interact with the interface is disabled [49]. Lemon *et al.* [6] have agreed with many of Allen *et al.*'s argument and claimed that powerful interaction between human and machine could be achieved through natural collaborative dialogues, hence, provided greater advantages over pure graphical information representation from a human-centred perspective.

Lemon *et al.* [6] had also identified several key aspects that supported their claim. An advantage of the naturalness of human-human speech dialogues, human UAV operators or mission commanders required minimal specialised training to interact with the machine agent. Furthermore, human operators could issue timely instructions to the machine through natural dialogues, encouraging seamless collaboration between the agents. Another advantage of speech interface identified by Lemon *et al.* is the hand-free capability, where physical interaction with the interface is not available. Hands-free operation enables the human operator is not limited to physical manipulation of controls, hence, lessens their physical demand and cognitive load.

Dialogue, defined by Lemon *et al.* [130, 129], adopted by Saget, Legras and Coppin, and Chen *et al. [128, 43]*, and supported by Allen *et al.* [6], as *a truly collaborative process between two (or more) participants, whereby references and tasks are negotiated and agreed on, often in an incremental manner.* In the context of this research, at any point of the interaction, the two participants are the human agent, and a machine agent. Although the problem of multiple machine/UAV agents are the focus of this research, at any point of dialogue, there was only one pair of participants participating in the process. The materials, or references/tasks being communicated was a two-way interaction, where one participant initiates the dialogue, it could only be concluded when both participants have agreed.

The points of view presented by Lemon *et al.*, Allen *et al.* and Shneiderman were assessed and evaluated. The use of dialogues to portray and communicate human and machine agent messages was an innovative mode of communication, as it draws upon the natural way humans talk to each other, which enables a more seamless collaboration between the agents.

A synthesis of the pros and cons inherited in both the graphical representation of information and the natural speech recognition capability was elicited. The conclusion was that natural language dialogue is an effective modality of the communication of system information. However, due to the current technology limitation, the natural speech mode of interaction are not yet sufficiently robust to be incorporated, while pure graphical

representation of information restricted the other modes of interaction for the human and the machine agents. For this reason, the advantages of using natural language for information communication is harnessed through direct text-based representation [43].

2.2.4 Authority Sharing

Authority sharing as described by Mercier *et al.*, is the concept where tasks and resources, determined by the mission goals, are collaboratively distributed to both the human and the machine agent, and that both parties not only are necessary to be aware of the others' role, but they must also be aware of the overall objective [146].

Two forms of authority sharing modes were identified in Sellner *et al.*'s work [207]: System Initiative Sliding Autonomy (SISA), and Mixed Initiative Sliding Autonomy (MISA). SISA enables the machine agents to share authority with the operators on demand, where the machine agent can initiate requests, seeing assistance by delegating some responsibilities to the human agent. MISA in comparison has a lower LOA, as the human agent can intervene with the machine agents' operation at any time.

The strengths of the machine agents and human operators were identified [172]. The machine agents have greater capacity to do computational/repetitive tasks, while the human operators are superior in dynamic problem solving. In order to effectively complement the strengths of the human operator and the machine agent, automation must be adaptively incorporated into authority sharing [63, 62].

Saget *et al.* in their study on cooperative interfaces for UAV swarms investigated the concept of authority sharing from the perspective of actor interactions through sharing of their internal status [197, 43]. They stated that an actor (a human operator or a machine agent) must acquire the understandings of the other actor's internal representations, and this is decomposed into three categories [197]:

- Models and representations of the situation: this refers to the representations of the world-model, wherein all systems are encapsulated
- Models and representations of the system: this refers to the representations of the system or the machine agent, where the autonomy is applied, and
- Models and representations of the operator: Similarly, this is refers to the representation of the human operator and his/her technical capabilities, cognitive abilities, performances etc.

Furthermore, the actor's understanding of the information carried in these models and representations must conform with their expectations. If conflicts arise between the system information and it's human understanding, hazards in the sharing of the mission authority are caused[107, 175, 221]. These situations cause the human actors' attention resources to be almost completely devoted to conflict resolution [62], neglecting other critical visual or auditory information [63], reducing the effectiveness of the sharing of decision authority between the actors. This phenomenon is known as attention tunnelling [93].

Cognitive load, and its allocation resources such as CW [37, 38, 89], SA [89, 197], and Autonomy Trust [197, 39, 37, 126] continue to be crucial elements to effectively use authority sharing in control and commanding UAVs. The effective exchange of both actors' status information is a challenging yet important milestone to address [128].

2.2.5 Belief-Desire-Intention Model

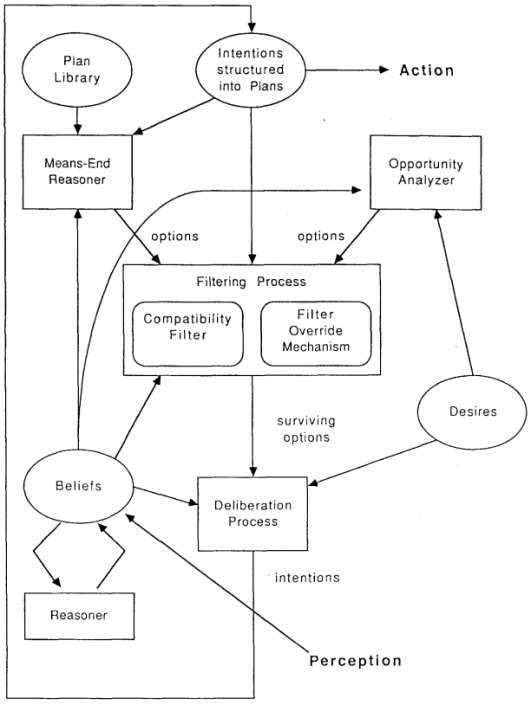


Figure 2.14: High level process for the BDI architecture (Published in [22])

The Belief-Desire-Intention (BDI) model is a software design architecture that aims to encapsulate the key principles behind human practical reasoning developed by Bratman [21, 22]. Illustrated in Figure 2.14, Bratman organised human practical reasoning into a repetitive process, encapsulating the steps starting from *Perception*, ending with *Action*. The stepping process can be summarised in the following list:

1. The *Perception* acquired from situations forms *Beliefs*, which then either "check" for any changes in the situation, or alternatively goes to the *Means-End Reasoner* where a queue of possible options from a *Plan Library* will be suggested

2. The next step is the *Filtering Process,* where plausible action plans are received and decided. There are two avenues to reach this stage:

   (a) Opportunity Analyzer: In the "checking" process, the *Opportunity Analyzer* presents situational updates, which are determined with the initial *Desires* and leads to the *Filtering Process*

(b) Means-End Reasoner: The possible options, generated through the *Means-end Reasoner* the are only considered if they provide the *means* to achieve the *end*. The options then are presented to the *Filtering Process*

3. The filtering process considers all plausible options presented from various avenues, considers the ultimate aim (or the *intention*), then structures it into *Plans*, and

4. Finally, plans are then executed in terms of *Actions*.

This process is considered important for the design of information representation because knowledge. This requires to be conveyed to a human with the purpose of allowing the human operator to make sound decision about the next-steps. However, in the application of this architecture, considering that the interest of this research lies in the representation of machine agents' autonomy information, only the *Desire* and *Intention* steps of this process need to be considered as the *Belief* step are performed inside the system [181].

2.2.6 Adaptive Automation

Opperman initially defined adaptive or adaptable automation as the information or decision support that is not fixed at the design stage but varies appropriately with context in the operational environment [164]. This posed a solution to the issues associated with fixed autonomy levels, which do not accurately describe automation behaviour of machines in the operational environment [107, 204, 169].

Adaptive automation was identified to have a positive effect in reducing CW and increase SA of the human operator during manual operation (low LOA) [196, 151]. On the other hand, particularly in command and control systems involving the interaction with multiple machine agents under stress and pressure, the effective human-autonomy collaboration can only be maximised when it is appropriately applied to the information-gathering stages of the human information processing model [128, 48, 173], with the autonomy information adequately and appropriately relayed back to the operator [39, 42, 43].

Limitations were identified that with inappropriate allocation of autonomy to the system and inappropriate information feedback to the human operator. High LOA could lead to a negative impact on the operator's cognitive capacities, such as the loss of SA for autonomous tasks [196, 42, 40, 43], and that the operators are prone to be caught in a surprise when autonomy were incorrectly triggered, and task states change as a result. This usurps the delegation authority from the human operator [16].

Parasuraman *et al.* proposed a four-stage information processing model with the scope of applied LOA [173] (Figure 2.2). This model divided the stages into two groups: Information, and decision. The information group included Stage 1 and 2, which are information acquisition and analysis. Whereas the decision group included stage 3 and 4, which are decision selection and action implementation [168, 15]. Furthermore, Parasuraman *et al.* suggested that very high LOA can be applied to the information stages (acquisition

and analysis) without significant impact to the operator's cognitive performance [173]. However, effective interaction between the human and the automation cannot be achieved if transparency of the agents' autonomy is not established. Past studies has shown that human operators experienced difficulty in understanding their autonomy agents' behaviours, reasoning process and expected outcomes [61, 126]., This leads to greater cognitive resources expended on determining the agents' capability, accuracy and effectiveness in their actions [18, 208, 119].

2.2.7 Autonomy Transparency

Transparency, traditionally had been identified to contribute significantly to the effective interaction between the human operator and the machine agent(s) in improving CW, SA, automation surprises, and human trust [39, 149, 196, 151, 16, 134].

Miller stated that the concept of transparency is naively understood as a straightforward property that a system's functions and behaviours, as well as the rationale behind them, are available and obvious to human users. That is, the system's inner workings are considered to be completely "see-through" [149]. However, to achieve such transparency, many predictable and unpredictable challenges in the design and implementation of human machine interface must be overcome. Furthermore, complete transparency can only be achieved when the human operator has a complete awareness[1] of the machine agents' states, as this defeats the purpose of autonomy [149].

Mercado *et al. [145]* and Chen *et al.* [39] had recently evaluated autonomy transparency based on the SA-based Agent Transparency (SAT) model. Their work extends the three most common challenges in human-automation interaction as identified by Sarter and Woods [202]: p*erception of automation system states, comprehension of their intentions based on their behaviour, and the projection of their future events.* These three challenges relate closely to the three level of SA proposed by Endsley [75, 76]: Situation perception, comprehension, and projection.

Chen *et al.* have responded to these challenges in their SAT model, which incorporates the system's *Purpose, Process,* and *Performance* (3Ps) [39]. The information from the 3Ps and its history, when communicated to the human operator will improve the operator's understanding of the system [126].

The SAT model consists of three levels of agent transparency as illustrated in Figure 2.15. The first level of the SAT model consists of the status of the world which the machine agent(s) is(are) in and their current state/goals, intentions, as well as their proposed actions. The second level consists of information which connects the machine's goals and intentions to support the justification of their proposed actions. The third level consists of projected states, that is, events or consequences that the system is projecting based on the current status and conditions [36, 35, 145].

[1]The term *awareness* used in transparency context denotes the subject's understanding of some particular entity. It is not to be confused with the cognitive metric of Situation Awareness.

Figure 2.15: SAT Model (Published in [39])

The impact on the human operator's performance was evaluated through a series of experiments which injected multiple Unmanned Vehicles' (UVs') autonomy information based on the SAT model which was conducted at the U.S. Army Research Laboratory with a modified FUSION simulator system [145]. The experiment included three interface configurations:

- SAT Level 1: Basic information only;
- SAT Level 1+2: Basic information and reasoning;
- SAT Level 1+2+3: Basic information, reasoning and projection of uncertainty;

The authors have demonstrated overall improvement in the human operator's calibrated trust at increased level of SAT without increasing cognitive workload [145].

This research adopts the definition of autonomy transparency proposed by Chen *et al.* as the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process [145, 39]. However, investigation was conducted on a less sophisticated and complicated simulation platform which Mercado *et al.* had available to them, to harness the effect of autonomy transparency on human operators, through representing autonomy information using natural language representation.

2.2.8 System and Agent Transparency

Many authors had identified the importance of transparency beyond autonomy. Saget *et al.* had stated that interactions between the human operator and the machine agent is crucial to maintain transparency between the actors in concept of authority sharing.

This importance of considering the system's transparency in the context of human-machine interaction was initially motivated by the influence on human trust in the machine's automation. Research had shown that human operator's perception of automation reliability is influenced by their trust in automation [226]. Furthermore, greater transparency was achieved through the feedback of the machine's limits and boundary information, promoting a higher automation trust recovery [72]. However, even inappropriate transparency (not-optimised type and amount of information), could impact trust. This was evident in Kim and Hinds's study where autonomous information was only available in high or low, that is, the human operator only had knowledge of whether the autonomy agent had a high level or a low level of autonomy. This attributed to the blaming of system for error [120].

Lyons defined transparency in the human-robot interaction context as the *the shared awareness or intent* between a human operator and a machine agent [133] and encapsulated the human-robot transparency in four models: an intentional model, a task model, an analytical model, and an environmental model [133]. An intentional model of the autonomy (robot) agent portrays the intention of the robot. That is, the purpose of why the robot would exist; such as to perform actions that are harmful to human beings, or perform high frequency repetitive tasks. This model includes information which provides the design, intent, and purpose of a system to the human operator. Furthermore, the human operator must also understand the robot's design principles, morals and rules of interaction. That is, it is crucial for the human operator to understand the bounding condition of the robot. For example, the reasons why the robot cannot be vetoed by the human operator in certain conditions [133].

A task model of the agent includes four components: The comprehension of a particular task within a cognitive frame, its goal at an epoch, the progress of the goal, the awareness of the robot's capabilities and errors. It is crucial that a shared awareness is established between the robot agent and the human operator by communicating the robot agent's understanding of the situation and its tasks to the human operator. At the first level, this can be achieved by communicating its task adherence process to the human operator. Furthermore, another crucial facet of this model is the robot agent's awareness of its functional capabilities. That is, the agent's understanding of it's own capabilities, as this contributes greatly towards the appropriate trust the human operator has towards the robot agent [133].

Analysis, or data processing is encapsulated in the analytical model. The model aims to support the human operators to make better decisions through the communication of its underlying analytical principles (i.e. how it performs analysis). This is very useful in situations where the robot agent performs certain acquisition, compilation, and processing tasks and the human operator can then be confident of the process or lacking thereof [133].

An environmental model of the agent encapsulates the state of the environmental conditions it is in. One of the primary aims of deploying robotic agents in the field is to be exposed to the environmental elements that may be harmful or high-risk for a human

agent. However, with the human agent/operator disconnected from the environment, it reduces their SA greatly, and the aim of communicating the environmental model is to increase operator SA by displacing the robot agent's representation into its context [133].

The communication of the status of these models presents transparency of the system in a human-robot system. This transparency must be effectively represented in command and control interfaces used by the human operators [133, 10, 151, 145, 134].

2.3 Cognitive Constructs

An effective human-machine interaction involving multiple machine agents requires not only a suitable amount of autonomy applied to the information acquisition and analysis stages of the machines [173] and the communication of this information [28, 37, 39]. The cognitive impacts associated with the information [37, 39, 42, 43], if not appropriately monitored or calibrated [126], a number of cognitive shortfalls could emerge such as the increase in CW which leads to a reduction in SA [74], complacency, skill degradation [127, 173], and the misuse, disuse and abuse of automation [167, 125].

Many cognitive attributes have been investigated in prior studies including automation over-reliance, decision biases, mistrust, and complacency [168, 15, 172, 203, 210, 231]. However, Parasuraman, Sheridan and Wickens [174] had identified and focused on three particular constructs: CW, SA, and Automation Trust (Trust). The authors concluded that these constructs are valuable in the understanding and predicting of the human's cognitive performance in complex systems, and heterogeneous multi-agent systems in the scope of this thesis is considered as a complex system [43, 165, 37, 38, 58].

The impact of human-machine interaction on the human operator's CW, SA and Trust in automation had been investigated individually [58, 163, 39, 134, 61, 96] and collectively [43, 42, 44, 165, 37, 38, 174, 195]. These studies had identified the importance, hence, conducted evaluative and comparative studies using these three constructs to gauge the success [58, 45, 101, 68, 206]. This section presents the significance of each cognitive construct in the context of human interaction with machine's autonomy information, and discusses the main measuring techniques and procedures.

2.3.1 Cognitive Workload

Cognitive workload was identified to be a multidimensional construct incorporating many aspects of both physical and psychological. Reid and Nygren described workload with three dimensions: *Time Load*, *Mental Effort Load*, and *Psychological Stress Load* [189] in their proposal of the Subjective Workload Assessment Technique (SWAT). However, Hart and Staveland had further encapsulated these dimensions and proposed six dimensions to describe workload in their proposal of National Aeronautics and Space Administration (NASA) Task Load Index (TLX) [97]: *Mental Demand, Physical Demand, Temporal Demand, Own Performance, Effort,* and *Frustration.* Similarities can be observed from four of these

dimensions (Mental Demand, Temporal Demand, Effort and Frustration) to the three dimensions described in SWAT.

Due to the complexity and capability of autonomous systems, much of the work during a mission is currently, or can be performed autonomously. The Human operator may not only need to interact with tasks that are non-automated, they must also monitor, and at appropriate times, interact with these automated systems as well. These systems impose increased CW on the operator and they are a crucial construct to effective autonomy management [43, 173]. A number of CW assessment measures, each owing its advantage to focusing on a different part of an operator's CW were proposed in the past with three of which are well recognised [194]. A number of comparative studies were also conducted, assessing the measure's psychometric properties [194, 97, 99, 232].

This section reviews three out of the five most significant multidimensional subjective workload assessment instruments [194]: *NASA-TLX* [97], *SWAT* [188, 190], and *Workload Profile (WP)* [218]. All these instruments share the four common psychometric properties [194]; i*ntrusiveness, sensitivity, diagnosticity* and *validity*.

Subjective Workload Assessment Technique

SWAT was first proposed by Reid [188] in an effort to address the need for a *workload measure with known metric properties that is useful in operational or "real world" environments* [190]. It is a multidimensional workload assessment technique. There are three dimensions used to assess CW in SWAT [188]: Time Load (T), Mental Effort (E) and Psychological Stress Load (S), and each of these dimensions have three levels associated:

- Time Load: This dimension focuses on the availability of spare time and the overlapping of task activities; how much spare time, interruption or overlapping of task activities does the operator experience during the experiment. The three levels associated with this dimension (as directly quoted from Reid [188]) are:

    1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all
    2. Occasionally have spare time. Interruptions or overlap among activities occur frequently, and
    3. Almost never have spare time. Interruptions or overlap among activities are frequent or occur all the time

- Mental Effort Load: This dimension is an indication of the amount of attentional or mental demands that are required for the experiment or a particular task. It is assumed that with lower mental effort load; the levels of concentration and attention of the operator is very low, thus performance is very high. The inverse effect applies when the mental effort load is increased, the levels of concentration and attention of the operator also increases [188]. The three levels associated with this dimension (as directly quoted from Reid [188]) are:

1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention
2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required, and
3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention

- Psychological Stress Load: This dimension refers to the three common psychological conditions which could cause accomplishing the task or experiment to be more difficult, these conditions are confusion, frustration, and/or anxiety. With lower levels of stress, the operator will feel relatively relaxed. With an increase in stress, these factors will also increase, causing loss of motivation, fatigue, fear, degradation of skills, discomfort with the temperature, and increase of ambient noise, vibration and comfort [188]. The three levels associated with this dimension (as directly quoted from Reid [188]) are:

  1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated
  2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance, and
  3. High to very intense stress due to confusion, frustration, or anxiety. High to extreme determination and self-control required

NASA Task Load Index (NASA-TLX)

The NASA-TLX uses six dimensions to measure CW as defined in Table 2.13. These dimensions are used to describe mental workload experienced by the operator. As such, CW of a particular task using the NASA-TLX instrument can be assessed.

Procedure [194, 97]: The general procedure of this assessment instrument requires the human participant to complete a questionnaire during the conclusion of each of the experimental tasks. This will produce a CW score (TLX) from 0 to 100, indicating their CW level during the task. Prior to assessing the CW of the operator during the task, a Weighting Procedure must be completed. This procedure requires the operator to perform a Paired Comparison of all pairs of the dimensions in Table 2.10; choosing a dimension (out of each pair) which is more significantly relevant to the task. The number of times each dimension was chosen (15 pairs in total) becomes the Weighting (W) of that dimension. The operator then assigns a Rating (R) (a scale of 0 to 100 with a resolution of five; a total of 20 steps) to indicate the levels of workload for each dimension they experienced during the task.

Table 2.10: Rating scale definition and endpoints from NASA-TLX (Adapted from [97])

Attribute	Endpoints	Description
Mental Demand	Low/High	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low/High	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low/High	How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satised were you with your performance in accomplishing these goals?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	Low/High	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratied, content, relaxed, and complacent did you feel during the task?

For each dimension, the Weighting (total selections during paired comparison) then must multiply by the rating (from 0 to 100 in increments of five) to form the individual Scale Score (SS). The sum of these scale scores (a total of six scales), dividing by 15 (total number of paired comparison) will produce the CW score, or TLX.

Past Applications: This method has been successfully applied to a variety of experiments. These include real and simulated fight, air combat and UAV experiments [212, 30, 42, 44, 43, 89, 195, 196]. Due to the simplicity and the versatility of applications this method has demonstrated, it was used in this research for CW assessment.

Workload Profile (WP)

WP's dimensions are defined according to the resource dimensions proposed by Wickens in multiple resource models [230]: perceptual/central processing, response selection and execution, spatial, processing, verbal processing, visual processing, auditory processing, manual output and speech output. Upon completion of all tasks that the operator is required to perform for the experiment, a rating between zero and one for each of these dimensions are given; zero denotes task placed no demand on the dimension being rated [194], and one denotes the task required maximum attention [194]. Table 2.11 is the workload profile rating sheet, used during this questioning process. Once the scores for each dimension of all the tasks are collected, the sum of all these dimensions for each task

is obtained to form the overall workload rating, with a maximum score of eight.

Discussion

A study comparing NASA-TLX, SWAT and WP was conducted by Rubio *et al.*; where an individual and a combination of the Sternberg's Memory Search Task and the Tracking Task are carried out.

Sternberg's Memory Searching Task: This task required the participants to memorise a set of consonants prior to each test. There were two levels of difficulty in this test, involving memorising two and four alphabetical characters. During the experiment, the participants were required to indicate if a character that they were required to memorise is displayed in a sequence of letters. Dependent variables that were collected involved correct identification, errors and response time.

Tracking Task: In this task, the participants were required to maintain a cursor within a moving path with arrow keys. There were three levels of difficulty associated with this task, which was defined by the width of the path; level three denotes the widest path, which was the easiest level, level one denotes the narrowest path, which was the most difficult level.

Dual Tasks: This task was a combination of both the Sternberg's memory searching task and the tracking task, which required the participants to pay even attention to both tasks during the experiment.

Through various statistical tests, Rubio *et al.* concluded that the psychophysical ratings, NASA-TLX and SWAT were most sensitive to task demand manipulations, while WP was outperformed in this parameter [194]. However, WP was able to determine between memory and tracking tasks, having significant sensitivity [194].

Rubio *et al.* also suggested that WP should be used if comparison between the mental workload of two or more tasks with different objective levels of difficulty are of interest [194]. If predicting the performance of a particular individual in a task, NASA-TLX should be used [194]. Finally, if an analysis of cognitive demands or attention resources demanded by a particular task must be conducted, then SWAT should be used [194].

2.3.2 Situation Awareness

Endsley defined SA as a person's *perception of the elements of the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future [75, 76].* This encapsulated three aspects of SA, where they were commonly recognised as the three levels; *perception*, *comprehension*, and *projection*.

Level 1 SA is the lowest level of SA. It defined the person's perception of the situation or world. This involves the acquisition and analysis of the sensory data information acquired by the human. Level 2 SA is the middle level of SA, and this level can only be reached if level 1 SA was satisfied. It defines the person's understanding of the situation based on aggregating historical and present information acquired from SA level 1. Finally, level

Table 2.11: Workload Profile Rating Form (Adapted from [194])

	Workload Dimension							
	Stage of Processing		Code of Processing		Input		**Output**	
Task	Perceptual/Central	Response	Spatial	Verbal	Visual	Auditory	Manual	Speech
$Task_1$								
$Task_2$								
$Task_3$								
$Task_n$								
$Task_{n+1}$								

3 SA is the highest level of SA, and likewise, this level assumes level 1 and 2 SA have been reached. This level of defines the projection of situations based on the situation currently established in level 1 and 2, that is, at this SA level 3, the human should pertain the capability to anticipate future situations. Endsley had also defined SA in three zones of interest as illustrated in Figure 2.16; immediate, intermediate and long-term. These zones can be viewed as the importance of the any SA element; the closer elements are to the operator, the more important it is.

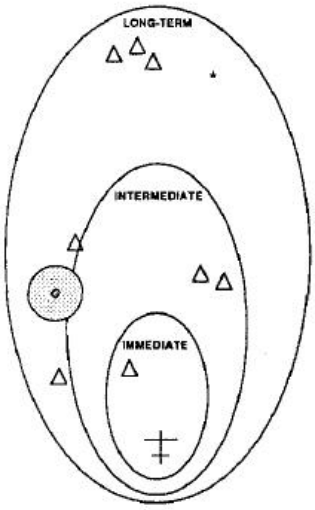


Figure 2.16: The three SA zones of interest (Published in [73])

This suggests that an operator's SA is crucial to the success and survivability of the mission, with a maximised SA, operators may provide greater overall human-machine interaction performance [73]. Therefore, to understand the human factors in a heterogeneous multi-agent system, it is important to understand the SA associated with the human operator.

SA is influenced by the functions of automation, where automation is used for substituting human involvement in the task all together. That is, in human-out-of-loop situations, SA was shown to reduce [139]. However, SA was shown to improve for automation functions that helps the human operator's cognitive resource to refocus on higher level tasks, such as monitoring or supervising of systems [163].

This section presents four generic methods to measure SA: physiological techniques, performance measures, subjective measures, and questionnaires. Followed by a discussion of the two common SA measurement techniques; SA Global Assessment Technique (SAGAT) and Situation Awareness Rating Technique (SART).

Physiological technique to measure any form of cognitive activity, such as SA, is a technique which involves external instruments to detect the human brain's physiological activity and interprets them to form conclusions on its cognitive information about the environment. Electroencephalography is one of the physiological techniques, which detects the electric currents generated by firing of neurons within the brain [100]. Electro-

encephalographic measurements were possible for a number of years, and it was proven trustworthy for determining cognitive information registration. Although electroencephalographs are capable of determining the operator's perception and processes of information about his/her environment, this method is not capable of detecting the amount of information which remains in the memory, or whether the information is even registered correctly. Although physiological techniques may provide useful data for other purposes, they are not very effective or accurate when measuring the human operator's state of knowledge [75].

Performance measures are generally considered objective and nonintrusive [75] . However, limitations of the performance measures are present in the following measures.

Global Performance Measures: Global measures of performance generally produce results following a long string of cognitive processing, providing very little information on the causes of poor performance and the reason for this performance within the given surroundings. Reasons for poor performance may be caused by poor sampling strategy, improper integration or project, heavy workload, poor decision making or action errors, or even a combination of multiple reasons. Many of these reasons may not even be related to SA.

External Task Measures: This form of measure requires removing of certain elements or pieces of information from the test subject's peripheral, then detect the amount of time required for the subject to react to this action [201]. This process is considered to be very intrusive, as the subjects are required to detect the missing information, react and reorientate himself or herself, as well as maintaining an unaltered, satisfactory performance with the initial experiment task [75]. The effects of the intrusive actions may not only lead to misleading results, it may also affect the attention of the subject, thus the SA and decision making [137].

Imbedded Task Measures: The ability to detect an operator's SA is not only limited to taking measures off the operator, but to examine the performance of the operator's particular subtasks that are of interest.

Subjective measure of SA can be viewed as the rating of the operator's SA from either his/her own subjective views about their SA, or with an expert's interpretation of the operator's SA. The advantages of these techniques are that it is cheap, and can be easily carried out. The limitation is that they are very "subjective" and may not accurately reflect the operator's true SA [73].

Self-rating: As the name suggests, the technique of self-rating is the operator's ability to rate their SA. This is limited by the lack of knowledge the operator has of the surroundings, or their knowledge of what they are expected to be aware of may be incomplete or inaccurate.

If results were to be collected post trial, then there is a possibility that the results will be objectively modified depending on the performance of the operator [222] as they might feel that their performance was satisfactory, thus they have a sound SA [159], or even purely good luck. This technique can be seen as a measure of a subject's confidence level

regarding their SA. Efforts had been made to improve this through the development of SART by Taylor [215], *which allows the operator to rate a system design based on the amount of demand on attentional resources, the supply of attentional resources and understanding of the situations provided.* [75].

Observer-rating: Another form of subjective measure technique is observer-rating. Trained observers observe the operator's performance and rate their SA. This method is slightly more accurate than self-rating, as the trained observers might have more information about the trial and the surroundings than the operator [75]. The limitation of this technique is that regardless of how well trained and experienced these observers might be, they will only have a very limited knowledge about what the operator's concept of the situation is, as the only indication of the operator's concept of the situation is through the interaction of the operator with the experimental interface. Therefore, this cannot provide a complete and accurate representation the operator's true SA [75].

Questionnaire: The questionnaire technique is one of the most common techniques to assess the human's cognitive states. This technique requires the operator to answer a certain set of questions from which can objectively assess their state of SA. This method is considered a more direct measure of SA as it taps into the operator's perception rather than inferences. There is no capacity for subjective judgment of SA [75].

Post Test: The measurement of the operator's SA is conducted immediately after the experimental trial. The test is arduous as it attempts to capture the detailed information about the operator's SA. The limitation of this method is that people are usually not great at recalling information about the past. Therefore, there is a tendency of overgeneralising and overrationalising their true perception during the experiment [75].

Online: Another method of questioning is conducted during the operator's simulated task. The limitations of this method is that the subjects will be placed under a very heavy workload, as the continuous answering of the questions will form a secondary task to the operator [75]. Furthermore, the questions asked may hint the operators to focus more or less on certain information of the experiment, thus altering the operator's true SA. This task is concluded to be highly intrusive on the primary task of the operator's system operation [75].

Freeze technique: This technique attempts to overcome the shortcomings of the other techniques mentioned above by asking the operator certain questions to determine their SA, while temporarily pausing the experiment. This will not cause undesirable workload to be placed upon the operator, while still achieving a more objective rating of their SA without needing to recall long periods of information after the test [75]. As studies had shown by Endsley that the 30-second pause of the simulation will not exceed the short term memory storage limits, and using this technique for up to 5 or 6 minutes had shown no apparent memory decay of the operator. Therefore, queries may be asked and answered during this time [75].

Endsley had proposed a method which adopts the freeze technique questionnaire assessment technique, SAGAT [76], to measure the SA of the operator.

Situation Awareness Global Assessment Technique

SAGAT was originally developed for the purpose of providing an objective measure of a pilot's SA with a given aircraft design and scenario. This method had become one of the most recognised questionnaire-type techniques to measure SA [76, 75] because it is able to measure the operator's knowledge of the situation, SA and other non-dependent sources.

The implementation of SAGAT involes a set of queries that are created to register the operator's knowledge of the situation at a given time. The query responses provided by the subject will be compared with the true information on bipolar scale (true or false). During query pauses, a selection of these queries covering the three levels of SA (perception, comprehension, and projection) [79] are administered [73, 76, 75, 77]. As more samples are collected from the subjects' answers to the queries, a clear understanding of the operator's SA is established.

SAGAT applies the analysed results to these zones to provide better picture of the operator's SA [73]. The general process of SAGAT can be summarised as [73]:

1. The operator performs the designed experiment as planned
2. During random times, the simulation will be paused and the experimental display will be blanked while a series of queries are administered, in order to determine SA
3. From a list of queries designed to capture the operator's perception of the situation, random samples are administered to the operator. This random sampling method allows consistency and statistical validity
4. After the experiment, the queried answers are evaluated based on the events of the experiment, a comparison of the real and the perceived situations will provide an objective measure of the operator's SA
5. Determine the SAGAT score using 3 zones of interest (Figure 2.16), and
6. This process is repeated for the number of times for different operators as subjects conducting the same experiment. SAGAT Advantages Due to the method of conducting SAGAT, there are number of advantages associated. By temporarily pausing the simulation and collect the operator's perceptive data during this time provides a snapshot of his/her SA. Rather than collection of data after the experiment [73], which could lead to limitations of the Post-Test measurement technique

Due to the global SA nature that SAGAT can measure, it takes an inclusive look at all contributing elements of an operator's SA, these then are objectively collected and evaluated [73], eliminating the limitations of any subjective measures. As studies have shown, this method will not cause any uncomfortable level of intrusiveness on the operators [75] who have not been previously trained or practiced a simulated mission involving SAGAT's workload or perception of the situation. As such, an accurate measure of the operator's SA can be achieved. To reduce the effect of stop and start routine of the simulation experiment

during a querying pause. Endsley had suggested that training should be provided to the test subjects for interrupts and getting used to the questioning routine [73].

It is also recognised that a number of advantages are present by administrating SAGAT as a form of measuring SA, but there are also limitations identified. Two of the most obvious limitations are that the experiment must be paused temporarily to allow data collection. This was later proven to not have great impact on the operator's workload and workflow [75]. Furthermore, if the simulation was to carry on without pauses, or temporary blank of the simulation screen, the operator has the opportunity to "cheat", thus inaccurate measurements could be taken [73]. If precautions are taken during design and conducting of the experiments, these limitations can be minimised or even eliminated.

Situation Awareness Rating Technique

SART developed by Taylor [215], is a subjective SA assessment technique designed for the assessment of SA in aircraft flight deck designs. SART has 14 components which are used to form the SA rating. Operators rate their mental resources and the understanding of the situation on a series of bipolar scales [79]. SART can be said to be a rating technique for operator's confidence level.

SART Advantages: There are two major advantages of SART over SAGAT. One of the advantages is that this technique is non application-specific, meaning no extra customisations will be needed for assessment of SA in different situations. The other advantage is that it can be administered to a wide range of task types.

SART Limitations: Due to the subjective rating nature of this technique, operators will not have a complete knowledge about the true situation. Therefore, results from this rating technique may not be a true representation of the operator's SA. This means, SART measures the operator's perceived quality of the situation as oppose to derived SA based on the factual/perceptual knowledge of the situation environment measured in SAGAT [76].

2.3.3 Automation Trust

The concept of the automation trust had been defined and discussed in detail in many previous studies and numerous trust measurement techniques had been published, mostly for human-human relationship [112, 124, 193, 191]. Two fundamental concepts were associated with the definition of trust: expectation (of outcomes) and intention/willingness (to act). Rotter [193], Rempel *et al.* [191], and Barber [11] commonly identified and defined trust as *the expectation or attitude towards another party's behaviours or outcomes to be favourable* [126]. Johns [112], Moorman *et al.* [152], and Mayers, Davis and Schoorman [141] extended the former definitions and identified trust as *the willingness or intention to behave in a certain manner or to enter into a state of vulnerability* [126].

Mayer, Davis and Schoorman defined trust as *the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a par-*

ticular action important to the trustor, irrespective of the ability to monitor or control that party [141]. In the context of automation trust, irrespective of the ability to monitor or control autonomy, in trusting it, the human operator enters a state of vulnerability to the autonomy agents' actions and behaviours.

Trust (or distrust) had also been identified to be the state which a human had been placed in as a result of certain actions performed by the other party or parties [66, 126]. This was encapsulated in Kramer's definition of trust as a state of perceived vulnerability or risk that is derived from an individual's uncertainty regarding the motives, intentions, and perspective actions of others on whom they depend [121].

Finally, Lee and See [126] summarised all the previous understanding of trust into a simple, yet conclusive definition as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. Only by understanding the implication on human operator's trust towards the autonomous party (agent) can one appreciate its significance.

Human operator's trust on autonomous agents/systems has been one of the most motivational cognitive constructs in humans' interaction with autonomous agent(s) [96, 126, 172, 205, 39, 72, 145]. The level of automation trust directly influences human's usage of the automation and the resulting consequences of the usage [12]. The appropriate level of trust, or calibrated trust, must be achieved to harness the full potential of human-automation collaboration; uncalibrated trust, such as the human overtrusting the automation, leads to the inappropriate use of automation [126, 172].

Parasuraman and Riley had identified a few types of inappropriate use of automation: use, misuse, and disuse [172]. *Use* was defined as the voluntary activation or disengagement of automation by human operators [172]. This sets the baseline understanding of human operators' engagement with automation, and trust significantly affects the decision in this use [209, 234], which Beck defined as Automation Usage Decision (AUD) [12] and adopted by Chen *et al.* in their investigation on understanding the influence of autonomy transparency on human trust [39]. *Misuse* was defined as the human operators' overreliance on automation, which can result in failures of monitoring or decision biases [172]. Overtrusting automation is potentially damaging to the human operator's SA, as automation may not always function optimally, while the human operator is overreliant on the automation and fails to notice indicators or behavioural patters of the system's underperformance [39, 126]. *Disuse* was defined as the neglect or underutilization of automation; it is commonly caused by alarms that activate falsely [172]. This described the lack of trust in the human operators towards automation, which led to the underuse of automation. One common reason for this lack of trust was the repetitive false alarms or incorrect behaviour pattern produced by automation, which did not align with the human operator's expectations. Although the automation was functioning optimally, making sound decisions, without calibrated trust, automation could not be used to aid human operators [43, 39]. Automation *abuse* was defined as the automation of functions by designers and implementations by managers without due regard for the consequences for

human performance, tends to define the operator's roles as by-products of the automation [172].

The effects and relationships between trust and automation-use had been acknowledged in many studies in the past [72, 133, 134, 43, 120, 151, 172]. Hence, the trust in automation is an important cognitive construct that must be incorporated in the design of autonomy visualisation methodologies to enhance system and autonomy transparency [39, 145, 72, 43].

Trust, as a quantifiable and measurable cognitive construct, has been generally accepted as a multidimensional and dynamic concept [126, 17, 111]. The dimensions of expectation (predictability), and motivational relevance was initially considered [66], while the reliability on the trustor could not be neglected [193]. Rempel *et al.* proposed that trust was a process of predictability (the expectation that autonomy will act), dependability (relying on the autonomy's actions), to faith (the autonomy will act favourably). And Muir and Moray further extended these three factors with competence, responsibility, and reliability [155].

During Jian, Bisantz and Drury's development of a definitive questionnaire to encapsulate automation trust, they had identified three limitations in the existing methods to measure trust [111]: 1) The questionnaires prior to 2000 were designed based only on the different, and contradictory theoretical foundations of trust [126], not empirical analysis on the components of trust [111]. 2) The concept of trust and distrust were assumed as bipolar. That is, the notion of trust and distrust was assumed to be completely opposites of each other. 3) There were a lack of empirically evaluated methods to measure trust between humans and autonomous systems, only in human-human interaction settings [124, 191]. To which end, Jian, Bisantz and Drury empirically determined and critically reviewed a scale for the trust in automated systems in a form of a questionnaire, and this questionnaire was employed by Mercado*et al.* in their most recent study to evaluate the significance of SAT on human operator trust [145].

An example of the questionnaire is illustrated in Figure 2.17. This questionnaire was developed in three phases [111]: 1) a word elicitation study, where various keywords were harnessed from the concept of trust and distrust, 2) a questionnaire study, where the words representing the concepts of trust and distrust were analysed to determine their pairwise relationships applied to the contexts of general trust, human-human trust, and human-automation trust, and 3) a paired comparison study, where experiment participants were requested to rate the similarity of the word pairs.

The results from their study had suggested that the concept of trust and distrust are indeed on the same scale [111]. However, that is not to say that these two concepts are bipolar, their results provided evidence to say that one continuous scale can be used to rate trust/distrust, where they are the extreme ends of the same concept, similar to automatic and manual. Furthermore, the results also suggested that the concepts of general trust, human-human trust and human-automation trust are similar. Hence, many previous studies on human-human trust can be considered in evaluating human-machine trust.

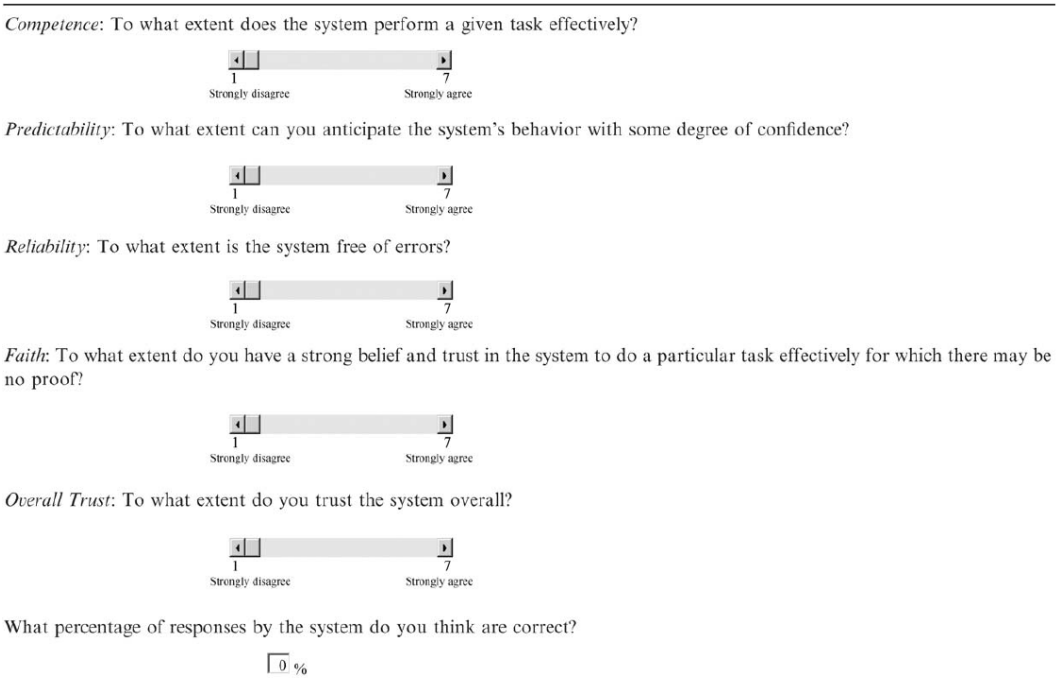


Figure 2.17: An example of the trust rating questionnaire (Published in [220])

As illustrated in Figure 2.17, there are 12 separate questions, each rated on a Likert scale from 1 to 7, where 1 denotes the lowest affirmation. An instantiation of this questionnaire was developed by Mercado *et al.* by incorporating the four stages of information processing to each of the questions [145]. Furthermore, they had added several extra questions, each also incorporated the four stages of information processing [145], forming an in-depth and thorough questionnaire to capture the human operator's subjective perceptive of their trust.

Although the questionnaires of both Jian, Bisantz and Drury, as well as Mercado *et al.* were thorough and inclusive [145, 111], experience obtained from previous studies conducted for this research showed that completing questionnaires absorb a large amount of time [44, 42], which was a major constraint in the research experiments for this topic [44, 42, 43]. Hence, another simpler trust questionnaire was required [43].

Uggirala *et al.* agreed with many points previously discussed by Jian, Bisantz and Drury [111] and had proposed to develop a trust assessment technique based on the concept of uncertainty. Their study had utilised Master *et al.*'s trust measuring questionnaire [140], which included five dimensions: competence, predictability, reliability, faith, and overall trust [220]. The dimensions covered the main aspects of trust in systems, and it is significantly simpler than Jian, Bisantz and Drury's questionnaire, hence reducing the data collection process of this research's experiments significantly [43].

2.4 Discussion

An ultimate aim was to understand the visual representation to reflect a UAV's capability (both functional and autonomy) to autonomously complete tasks corresponding to the relevant functional subsystems.

Research showed that with a single or multiple UAVs operating under a high LOA, human operators experience a reduction in CW, however, they also exhibited a reduction in SA as well [126, 170, 74], and without appropriate management techniques, the human operator will feel complacent [173, 127], reduce trust in the system [126, 37, 145], and consequently degrade in their mission performance [43, 42]. However, with a single or multiple UAVs operated at a much lower LOA, the operators reported a significant increase in CW, although their SA was maintained [116]. These evidences were confirmed by Villaren *et al.* in his study which involved interviewing 18 professional French UAS operators. In his interview, the operators reported that they commonly operated the UAVs with very minimal to no autonomy [225]. That was to say, the operators preferred to carry out manual operations of the UAVs, as a lack of the UAV's autonomy transparency led to hesitation to operate the vehicles at higher LOAs [172, 149].

Villaren *et al.* had focused on the macro-task of *Trajectory Management* in his study and a simulation environment was developed to assess the different modes of autonomy and the effects of transitioning between modes on the cognitive behaviours of a dual-operator team (one pilot and one payload operator) [225]. However, their study was limited to multiple human operators operating a single UAV, the effects of visualising multiple autonomously heterogeneous UAVs on a single human operator's performance is not yet understood.

Fuchs *et al.* had demonstrated that through the manipulation of the Ground Control Station (GCS) based on the the operators' preferred functional information, there was a significant increase in the mean operator SA score and a decrease in the mean CW score [86]. This suggests further understanding of the impacts of visualising the UAVs' functional autonomy capabilities without needing to know their exact LOA on human operator's performance was necessary to establish an effective medium of communication in a multi-agent system.

Coppin and Legras had proposed the concept of *autonomy spectrum* [51, 127] to address the challenge which involved in task-switching of UAVs under heterogeneity of autonomy. Villaren *et al.* studied the effect of a mixed operative (autonomy) modes for a single UAV involving multiple operators through managing the macro-task of *trajectory management*, and Fuchs *et al.* investigated the effects of modifying UAV's functional information based on the operator's preferences.

The research presented in this thesis ultimately combined these three areas of studies and the theoretical contribution is discussed in Chapter 3.

2.5 Conclusion

This chapter presented an in depth review of the theoretical knowledge that contributed to this research; from the autonomy framework used to described the UAVs (Section 2.1), to the human cognitive constructs (Section 2.3) that are a direct factor influenced by their interfaces and interactions (Section 2.2).

Research Question 1 is associated with the proposal of a new functional capability framework to describe UAVs' functional subsystems. The derivation of the proposed framework invoked HACT [28] to describe the UAV's functional branches and EID's abstraction hierarchy to describe the layers of information abstraction [182, 183]. The framework's feasibility was evaluated based on the human operator's cognitive workload and situational awareness. CW was measured using NASA-TLX [97] and SA was measured using SAGAT [76].

Research Question 2 is associated with the aim to understand the effects of relaying UAV's flight path replanning autonomy through the visualisation of alternate trajectories, this in turn reflects the UAV's autonomy capability to the human operator. The proposed experiment to evaluate the impacts incorporated the functional capability framework established in Research Question 1, with the UAVs' autonomy configurations designed using autonomy spectrum [128, 53], and tested using the cognitive constructs of CW (measured using NASA-TLX [97]) and SA (measured using SAGAT [76]).

Research Question 3 is associated with the aim to investigate the effects of autonomy transparency on human performances. The proposed method to increase autonomy transparency draws on the concept of adaptive automation [148, 150, 171] and mixed-initiative dialogues [154, 46, 6]. The constructs used to evaluate its impact on human operators are CW (measured using NASA-TLX [97]), SA (measured using SAGAT [75, 76]) and Trust (measured using Master el al.'s questionnaire [140]).

Evidently, this literature review had thoroughly reviewed all aspects of literature that supported the motivation and the methodology necessary for the completion of this research.

CHAPTER 3

Theoretical Foundation

This research incorporated three primary areas of study and investigated on the effective collaboration of the human operators and the Unmanned Aerial Vehicle (UAV) agents in a *Hybrid System*[1] (Figure 3.1). The hybrid system involves multiple UAV agents operating in close proximity to be managed by a human operator. Increasing the UAV's information transparency is achieved from two perspectives proposed in this thesis: the visual representation of of the UAV's functional capability throught a framework, and the text-base dialogue system used to exchange UAV's autonomy capability.

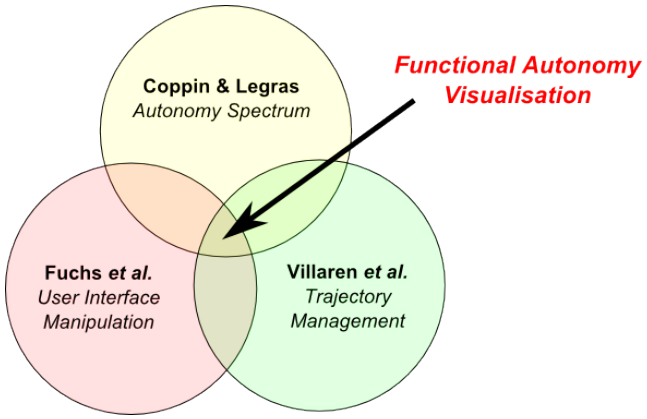


Figure 3.1: Overlap of existing research in multiple heterogeneous UAV management

3.1 Capability Transparency

Capability in this thesis encompasses functional capability and autonomy capability. The functional capability refers to physical systems and subsystems of the UAV (agents), its internal states, and functional information. The autonomy capability refers to the agent's ability to perform different tasks autonomously in the information gathering and decision processes; this includes the agent's intentions (goals) and desires (plans). However, the effective management of multiple agents requires the human operator to have an internal operative state of the agents. Hence, capability transparency must be achieved.

[1] A *Hybrid System* in this research is defined as a system that includes a human entity (human operator), multiple agents (UAVs), and the interfacing agent (an implementation of the *PAC* model)

The transparency of these capabilities are presented in two ways: The amount and the depth of the agents' abstract functional subsystem capability information (discussed in Section 3.2), and the agents' autonomous capabilities to regenerate alternate objectives when confronted with unexpected events (discussed in Section 3.3), displayed at a user interface level [42].

The implementation and effects of the capability transparency on a human operator's cognitive performance was evaluated in a series of three experiments involving a human operator managing four simulated UAVs to complete search tasks. The human operator was needed to interact with a software system (Figures 3.2 and 3.3) which included two components: The environment, and the User Interface.

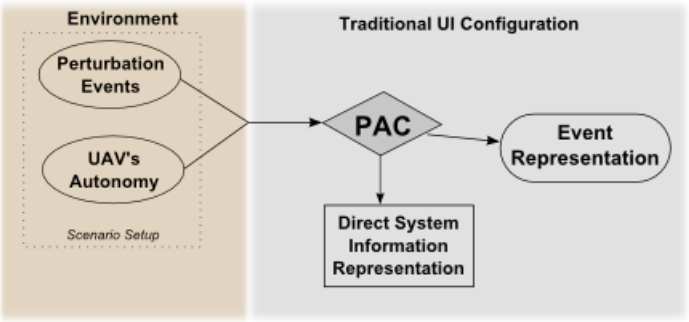


Figure 3.2: Autonomy visualisation model for lower levels of functional autonomy capabilities

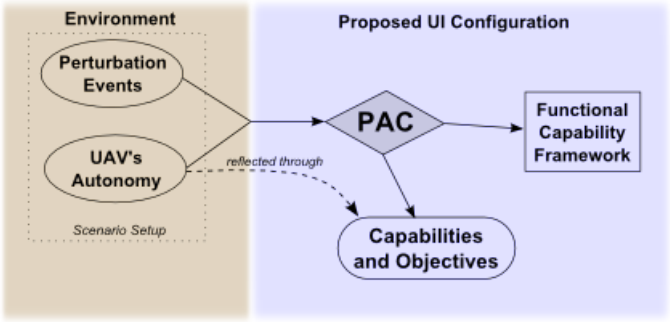


Figure 3.3: Autonomy visualisation model for high levels of functional autonomy capabilities

3.1.1 Environment Grouping

The environmental grouping in this research defined the elements that were present in the "uncontrolled" environment. This was simulated in the software prototype. The two elements in this group are *Perturbation Events* and the UAV's *Functional LOA (F-LOA)*. The *Perturbation Events* in this research were described as the unexpected events, interruptions or obstacles that altered the pre-planned/original flight profiles of the agents. There were different types of events which influenced a agent's functional subsystems separately or together, causing a malfunction of the vehicle and thus reducing its autonomy or work

capability. The UAV's *F-LOA* described as the amount of autonomy present for each of the UAV's functional subsystems. Autonomy levels apply to each functional subsystem separately, meaning a combination of LOAs coexisted in one agent, and consequently in a Multi-Agent System (MAS), there were agents with autonomy heterogeneity.

3.1.2 User Interface Grouping

The UI was designed based on the traditional *Presentation-Abstraction-Control* (*PAC*) model as discussed in Section 2.2.1 (Figure 3.2). The *Presentation* component displayed all of the information of the experiment in its most basic form, as well as supported interactions with the human's hand gestures and the system. The *Abstraction* component contained the UAV attributes such as the autonomy capabilities and functional subsystems. The *Control* component mediated all the gesture inputs and software outputs from the *Presentation* to the *Abstraction* component.

Research Contributions 1 and 2 aim to transform the traditional implementation of UAV functional information into one that is able to promote capability transparency in a hybrid system (Figure 3.3) with the following proposals:

- Modified *Presentation* component: Proposed to reconfigure the UI display to incorporate capability information, this includes functional and autonomy capabilities, as well as its physical interaction capabilities
- Modified *Abstraction* component: Proposed to arrange the UAVs' functional subsystems into the Functional Capability Framework (FCF), and organise the autonomy status to be communicated in the *Presentation* component, and
- Modified *Controller* component: Proposed to augment the interactions between FCF and autonomy information

Two configurations of interfaces are presented: A traditional *PAC* configuration and a novel *PAC* configuration to support capability transparency. The traditional *PAC* configuration is represented in a modern-day Ground Control Station (GCS) [219] where the human operators receives no indications to suggest the UAV capabilities as illustrated in Figure 3.2. This configuration limits the visual display of the environmental situations (such as any unexpected obstacles) that the UAVs are exposed to, and are only reflected through indications and situation behavioural trends to the operator. Based on these information types, the human operators experience a higher level of perceived workload in interpreting the UAVs' autonomy status. Given the inverse relationship of SA and CW, the level of operator SA of the UAVs' overall status was lower [116], which was a challenge in any multiple UAV management scenario. The proposed *PAC* configuration represents UAV capabilities in a form of direct notification as illustrated in Figure 3.3. This configuration adds two key visual components: A UAV's functional capability framework in a layered/tag-centred form, and a UAV's autonomy capability visualisation. These components were

studied independently to understand the effects of UAV capability information on human operator's cognitive performances.

Between these two arrangements, the UAV's FCF is a common element. This framework (further discussions are presented in Section 3.2) consisted of a classification of a UAV's functional subsystems according to the subsystem roles, and it further divided these classifications into three layers of abstraction. During mission design and scenario trial, the Level Of Details (LOD) of the UAVs could be changed by the experiment participant depending on the functional autonomy of the UAV system or the operator's preference.

3.2 Functional Capability Framework

With the prevailing notion to invert the paradigm from multiple human operators operating a single UAV to a single human operator managing multiple UAVs [113, 50**?** , 127], Fuchs *et al.* had investigated the effects of manipulating the visualisation of the UAV functional information on a pre-existing GCS platform to accommodate for multiple heterogeneous UAV management [86]. In their study, a survey (that was similar to the process of Information Abstraction (IA) initially proposed by Chen *et al.* [40, 41]) was conducted with several certified UAV operators to abstract functional information about their UAVs in varying scenarios.

This research proposes an FCF which aims to deliver the UAVs' functional subsystem information in a structure based on the amount of abstraction. The specific functional level of abstraction (or detail) about any of the UAV's subsystems is referenced and accessed during mission planning or operational stages. It also provides a unified and adjustable foundation for the software visualisation implementation of a UAV's functional subsystem capabilities in a hybrid system.

Two primary requirements were necessary to provide a practical architecture for the development of the experiment software prototypes (discussed in Chapters 5, 6 and 7) used throughout this research, and to derive this framework: The structure of the functional subsystem abstractions, and the method of layer indexing; where they were used to validate the impact of this framework on operator CW and SA.

3.2.1 Requirement 1: Functional Subsystem Abstraction

The first requirement was to obtain a breakdown of a UAV's functional subsystems. The breakdown was classified based on the volume of information through information abstraction. Figure 3.4 illustrates an example of a functional framework, where it was interpreted in two dimensions; a horizontal division of the UAV's functional subsystems and a vertical division in a level of information abstraction.

The horizontal classification was presented in a branch format where the UAV's collection(s) of functional subsystems that were related to each specific branch was abstracted. Each branch was then further divided into multiple layers based on each subsystems' layers

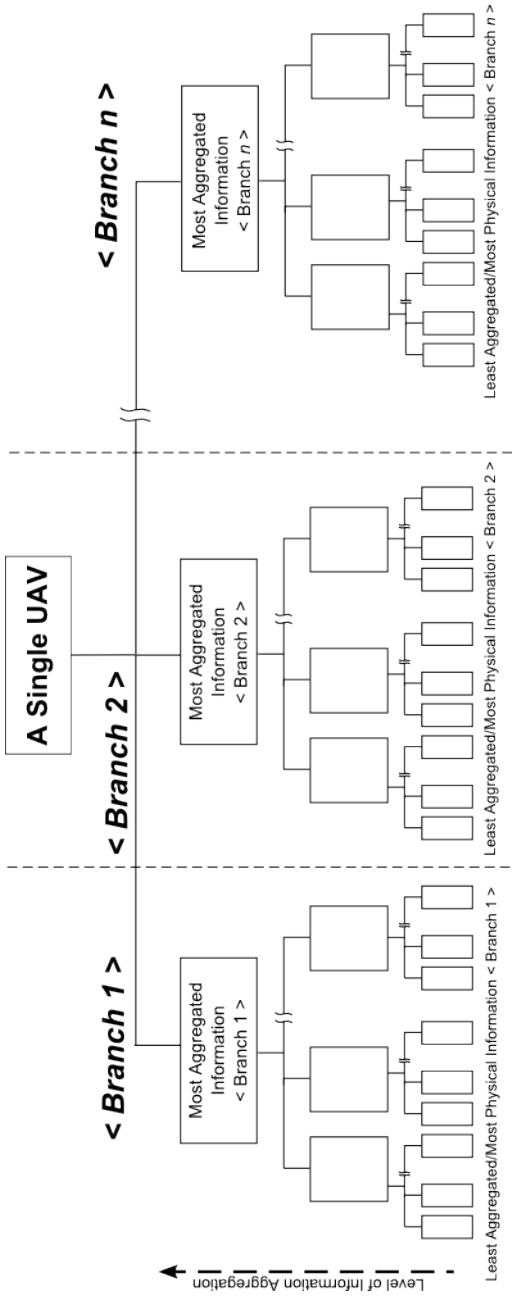


Figure 3.4: A functional framework structure wireframe illustrating the two dimensions of classification. Horizontal: Abstracted in functional subsystems, Vertical: Abstracted in aggregation of information;

of abstraction. The amount of abstraction layers remained the same across the different branches. Therefore, each functional subsystem can be referenced with a consistant indexing method.

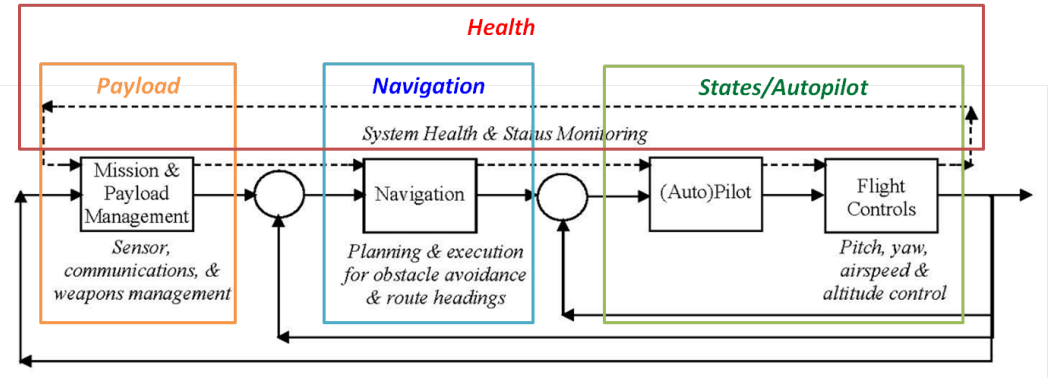


Figure 3.5: Four Functional Subsystem Branches mapped to the Automation Architecture for Single Operator Multiple UAV Command and Control

The functional subsystem classification was derived from the automation architecture for a single operator to command and control multiple UAVs initially proposed by Cummings' team [56, 28]. The control loops in this architecture, illustrated in Figure 3.5, were then extracted to form four different functional branches of a UAV: Health, Navigation, States/Autopilot, and Payload. These branches classify the division of the UAV's functional subsystem (i.e. the horizontal division in the FCF). The information's levels of abstraction forms the vertical division.

The UAV's level of information abstraction was derived from the five levels of abstraction: The functional purpose layer, the abstract function layer, the generalised functions layer, the physical functions layer and the physical form layer, proposed by Rasmussen [184].

- Functional purpose layer: This layer describes the functionality of a certain subsystem and its purpose in the open world
- Abstract function layer: This layer symbolises the functional purpose of the particular subsystem
- Generalised function layer: This layer describes the general process of a particular subsystem in which it presents the subsystem's behavioural structures
- Physical function layer: This layer describes the specifics of the systems and interacting subsystems such as the electrical, fuel, hydraulics, avionics etc., and
- Physical form layer: This layer describes the physical state of the components and subsystems - the state where the subsystem interfaces with the external environment

The levels of information abstraction in this research was formulated by simplifying the five levels into three, which, in the context of this research, was sufficient in terms of the abstraction of information display consideration. Table 3.1 presents the level mapping of

the LOD and Rassmussen's five levels of abstraction [182, 224]. The functional purpose layer and the physical form layer was omitted. The reason for this was that; at the highest of abstraction, the functional subsystem was defined and was not required to be visually presentable, and the physical form was the physical asset, which was not applicable in the scope of this research.

Table 3.1: Equivalent mapping of the five levels of abstraction proposed by Rasmussen [182] to the three levels of information abstraction proposed in this research

Levels of Abstraction	Information Abstraction
Functional purpose layer	-
Abstract function layer	High LOD/Abstraction
Generalised function layer	Medium LOD/Abstraction
Physical function layer	Low LOD/Abstraction
Physical form layer	-

High LOD or high level of information abstraction was proposed to encapsulate the highly aggregated functional information about a UAV, i.e. the overall health status, flight path, origin point, destination point, waypoints, UAV altitude, speed etc. The aim of this level of abstraction was to provide the most fundamental information about the UAV's status and performance, and to reduce information complexity and workload. The typical high-level information that could be expected to be used as a basis point for any UAV is presented in Table 3.2. The functional branches and its associated generic High LOD components are listed. This method could be applied as presented in the examples in Sections 3.2.3 and 3.2.4.

Branch	Component
Health	Overall Health Status
Navigation	Flight path
	Origin waypoint
	Destination waypoint
	Nav. waypoints
States/Autopilot	UAV Altitude
	UAV Speed
	UAV Attitude
	UAV Heading/Direction
Payload	Communications
	Purpose system (i.e. Cameras)

Table 3.2: Functional branch to fundamental High LOD components of a UAV

Med LOD or medium level of information abstraction required an extension to the existing fundamental functional components and they are platform dependent. As presented in the examples in Sections 3.2.3 and 3.2.4, different subsystem elements were designed for the Med LOD layer.

Low LOD or low level of information abstraction required complete or near complete representation of all the subsystem component information in the least aggregated, while humanly comprehensible form. For example, fuel level could be represented in units of Litres or percentage, communication signals could be represented in strength and integrity of the signals in percentage.

The following subsections present examples of the process of deriving the FCF for a real world B-HUNTER UAV and a hypothetical UAV that was used for the purpose of this research.

3.2.2 Requirement 2: Level Of Detail Indexing Method

The *LOD* is an indexing method that is proposed to reference the abstraction level of subsystem information [40, 41]. A *high* LOD denotes a high level of abstraction, which means that high-level information about the subsystems is available (and subsequently, raw and unprocessed subsystem information, such as that appear in *physical form* or *physical functional layer* [182] is hidden). This is not to be confused with the amount of detail (or information) presented. Rather, it denotes the level of information abstraction. A *low* LOD is the contrary; it denotes a low level of abstraction, which means that a large amount of raw and non-aggregated subsystem information is presented, and highly abstracted information is hidden.

Figure 3.6 illustrates an example of the health subsystem of a hypothetical UAV and its LOD breakdown. At the highest LOD, information about the overall health status of the UAV that are highly abstracted is presented using a coloured pie chart, where *Green* denotes *good* health and *blinking red* denotes *bad*. The transition between green to red is gradual, where the colours fade from bright green to solid red as the UAV health level reduces. Finally, a solid red that pulses on and off denotes that the health of the UAV has reached a critical level.

The next LOA, *med*, presents a lower level of information abstraction; it presents a generalised and graphical description of the subsystem information. Figure 3.6 illustrates an example that presents only three subsystems. This level includes the health status of each of the subsystems at an aggregated level. Colour coding and status bars present the detailed information.

The final LOD, *low*, presented low/sensory level subsystem information about the UAV. A very low level of abstraction of raw information was available in this level, revealing all the sensory data about every aspect of the UAV. This level presents no aggregated information.

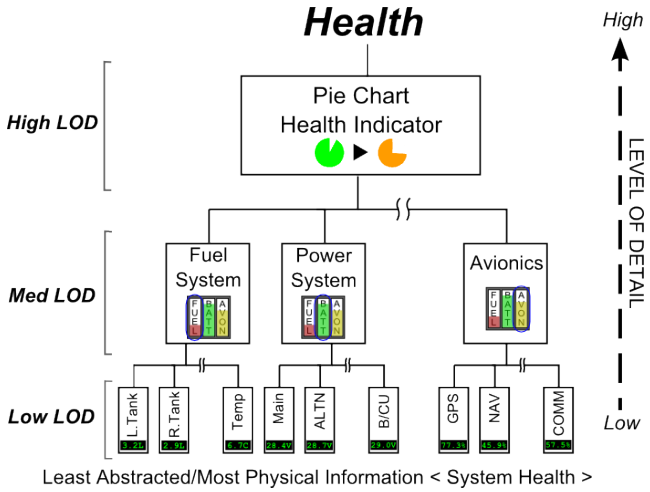


Figure 3.6: An example branch of an FCF of a hypothetical UAV's health monitoring subsystem with an LOD index illustration

Figure 3.7 illustrates an example of a hybrid LOD configuration for two UAV during mission planning stage. The four categories spanned horizontally are the functional capability categories. The three vertical layers represent the LOD (or amount of information abstraction from a high level/minimal information to a low level/maximal information). The labelled and coloured nodes represent the visual configuration of the UAVs' functional capability information. In this example, UAV_1 was configured with a medium LOD for both the *Health* and *Payload* functional categories, a high LOD for its *Navigation* category and a low LOD for its *State* category. Similarly, UAV_2 was configured with a low LOD in both *Health* and *Navigation* category, a medium LOD for its *State* category, and a high LOD for its *Payload* category.

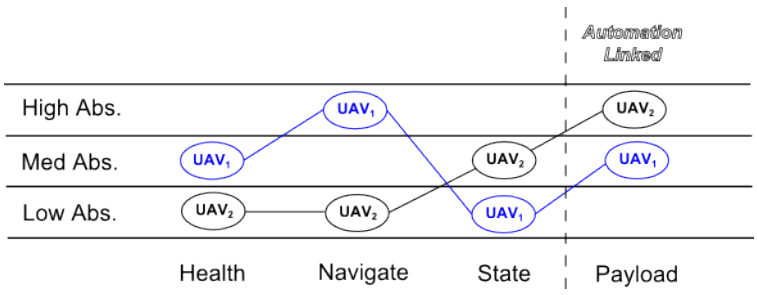


Figure 3.7: An example of the Level Of Detail (LOD) in a hybrid configuration for two UAVs across the four functional capability categories. Here, the term Abstraction (Abs.) is used

3.2.3 Example 1: B-HUNTER UAV

The B-HUNTER UAV, initially designed and developed by TRW Inc. (currently acquired by Northrop Grumman) for the United States Army as a short range tactical UAV. This model was a derivation of the Hunter UAV designed in a joint venture between TRW and Israel Aircraft Industries Ltd. (IAI) [65]. The B-HUNTER UAV was purchased by the Belgian

Table 3.3: B-HUNTER UAV Subsystem mapping to the FCF functional branches, nomenclatures in this table are: Digital Central Processing Assembly (DCPA), Air Data Terminal (ADT), and Air Traffic Control (ATC)

Subsystem	Branch	Description
Fuel & Propulsion	System Health	Fuel level & propulsion capability
Electrical Power	System Health	Electrical & power system integrity
DCPA	States/Navigation	UAV's onboard avionics systems
Sensors	States/Payload	Help acquire UAV's states & payload
Electromechanical	System Health	Actuators & controls health
ADT	Payload	Payload communication systems
ATC Transponders	System Health	UAV communication systems
Flight Termination	Navigation	Emergency flight profiles

Army Ground Forces, primarily used for information gathering and reconnaissance.

The B-HUNTER UAV's is a twin two-cylinder piston engined, fix-winged UAV which has a wingspan of 8.9 metres, a length of 6.95 metres and a weight of 725 kilograms. It has a cruise speed of 196 kilometres per hour or loitered at 165 kilometres per hour with a service ceiling of 4570 metres and an endurance of 12 hours [114].

Figure 3.8 presents the major system components of the B-HUNTER system, which contains Air Vehicle (AV), Ground Support Equipment (GSE), Ground Data Terminal (GDT), GCS, Portable Ruggedized Control Station (PRCS) and take-off/landing equipment (RAPS & RATO).

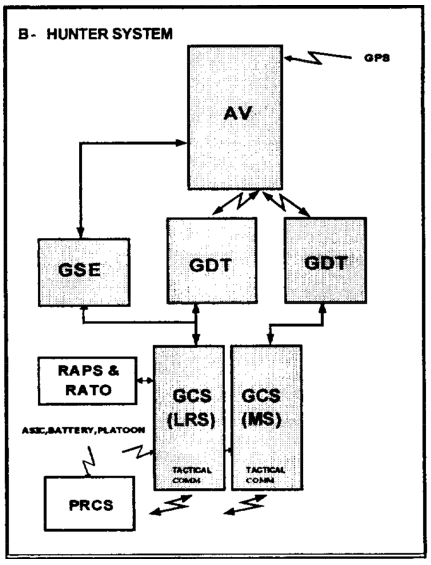


Figure 3.8: B-HUNTER System overviews and interfaces (Published in [114])

The derivation of the FCF for the B-HUNTER system focused primarily on the AV system. Table 3.3 presents the major subsystem and the FCF branch to which they belong.

Table 3.3 provides a summary of the UAV subsystem components at a high level, and the combined/integrated data information helps to form the High LOD layer. The Med

LOD layer is also formed from the status level of each subsystem individually.

Each of the major subsystem components of the B-HUNTER UAV contains a number of functional elements. Table 3.4 presents a summary of these subsystems and their unit of measure.

Table 3.4: Functional element layout of the B-HUNTER UAV's subsystems, nomenclatures used in this table are: Engine Temperature Sensor (ETS), Fuel Level System (FLS), Avionics Computer (AVC), Air Data Unit (ADT), Vertical Gyro Unit (VGU), Global Position Sensor (GPS), Nose Wheel Steering (NWS), and Identification of Friend or Foe (IFF).

Subsystem	Functional Element	Measures
Fuel & Propulsion	Front Engine (with ETS)	Engine Status
	Rear Engine (with ETS)	Engine Status
	L. Fuel (with FLS)	Quantity
	C. Fuel (with FLS)	Quantity
	R. Fuel (with FLS)	Quantity
Electrical Power	Generator	Volts
	E/BATT	Charge
DCPA	AVC 1	Integrity %
	AVC 2	Integrity %
	Communication System	Signal Strength %
Sensors	ADU	Airspeed/Altitude
	VGU	Attitude
	Gyros/Accelerometers	States Backup
	FVU	Heading/Direction
	GPS/Antenna	Signal Strength %
	Emergency Accelerometer	Integrity %
Electromechanical	Aileron/Rudder/Elevator	Servo Position
	Throttles	Servo Position
	Flaps	Servo Position
	NWS	Servo Position
ADT	Control Commands	Signal Strength %
	Payload Comms	Signal Strength %
	Video Streams	Live Footage
	Primary Receiver	Signal Strength %
	Backup Receiver	Signal Strength %
ATC	IFF Transponder	Transponder State

The Low LOD layer of the FCF for the B-HUNTER UAV was further extended from the Med LOD, and is presented in Figure 3.9. The subsequent LOD's included functional subsystem elements were determined by the sensibility, that was to say, the capability of the elements in providing status information was crucial, as this enabled information to then be visualised as a part of the FCF.

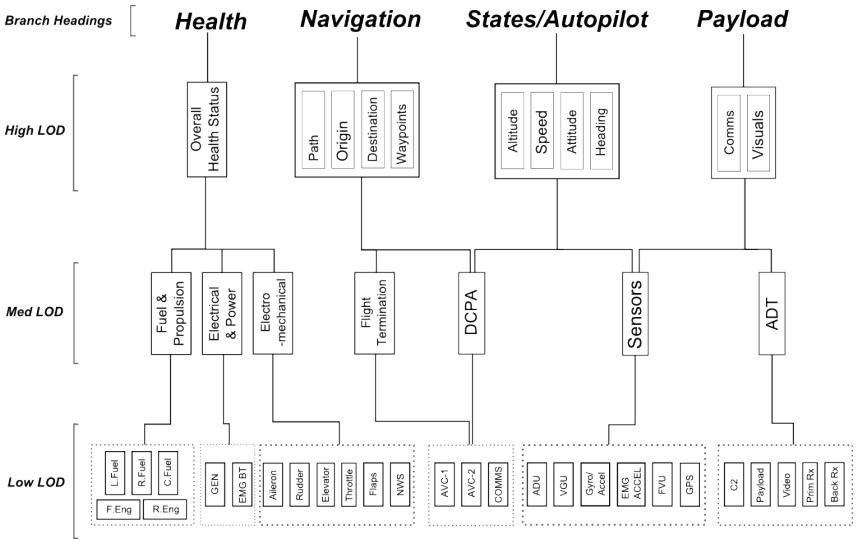


Figure 3.9: Example FCF of a basic B-HUNTER UAV configured as a reconnaissance platform

Given that the general FCF of the B-HUNTER UAV's functional systems components and subsystem elements had been formed, the interface designer could proceed with the visualisation designs of these components - which is beyond the scope of this example.

A limitation in the derivation of this FCF which prevented the illustration of a more functional FCF element, was the limited available literature on the detailed components of the UAV. Hence, only a high level systems understanding could be adapted in the derivation process. In the second example, a less descriptive approach related to the UAV's physical subsystems and more direct functional attributes were included in the derivation process as the design of the UAV components were conducted as a part of this research.

3.2.4 Example 2: A Generic Tactical UAV for this Research

The second example presented in this section details the derivation methodology that is similar to that used to derive the FCF for the B-HUNTER UAV example. This example UAV is a hypothetical, rotary-winged UAV that is similar to the B-HUNTER Tactical UAV system. The generic tactical UAVs necessary for the experiment had three basic functional requirements:

1. Hovering Capability: The UAVs were required to be able to hover over a specific location, and this requirement was satisfied by deploying (implementing) rotary-winged platforms

2. Fast & Agile: The landmass of the test arena was vast, hence, in order to cover such an area, while having the ability to instantaneously change course due to unforeseen reasons, the platform must be able to fly quickly and with agility, and

3. Searching Capability: A primary aim of the experiment mission was to conduct aerial searches of particular zones. Hence, the payload must have a video imaging

capability

In summary, the hypothetical UAV platforms used for this research were rotary-winged, highly agile and had a high operational speed range, as well as simulated onboard video imaging capabilities. Table 3.5 presents the major onboard systems of the test UAV platform, however, its functional capabilities form the elements in the FCF for this UAV as illustrated in Figure 3.10.

Table 3.5: Hypothetical UAV subsystem used for this research, its mapping to the FCF functional branches

Subsystem	Branch	Description
Fuel	System Health	Fuel quantity
Electrical & Avionics	System Health	Electrical and avionics system integrity
Control / Propulsion / Hydraulics	System Health	*Not implemented*
Navigation Computer	Navigation	Plan and execute navigation decisions
Air Data Systems	States	Data to operate aircraft (airspeed, altitude etc.)
Search Cams	Payload	Payload systems

Table 3.6 provides a summary of the hypothetical experiment UAV subsystem components at a high level, and the combined/integrated data information helps to form the High LOD layer. The Med LOD layer is also formed from the status level of each subsystem individually, and in this example, the visualisable elements are illustrated in Figure 3.10.

Table 3.6: Functional element layout of the experiment UAV's subsystems

Subsystem	Functional Element
Fuel	Fuel
Electrical & Avionics	Electrical
Control	*Not implemented*
Propulsion	
Hydraulics	
Navigation Computer	Waypoint (wpt)
	Path
Air Data Systems	Airspeed
	Altitude
	Thrust
Search Cams	Visualisation

Table 3.6 presents a functional breakdown of the visualisable information available for each of the subsystems of the experimental UAV. The subsystems *Control/Propulsion/Hydraulics* were included visually. They were, however, statically (not functionally) implemented. Figure 3.10 presents the FCF of the experiment UAV, including the visualisable

elements of the FCF rather than the physical systems as presented in Section 3.2.3 (due to limited resources on the platform).

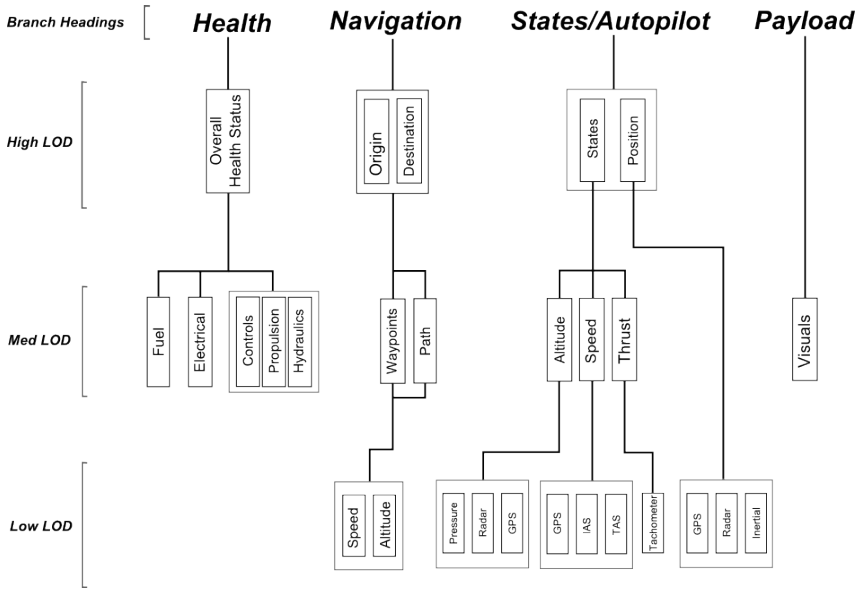


Figure 3.10: Example FCF of a hypothetical experiment UAV configured for the experiment mission requirement

As the FCF of the hypothetical experiment UAV's functional systems components and visualisable components had been derived, the interface specification were designed and implemented for the experiments used in this research. These implementations are presented in Chapters 5, 6 and 7.

3.3 Autonomy Transparency in Hybrid Systems

Many cognitive challenges are associated with the proposed management ratio inversion [43, 42, 127, 83, 113]. They are required to be addressed to enable further growth in the effective management of multiple UAVs. This thesis focuses on increasing the transparency of the UAV's autonomy such that the human operator obtains an understanding of the UAV's autonomy capabilities.

In order for the operator to obtain an understanding of the UAV's autonomy in a hybrid system, the agent's desires and intentions must be communicated. The agent's *desires* denote its current or alternate goals/objectives, whereby the current objective feasibility is being analysed and assessed. If infeasible, alternate objectives must be autonomously formed. From this, the agent's new *intentions* are then autonomously formulated from its desires by selecting and carrying out a methodology, or plan, to reach the *desire*.

Under situations where the original *desire(s)* can no longer be satisfied in situations such as hazard confrontations, alternative desire(s) need(s) to be reached. Subsequently, the agent's *intention(s)* (or the means to reach the *desire*) are in effect. The resulting actions the agents perform are dependent on their autonomy capabilities, and human-

agent collaboration can only be maximised when the human operator has an adequate knowledge of the agent's autonomy [37, 197] in the form of *desires* and *intentions*.

The agent's autonomy transparency can be achieved graphically or textually [198, 165] by communicating its desires and intentions to the human operator. Chapter 6 investigates the impact of autonomy transparency through graphical representation in the UAV's navigation branches which compared an interface configuration that did not present any autonomy status or intentional changes, with an interface configuration where UAV navigation autonomy transparency increased and presented the flight path information visually. Chapter 7 investigates the impact of autonomy transparency on the cognitive and objective performances with the addition of automation trust, through natural language representation in a textual form, incorporating all four branches of the UAVs' FCF (Section 3.2).

3.3.1 Functional Level Of Autonomy

Functional LOA describes a concept where each of the UAV's subsystems contain some amount of autonomy, while the concept of an LOA is used to describe a UAV in its entirety as the importance of classifying the amount of autonomy a system has was initially recognised by Sheridan and Verplanck in their of the SV Scale [211].

Recent studies suggest that autonomy should not wholistically apply to a complex system such as a UAV; the system has different capabilities and functionalities [48], and the autonomy capabilities of these functionalities should also be categorised, hence, the concept of Functional LOA was introduced.

The three levels of functional autonomy are: *High Autonomy (HA), Part Autonomy (PA),* and *Low Autonomy (LA). HA* denotes that no manual input was required; the UAV's functionality was able to make appropriate internal adjustment to perform the required task. *PA* denotes that partial input might be required by the human entity where the autonomous functionality could only make certain decisions based on the scenario. Finally, *LA* denotes that the human agent interaction is required as the autonomous agent's functionality has minimal to no decision-making capability available. These metrics were used in the design and implementation of the experiment scenarios presented in Section 6.1.

3.3.2 Information Transparency

The effective sharing of information between the human and UAV agents in a hybrid system enables effective human-machine interaction. This could be achieved with a mutual awareness of the other agent's status as illustrated by the dual-ended arrow in Figure 3.11.

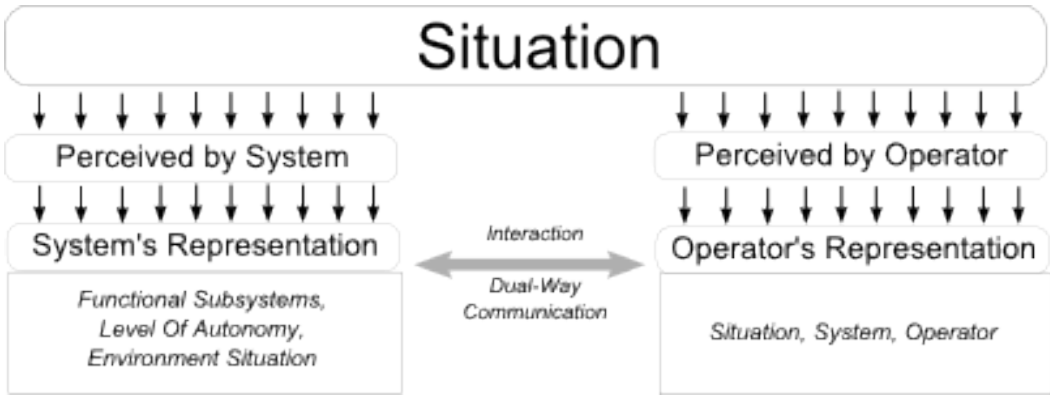


Figure 3.11: Authority sharing concepts in a single robot single operator interaction (Adapted from [197])

One form of UAV information was its autonomy capability, which was recently recognised to possess a tremendous impact on the UAV operator's CW, SA and Trust [149, 37, 38] as robots (UAVs) are no longer considered as just a remote asset in the field. Rather, they are considered as *teammate(s) [177].* To harvest the prime capabilities of both forms of agents to enable superior Human-Robot Interaction (HRI), a bi-directional communication channel throughout the mission to allow the other agents' operating statuses/models to be shared is currently a limitation of modern-day HRI systems[165].

3.3.3 Autonomy Spectrum

The autonomy spectrum, initially proposed by Legras and Coppin [51], extends the autonomy profile of an autonomous system to human-centered adaptive systems. This enabled multiple modes of operations within any adaptive system, such as a UAV.

A mission could involve multiple modes-of-operation. Specific spectra is to be applied to different operative modes. The spectrum consists of a four-step sequential model of human information processing: Information Acquisition, Information Analysis, Decision Selection, and Action Implementation. This also aligns with Boyd's four-step model of decision making [19]: Observe, Orient, Decide and Act. Each step consists of a linear, ten-level SV scale forming a two-dimensional framework for the UAV's modes-of-operation [128, 127].

Each step of the spectrum is then divided into the ten levels of autonomy, which described the agent's task responsibilities. At LOA 1, the human agent assumed all the responsibility of performing that specific step of the process. At LOA 10, the machine agent assumed all the responsibilities instead. Figure 3.12 presents an example autonomy spectrum and the possible configurations.

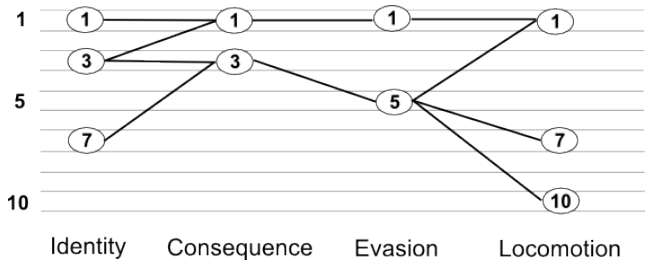


Figure 3.12: Example UAV autonomy spectrum for the hazard avoidance mode of operation

The spectrum presented in Figure 3.12 denotes a hypothetical, self-reliant UAV's hazard avoidance mode-of-operation. This mode consists of the four steps: Identify (Observe), Consequence (Orient), Evasion (Decide), and Locomotion (Act).

Identify denotes the step of hazard identification; the observation of the environment and classification of any incoming hazard. There are three potential autonomy levels for this step: LOA 1, the human agent assumes all of the responsibilities of identifying potential hazards from rudimentary onboard sensors, LOA 3, the human agent assumes some responsibilities in identifying potential hazards, however, the machine agent assists in the identification process, and LOA 7, the machine agent assumes responsibilities of identifying the hazard, however, the human agent also had some responsibilities in confirming the identification.

Consequence denotes the step of assessing the hazard's implications if it confronted the UAV. Two potential autonomy levels were available for this step: LOA 1, the human agent assumed all the responsibilities of identifying the consequences of the hazard to the UAVs, and LOA 3, the human agent assumed some responsibilities in identifying the consequences and the machine agent also assumed some.

Evasion denotes the step of assessing and selecting the appropriate evasive action to the hazard presented. Two potential autonomy levels were available for this step: LOA 1, the human agent assumed all the responsibilities of generating possible evasive profiles and selecting the most appropriate action, and LOA 5, the machine agent assumed the responsibilities of generating possible evasive profiles and the human agent selected the appropriate evasive action.

Locomotion denotes the step of carrying out the selected evasive action to physically evade the hazard presented. Three potential autonomy levels were available for this step: LOA 1, the human agent assumed all the responsibilities of flying and manoeuvring the UAV to follow through the selected action, LOA 7, the machine agent performs the flying operation mostly, while the human agent assumed some responsibilities in supervising the operation, vetoing if necessary, and LOA 10, the machine agent carried out all the operations from supervisory control, to manual operation of the UAV.

The possible arrangements were governed by the UAVs' onboard functional and autonomy capabilities, and the arrangements used in this experiment is presented in chapter 7 Section 7.1. Figure 3.13 presents an example autonomy profile based on the

spectrum presented in Figure 3.12. The blue profile lines define the configuration of the profile.

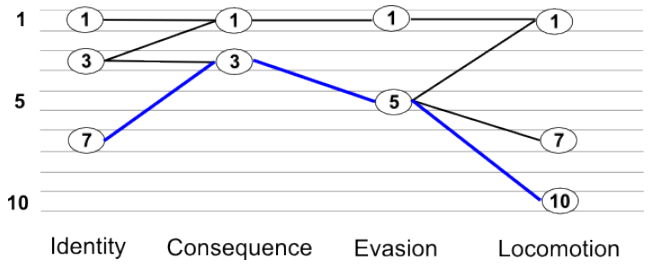


Figure 3.13: Example UAV specific autonomy profile for the spectrum's hazard avoidance mode of operation

In the example, the profile defines the following LOA progression: 7-3-5-10 (denoted by the profile lines), which denotes that the machine agent assumes most of the responsibilities when it came to hazard avoidance. It would first identify potential hazard situations with the human agent's assistance, then determine possible consequences along with the human agent, and decide the appropriate evasive action and carry through the action to physically evade the hazard fully autonomously. However, the autonomy profile may vary throughout the mission due to unforeseen situations and environmental influences. Hence, a way to visualise this responsibility must be determined and implemented to avoid automation surprises.

3.3.4 Model of Autonomy Transparency through Text-Based Representation

Figure 3.3 illustrates the model of autonomy transparency using text-based representation. *V'* presented previously denotes the proposed method of the visual representation of the UAV situation/scenario. Both the UAV's functional LOAs and the scenario situation (perturbation events) that impact the normal operations of the mission are reflected to the human operators as alternate objectives through a text-based dialogue (Section 4.2.3).

3.4 Implementing Capability Transparency

Figures 3.14 and 3.15 present the expanded models of the autonomy transparency configuration for high and low autonomy transparency. This thesis divides the hybrid system into three layers: Mission layer - representing the environment space with a high degree of environmental uncertainty, Visualisation Layer - the display interface where the mission layer was visually represented to the human operator, and Agent layer - denotes the human agent/operator and their performance (perceived cognitive and mission performances).

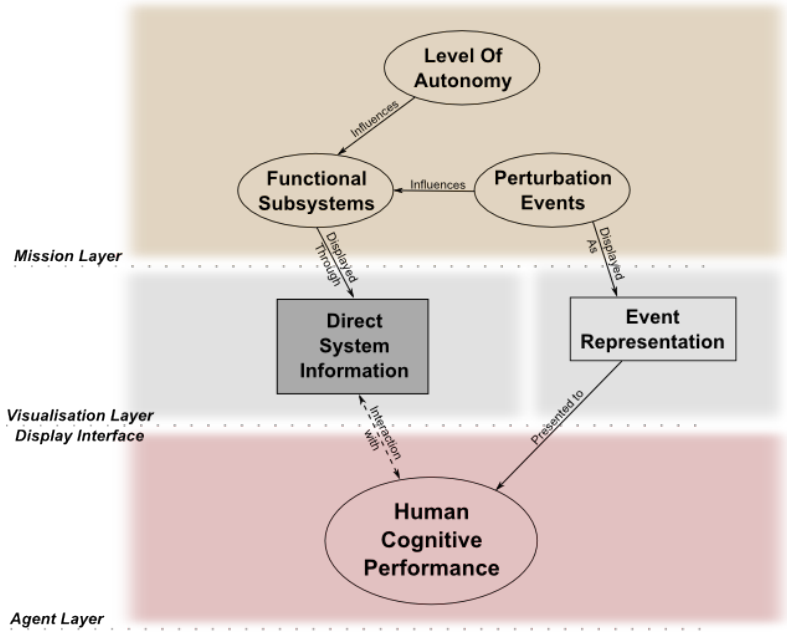


Figure 3.14: An expanded transparency visualisation configuration model as a baseline configuration for comparison

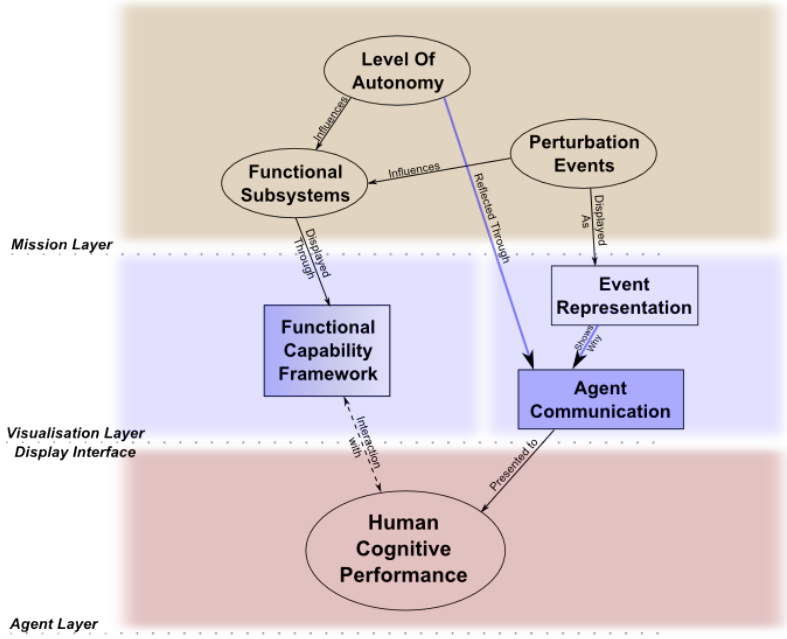


Figure 3.15: An expanded transparency visualisation configuration model for a evaluative configuration for comparison

3.4.1 Mission Layer

The *mission layer* represents the dynamic and uncertain environment that the UAVs operated in. Each UAV had the autonomy requirements and descriptions available to perform tasks [89] (functional *Level Of Autonomy*) and the physical capabilities of the functional subsystems and skillsets (*Functional Capability*) [89, 127] that formed the UAV.

The F-LOA defines the configurations of autonomy (example illustrated in Figure 3.7) for each functional category. This directly moderates the UAV's physical subsystems and components, and thus influences the UAV's functional abilities to perform tasks.

In a dynamic and uncertain environment, unpredictable obstacles or events such as deteriorating weather conditions, hazardous/high-threat areas and malfunctioning subsystems are inevitable [51]. A real world example of the unpredictable events could be seen in the incident involving a United States (U.S.) MQ-1B "Predator" in September 2012 [123], where the UAV experienced a catastrophic loss of power leading to a loss in its satellite datalink and impact with the terrain at approximately 3.25 nautical miles from the initial position where the link was lost. In this example, the UAV was operating in a typical dynamic and uncertain environment when it encountered an unpredicted error and was lost. The total loss was valued at approximately $4.4 million U.S. dollars.

These unpredictable obstacles or events directly influence the capability of the UAVs, and consequently challenge the human operator's cognitive boundaries [51, 127]. This thesis models these *Perturbation Events* as a variable of independence, and the studies are presented.

3.4.2 Visualisation Layer (Display Interface)

The *visualisation interface layer* represents the visual and control interface between the human and the machine agents. It consisted of three components that were integrated together; the functional subsystem visualisation, the functional autonomy visualisation, and the control interface.

The functional subsystem visualisation consists of the visualisation of the UAVs' subsystem information in a layered style where the operator can access different layers of details related to any particular subsystem through the FCF (Section 3.2). Different layers of subsystem details can be accessed using the control interface described in chapter 5.

The functional autonomy visual representation consists of two types: Graphical, and textual. The graphical representation enables a one-way communication channel from the UAV agent to exhibit its autonomy graphically. Experiment 2 explores the effectiveness of the the communicated autonomy information of the UAV's navigation system by visually presenting original and alternate flight paths to the human agents throughout the experiment scenario.

Further to the graphical representation of the UAVs' autonomies in Experiment 2 (Chapter 6) presents the effect of navigation capability's transparency on human operator/teammate's cognitive performance. Brick and Scheutz suggested that text-based natural language was the most effective and direct way to communicate between agents [25]. Hence, the effect of the UAV system's autonomy transparency was expanded by communicating its intentions and capabilities bi-directionally via a text-based dialogue system.

The text-based dialogue system enables a bi-directional communication channel that

was simple. It allowed effective textual exchange between the machine and the human agent. Figure 3.16 illustrates a simple bi-directional dialogue process between the human and the machine agent. In this process, the machine agent initiates a dialogue, sharing its autonomy information and makes requests, the human agent satisfies its requests by acknowledging the message and acting on the request.

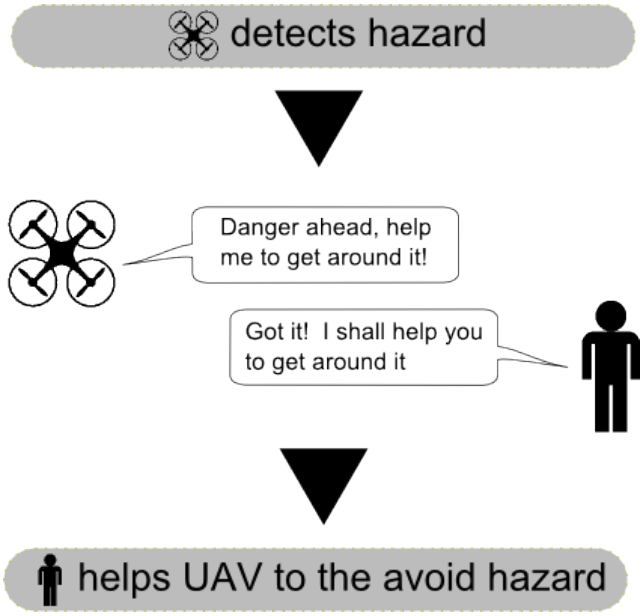


Figure 3.16: Flow illustration of the human and machine agent autonomy information and request dialogue

3.4.3 Agent Layer

The *agent layer* represents the human agent's performance and cognitive attributes. These attributes are used as quantitative measurements to gauge the human operator's performance. Experiment 1 (Chapter 5) evaluates the effect of the implementation of FCF based on the metrics of CW and SA. Experiment 2 (Chapter 6), evaluates the effects of autonomy transparency through graphical representation of the machine agent's alternate plans also based on the cognitive constructs of CW and SA. Experiment 3 (Chapter 7) evaluates the effects of autonomy transparency through text-based dialogue to represent the machine agent's capabilities. The evaluation is based on the cognitive metrics of CW, SA and Trust in automation. Objective metrics, such as initial response time, event response time, and the number of items of interests found during the experiment scenarios are also captured for analysis.

The human agent interacted with the hardware systems denoted by the *visualisation interface layer* based on the visual information presented and responded via the control and interaction-enabled designs, and the cognitive effects were then measured and analysed in the subsequent chapters.

3.5 Conclusion

This chapter presented the theoretical foundation that was used throughout the study. Three main areas include: The FCF and its derivation process, the autonomy capability and transparency model, and the experimental design to implement autonomy capability.

FCF provides the capability to systematically present the machine agents' (UAVs') functional subsystems. It is organised in layers, which is accessible through designated gestures. Each layer contained different levels of details related to each specific subsystem.

The autonomy capability model represented the concept of the visualisation of the UAVs' autonomy capability and consequently the increase in its autonomy transparency. Two models were presented: A generic model of the traditional UI design model with opaque autonomy transparency is presented in Figure 3.2, and an improved *PAC* model to support the communication of autonomy capabilities by increasing transparency to the UAV's autonomy (Figure 3.3).

Finally, the experimental design implementation represented the concept of the experiment software and hardware design used to validate the proposed implementation of the improved *PAC* model through the effects on the human operator's objective and cognitive performances. The next chapters present an in-depth experiment aim, methodology, scenario design, results, analysis and discussions of Experiment 1, 2, and 3.

CHAPTER 4

Experiment Details

The aim of this chapter is to present a detailed presentation description, set up, procedure, and data collection process related to the three experiments conducted through the course of this research. Since the three experiments are similar in their set up and procedures, the general procedure is described. Specific details of each experiment will also be highlighted throughout this chapter.

4.1 Experiment Overview

This section provides an outline of each study. The experiment flow is similar throughout the three studies. However, since each experiment aims to achieve different goals, minor differences were necessary.

4.1.1 Experiment 1

The aim of the first study (presented in Chapter 5) is to validate the implementation of the Functional Capability Framework (FCF) proposed in Chapter 4. The experimental methodology is summarised in 20 steps as illustrated in Figure 4.1.

The first stages of the methodology involved researching and specifying the goal of the study and determining suitable scenario designs to achieve the goal. Once the goal was defined, specification and development of the experiment software followed. Near the completion of the software implementation, experiment administrative tasks such as participant invitation and recruitment were carried out. With the necessary Ethical clearance obtained, the experiment was conducted.

This experiment was broken into three segments:

- Segment A: A scenario where only limited functional information details about a UAV was available to the participant,
- Segment B: A scenario where an adjustable level of functional information details were available, and
- Segment C: A scenario where all Levels of Detail (LODs) were available to the participant.

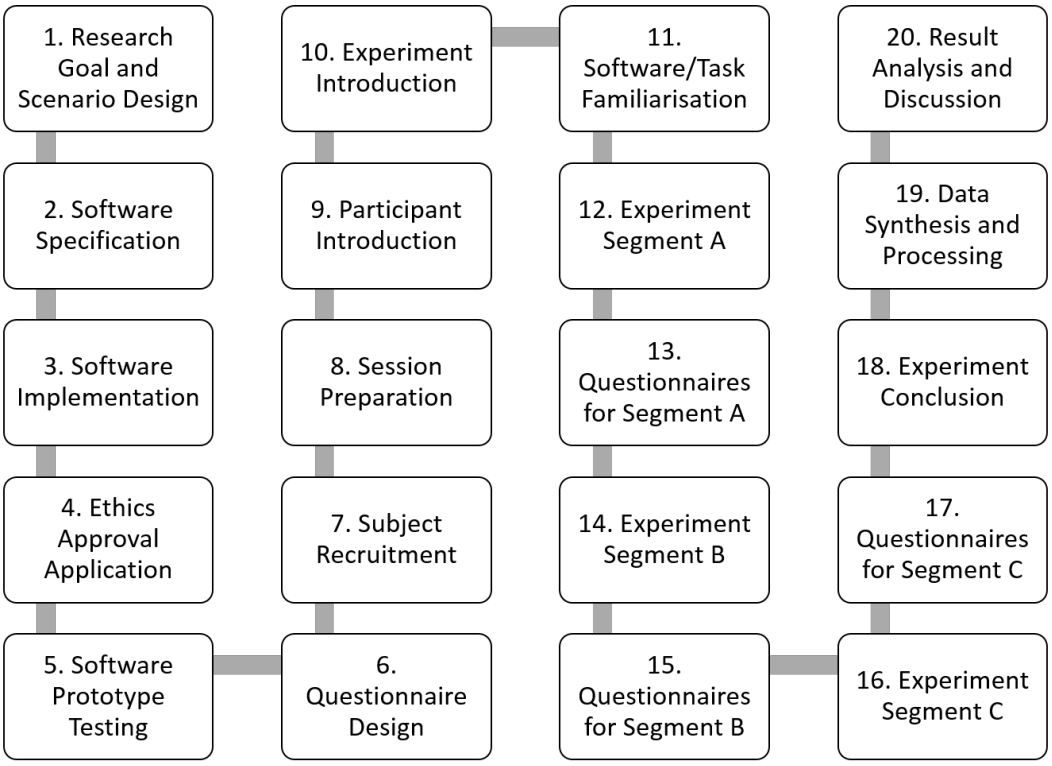


Figure 4.1: Workflow diagram of the experimentation process for Experiment 1 from initial proposal to final execution

The concept of LOD is not to be confused with a UAV's LOA, as the LOD relates to the amount of UAV's subsystem information visualisation, and the LOA relates to its onboard capabilities to perform tasks autonomously.

The aims were validated through capturing and comparing the participants' Cognitive Workload (CW), Situation Awareness (SA), Automation Trust, and objective mission performances.

The CW and SA scores captured in the segments with low LOD and high LOD interface configuration were used as a benchmark for comparison to the scores captured in the segment with a hybrid/adjustable LOD interface configuration.

It was hypothesised that the CW and SA scores captured from the evaluation segments would experience an improvement over both *Segment A* (where limited functional information were available) and *Segment C* (where all the information was available), as adjustable LOD would enable the participants to choose the LOD that they need for the task at hand.

4.1.2 Experiment 2

The aim of the second study (presented in Chapter 6) was to evaluate the influence of autonomy transparency on the operator's cognitive performance through graphical representation of the UAV's autonomous navigation capability. Since the workflow of the second study was very similar to the first, Figure 4.1 illustrates the experimental workflow used in this study. However, the *Baseline* and *Evaluation* scenarios were the equivalent of

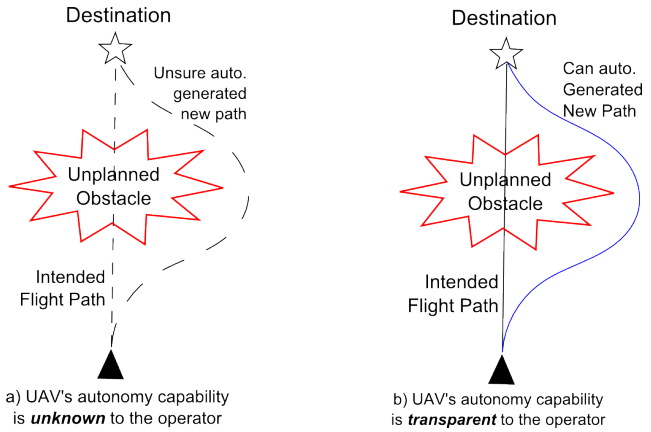


Figure 4.2: A model illustrating the autonomy transparency relationship between the two scenarios; (a) Scenario with opaque transparency, (b) Scenario with transparent autonomy information

Segment A and *B* in Experiment 1.

This experiment was a preliminary study that investigated only one FCF branch - the navigation branch. This branch was investigated through flight plan visualisation, and its influence on the operator's CW and SA during supervisory tasks (i.e. tasks that required long periods of observations and asset monitoring).

Two comparative trials/scenarios were performed: a baseline and an evaluation scenario. The baseline scenario set the benchmark for cognitive performance of the participants with a limited autonomy transparency about the autonomous navigation capability, where the evaluation scenario compared the cognitive performance which was influenced by an increase in the navigation autonomy transparency.

Figure 4.2 presents a model of autonomy transparency through flight path visualisation. Figure 4.2a presents a scenario where a UAV was tracking its predefined course while the flight path was not visually present, and upon confrontation with the unplanned obstacle (represented by the hazardous cloud described in Section 6.1), the operator had no information regarding the navigation autonomy of the UAV, although the UAV was able to generate a new path to avoid the hazard. Figure 4.2b differs in that the UAV's navigation autonomy was transparent through the visualisation of the navigation path. This scenario presented a UAV tracking its predefined path. However, prior to confronting the unplanned obstacle, the UAV was able to generate a new path autonomously, and this information was fed back to the operator, hence the autonomy was *transparent* to the operator.

The hypothesis was that the increased autonomy transparency in the evaluation scenario would result in an increase in the overall cognitive performance of the participants when compared with the baseline scenario, where the autonomy was opaque.

4.1.3 Experiment 3

Experiment 3 (presented in Chapter 7) aims to verify the effect of communicating the UAV agents' functional and autonomy capabilities to the human operators, in order to increase

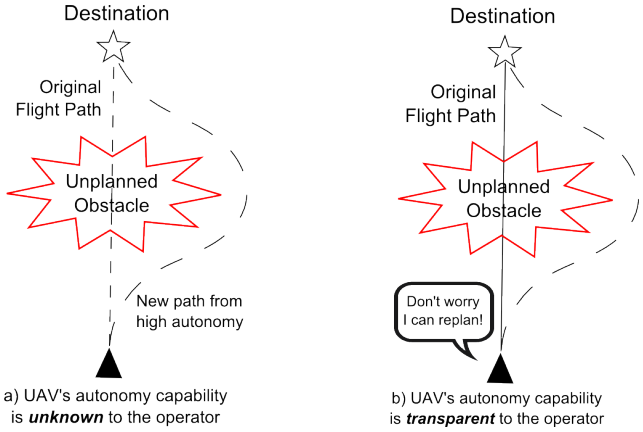


Figure 4.3: A example model illustrating the autonomy transparency relationship between the two scenarios; (a) Scenario with opaque transparency, (b) Scenario with autonomy communicated by the UAV

autonomy transparency. The procedure of this experiment was very similar to Experiment 2 (Section 4.1.2), but the main difference lies in the removal of the post-experiment interview with the addition of an extra familiarisation scenario.

The overall outcomes of Experiment 2 (Chapter 6) indicate that with the increased autonomy transparency in one of the UAV's subsystem capability, there is an improvement in the operator's cognitive performance. This experiment furthered this investigation by implementing autonomy transparency to all the functional capabilities for each *mode-of-operation*.

This chapter applies the concept of autonomy spectrum (as presented in Section 2.1.3 in Chapter 2) to the design of the scenarios. The success of each mode-of-operation was achieved with certain Functional-LOAs (F-LOAs) of the UAVs, hence, effective collaboration was necessary between the human and machine agents.

Two experiment trials were conducted to generate two sets of operator cognitive and mission performance results: a baseline scenario where the UAVs' autonomy capabilities were opaque to the operators, and an evaluation scenario where the autonomy was transparent. The baseline scenario set a benchmark of the operator performances for comparisons with the evaluation scenario. The transparency of the UAVs' autonomy was achieved through the use of a text-based message box, which included the UAVs' status message, an warning type of message where the UAV had a moderate autonomy, input/help was required by the human agent, and an alert type of message where the UAV detected immediate danger and it did not have the appropriate autonomy to mitigate the problem.

Figure 4.3 presents the experiment model for an example of functional autonomy capability and its transparency effect achieved through natural language dialogue. Figure 4.3a illustrates a (baseline) scenario where a UAV was tracking its predefined course while the flight path did not include any further information regarding its autonomy capabilities or intentions. As a result, although the UAV had the autonomy to reroute when confronted

with an unplanned obstacle, its capability was unknown to the operator. Figure 4.3b differed in that the UAV's autonomy was communicated to the operator using natural language, serving as a form of dialogue exchange. In this situation, the UAV was able to *talk* to the operator and explain its capabilities and intentions prior to confronting the unplanned obstacle, in this (evaluation) scenario, the UAV was able to generate a new path autonomously, and this capability and intention was exchanged with the operator, making the autonomy *transparent*.

The experiment hypothesis was that the increased autonomy transparency in the evaluation scenario would result in an increase in the overall cognitive performance of the participants when compared with the baseline scenario, where the autonomy was opaque.

Three types of cognitive and mission performances were investigated. The cognitive performances that were investigated were the Operator CW, SA, and Trust in automation. The mission performances that were investigated was the Initial Reaction Time (IRT), the Event Reaction Time (ERT) and the number of Items Of Interest (IOIs) found.

4.2 Experiment Software Prototype

The experiments used a software prototype designed and implemented by the author. This prototype captured and visually presented the information related to the capability and transparency of a generic UAV to the experiment participants. However, the focus of this software prototype is to understand the comparative effect of information transparency, therefore, less attention was placed on designing the interface by following specific design principles. Furthermore, specific gestures were also implemented to allow different LODs to be accessed during the experiment to accompany the touch-interactive table interaction device used throughout the three experiments.

This prototype followed the *Presentation-Abstraction-Control* (*PAC*) design model, where the *Abstraction* is described by the UAVs' FCF and the scenario configurations, the *Presentation* is described by the visual representation of components of the FCF and the in-prototype environment (i.e. the testing arena), and the *Control* is implemented by the DTIvy detection engine (described in Section 4.2.1) and the interaction specification (Section 4.2.1).

This section focuses on the design and the implementation of the software prototype used in the experiment, the hardware used, and the setup of the testing environment.

4.2.1 Software System Design

The software prototype was designed for use on a touch interactive tabletop device. This device enabled the participants to have more intuitive and immersive command and control experience. The architecture of the design is illustrated in Figure 4.4 .

The software prototype was implemented with the adaptation of two frameworks as well as a gesture detection engine specifically written for the DT104 MultiUser MultiTouch

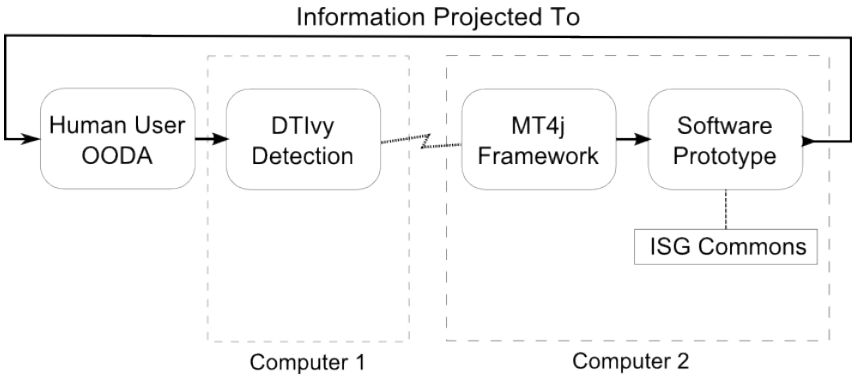


Figure 4.4: Process architecture of the software experiment prototype

interactive tabletop.

- The ISGCommons Framework - developed in the Interactive Systems Group (ISG) at the Australian Research Centre of Aerospace Automation (ARCAA), QUT, provides a suite of comprehensive interfaces and classes that can be implemented in the experiment prototype to simulate a UAV search mission.
- The Multi-Touch for Java (MT4j) Framework - developed in the University of Stuttgart, Germany, provides a comprehensive set of capabilities to enable multi-user, multi-touch capabilities to any Java project.
- DTIvy Gesture Detection Engine - developed in Télécom Bretagne, France, interfaces the hardware DT104 with a software host. Any touch signals detected by the DT104 is transmitted via a specific Internet Protocol (IP) and port number and processed by DTIvy.

The process loop involving the human participants and the software system is depicted in Figure 4.4. Information from the prototype was displayed through a projector onto the interactive tabletop which the participant perceives, comprehends, and makes a decision on the necessary action to take physically by applying fingered gestures to the touch surface. The signals of the human participant's physical touch, detected by the touch table, are then processed by DTIvy and transmitted to the software prototype via a remote connection. The software prototype (implemented through the use of ISGCommons) receives the information sent across by DTIvy, and interprets them using MT4j. Once the hand/finger gestures are received and processed by DTIvy, the input intention is transferred to the software prototype. The subsequent outcome will be reflected graphically and displayed to the participant through a projector.

The experiment scenarios (with the file extension **.scn*, e.g. *experiment_1.scn*) were described using an external/custom-designed file to enable robust and rapid scenario prototyping. The scenario file is a configuration file that includes a general header section (Algorithm 4.1), a UAV performance attribute section (Algorithm 4.2), a scenario event

description section (Algorithm 4.3), and a body which describes the UAV's flight paths (Algorithm 4.2).

The general header describes the general settings required by the prototype to function correctly. The UAV performance attribute describes the UAV's expected performance parameters. The configuration section, which defined the hazard scenarios; and a UAV flight path section, which defines the flight path of the UAVs by describing its waypoint coordinates, expected altitudes and speeds over that waypoint.

Algorithm 4.1 General attributes in the header section of the scenario description file; where *[...]* denotes more entries were inserted

```
1  <HEADER>
2    <GENERAL>                      ; General environment setup variables
3      ivyDomain=192.168.56:3456           ; The DTIvy gesture engine domain
           address
4      tilesPath=mapTiles/zoom14/            ; Path for caching the map
           tiles
5      base_pos=44.443398:6.256947:767, [...]       ; Base location in DD
           coordinates
6      target_pos=44.478298:6.259112:8:4:400:1:1, [...] ; Center of target:
           Number of people:UAV:Segment
7      terr_msa=450:1580               ; Terrain MSA
8      refuel=44.449715:6.280581           ; position of the refueling
           station
9    </GENERAL>
```

Algorithm 4.2 UAV Performance attributes in the header section of the scenario description file

```
1    <PERF>           ; Rotary UAV configuration variables
2      fuel_perc=5:100   ; Fuel quantity in percentage (min:max)
3      power=-1:-1     ; Power of the UAV, minimum:maximum, (-1) indicating
           n/a
4      altitude=+10:2000 ; (+X) indicates X m above the MSA or treetops
5      speed=0:300     ; Minimum and maximum operational speed in KPH
6      speed0=0       ; The speed at which the UAV is idling
7      vert_rate=1     ; Climb:Decent (mpm)
8      turn_rate=10     ; Turning rate in seconds for maximum radius (637m),
           2 min turns
9      accel_rate=0.5:0.5   ; acceleration:deceleration mp(s^2)
10     fuelRate=3        ; Fuel burning rate in percent/min
11   </PERF>
```

Two versions of this file were implemented. The first version was as presented above, the second version was proposed during Experiment 2. Two improvements were made to the original **.scn* file: 1) The F-LOA and LOD configurations were extracted from the main body of the code, and placed in the BODY section of the *scn* file, 2) The UAVs can autonomously reroute during the experiment to a preplanned path. These two improvements enabled the dynamic programmability of F-LOAs and LODs, as well as the dynamic programmability of scripted hazard mitigation plans. They were also used in Experiment 2 and 3.

Algorithm 4.3 Scenario event attributes in the header section of the scenario description file; where *[...]* denotes more entries were inserted

```
1   <CONF>
2     fuel=80:30:74:87                    ; Fuel level (%), for each UAV
          separated by ':'
3     hazard=44.468212:6.267498:100:60:-17:3:1:30, [...] ; Lat:Lon:Width:
          Length:Angle:UAV:Seg to show:% of UAV being lost
4   </CONF>
5 </HEADER>
```

Algorithm 4.4 UAV waypoint/flight path attributes in the body section of the scenario description file - Example illustrates UAV 1 and its partial waypoints, *[...]* denotes more entries were inserted

```
1 <BODY>
2   <UAV_1>
3     waypt_1=44.443398:6.256947:767:250:-:-:-:-:1
4     waypt_2=44.478298:6.259112:900:250:-:-:-:-:1
5     waypt_[...]=[...]
6     waypt_8=44.443294:6.254144:1100:220:-:-:-:-:H
7   </UAV_1>
8   [...]
9 </BODY>
```

The first improvement was implemented realising the LOAs and LODs need to be applied to each functional capabilities of each UAV separately. A section was added to the UAV subsection of the <BODY> section (Algorithm 4.5).

The initial line `loa=` denotes that the following attributes set the current UAV's LOA according to their functional subsystem. The subsequent `ha:ha:pa:ha` denotes the autonomy levels for the four functional subsystems: Health Feedback, Navigation Capability, State/Autopilot, and Payload Capability respectively. In the example presented in Algorithm 4.5, UAV_1 had a High Autonomy (HA) for its health monitoring and feedback system, HA for its navigation capability, Partial Autonomy (PA) for its auto-flight capability, and an HA for its payload/searching capability.

The following line `lod=` denotes that the following attributes set the LODs for each functional subsystem at scenario launch. The LODs could be changed by the participant during the trial to a setting which they felt comfortable with. The subsequent `hl:hl:ml:ll` denoted the detail levels for the same four functional subsystems respectively. In the example represented by Algorithm 4.5, UAV_1 had a High LOD (HL) for its health system, HL also for its flight path visualisation, Medium Level (ML) for its state sensory display and a Low Level (LL) information display for its payload/searching system.

This format was consistent for the remaining three UAVs where the detailed configuration is presented in Section 6.1.

The second improvement was the ability to dynamically define mitigation plans and

Algorithm 4.5 UAV F-LOA and LOD information added before the waypoint/flight path attributes in the body section of the scenario description file - Example illustrates UAV 1's LOA, LOD and its waypoints, *[...]* denotes more entries were inserted

```
<BODY>
  <UAV_1>
    loa=ha:ha:pa:ha
  lod=hl:hl:ml:ll
  waypt_1=44.451615:6.265188:1000:330:1
  waypt_2=44.472111:6.343017:900:330:1
  waypt_3=44.450622:6.265269:800:330:1
  </UAV_1>
  [...]
</BODY>
```

Algorithm 4.6 New REROUTE section of the scenario file presents the hazard mitigation condition and strategy. Code showed snippets of UAV_4's mitigation plan for the baseline scenario, where *[...]* denotes more entries were inserted

```
<REROUTE>
  <UAV_4>
  cond=cloud+tms1000      ; Condition dependent on cloud.  When
     cloud is present, +1ms delay shows the new path
  waypt_1=44.447055:6.286139:900:250:1
  waypt_2=44.448830:6.287148:900:250:1
  [...]
  waypt_8=44.444013:6.266895:800:180:1
  </UAV_4>
</REROUTE>
```

strategies for UAVs that were scripted to expect perturbation events. A new section <REROUTE> was added below the <BODY> section to script this behaviour. Algorithm 4.6 presents an example of a hazard mitigation strategy for UAV_4 in the baseline scenario of the second experiment.

The line `cond=cloud+tms1000` directly under <UAV_4> denotes the condition where this mitigation strategy is triggered. `cond=` denotes that the subsequent attributes describe the triggering conditions, while `'cloud'` describes the event that this UAV was expected to confront. Similarly, a low-fuel level event could also be triggered using the keyword `'fuel'` as seen in UAV_3's mitigation strategy (`'cond=fuel+lim25'`).

The addition sign (`'+'`) denotes the next attribute which sets the trigger criterion for which the new flight path becomes active. `'tms1000'` denoted that the UAV's new path was programmed to become active approximately 1000 millisecond prior to intersecting the hazard cloud. Hence, in practice, the UAV which had a high functional LOA in its navigation capability, autonomously generated a new flight path that would deviate from the hazard zone approximately one second prior to the collision with that hazard.

Similarly, UAV_3 has the notation of `'+lim25'` following the `'fuel'` condition. This denotes that the mitigation strategy (new flight path) activates when the minimum

fuel limit (`'lim'`) of 25% (`'25'`) of the UAV is reached. In practice, the UAV which had a high functional LOA in its navigation capability, autonomously regenerates a new flight path that would ideally reach a refuel zone and continue onto its original search zone or a new and closer search zone, or alternatively, deviate the flight track and land at a safe location.

4.2.2 LOA/LOD Visual Representation

A total of three LOD was the outcome of the final refinement of the FCF, and it was referred to a high LOD (LOD 3), medium LOD (LOD 2) and low LOD (LOD 1). Each element in each layer had its own respective visual representation that was structured in layers. Furthermore, the UAVs had autonomy heterogeneity in their capabilities, hence, the subsystems of each UAV had different LOA. This also visualised in the interface as illustrated in Figure 4.5.

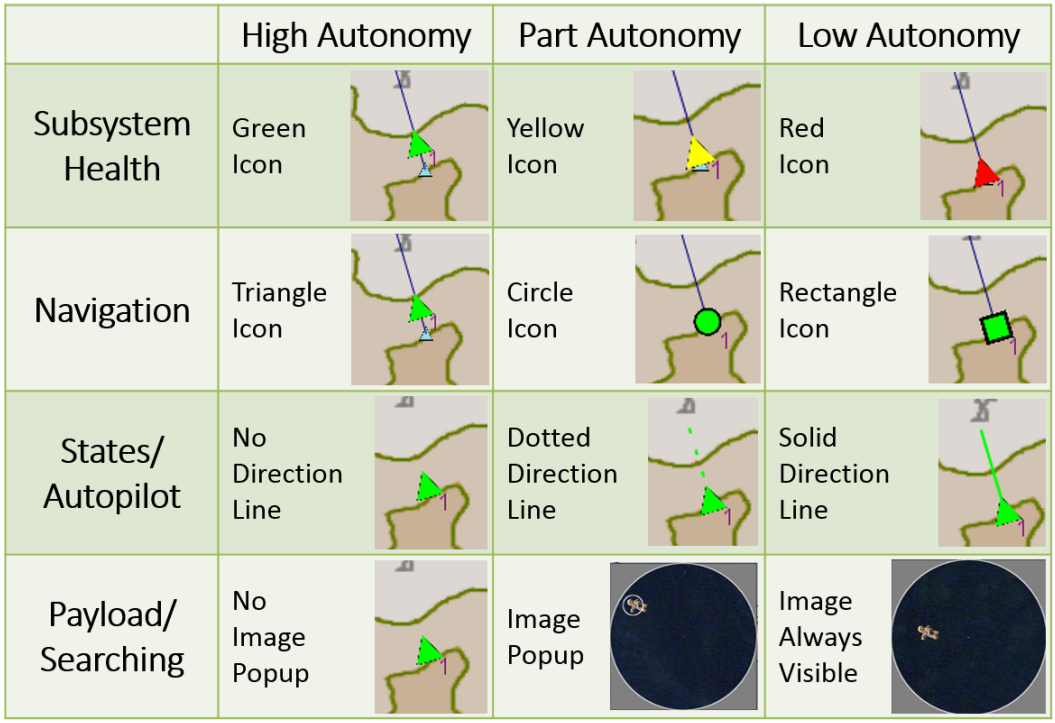


Figure 4.5: Icons that represented the various levels of autonomy with respect to the subsystems

The following sections describe the visual representations of both the LOD and LOA as it was appeared on the interface.

Health Functional Subsystem

Table 4.1 presented the descriptions of the three LOD of the *Health* functional subsystem and its equivalent visualisations.

Table 4.1: Three LOD descriptions for the *Health* functional susystem

LOD	Description
High	The highest level of health subsystem abstraction, no specific health data was visually represented.
Medium	A pie graph overlayed above the UAV icon. This pie graph represented the UAV's overall health subsystem information; the healthier the UAV was, the higher percentage of the pie graph was covered, and the greener the shade of the coverage was. Similarly, with a lower health level of a UAV, the lower percentage of the pie graph was covered, and the shade colour gradually became red.
Low	A detailed breakdown of the health subsystem of the UAV was represented. There were five subsystems; the Fuel quantity, the Electrical strength/health, the Controllability, Propulsion system and the Hydraulic system. However, only the fuel and the electrical health was implemented and used, while the remaining subsystems remained dysfunctional and static throughout the experiments.

Table 4.2: Three LOA descriptions for the *Health* functional susystem

LOA	Description
High	The UAV icon appeared green, denoting that a high autonomy in the UAV's health monitoring subsystem. A high LOA represents the UAV's autonomous ability to monitor the health status of the operational subsystems and is able to determine the appropriate time to alert the operator with necessary information about the health of the UAV subsystems.
Mid	The UAV icon appeared yellow, denoting that a moderate amount of autonomy is available in the UAV's health monitoring subsystem. A mid LOA represents that the health monitoring capability is capable of detecting any malfunctions within the UAV that may cause abnormal behaviours, but will not be able to determine which system, hence, the human operator is required to determine which subsystem is malfunctioning.
Low	The UAV icon appeared red, denoting that a low or no autonomy is available in the UAV's health monitoring subsystem. At this LOA, the monitoring of the UAV's health levels is the human operator's complete responsibility. The colour chosen does not reflect the UAV's health condition. It only reflects the autonomy of the UAV's health monitoring capability.

Table 4.2 presented the descriptions of the three LOA of the *Health* functional subsystem and its equivalent visualisations.

Navigation Functional Subsystem

Table 4.3 presented the descriptions of the three LOD of the *Navigation* functional subsystem and its equivalent visualisations.

The autonomous capability of this subsystem, where the degree of the UAV's ability to regenerate new flight paths when they are confronted with the hazardous environment,

Table 4.3: Three LOD descriptions for the *Navigation* functional susystem

LOD	Description
High	Waypoing and directional information is derivable from the UAV's trajectory.
Medium	Directional information such as specific waypoints of a UAV was available. Each waypoint contained a preferred waypoint altitude and speed which the UAVs were required to adhere to when overflying the specific waypoint.
Low	A flight path tracking line was also visually represented.

Table 4.4: Three LOA descriptions for the *Navigation* functional subsystem

LOA	Description
High	The UAV is able to determine the most appropriate flight path for any given scenario. A triangular icon is used to represent this autonomy.
Mid	The potential new flight paths are generated based on the realtime environmental situation autonomously. The human agent must choose the most appropriate option to carry through. A circular icon is used.
Low	New flight paths are required to be generated by the human agent manually. Under this autonomy, a square icon is used.

is represented by the shape of the UAV icon. The mapping of the icons was selected with the aim to differentiate the LOA of the navigation capability. However, further research is required to identify the optimal graphical representation of navigation autonomy.

Table 4.4 presented the descriptions of the three LOA of the *Navigation* functional subsystem and its equivalent visualisations. The shapes of the icon are the representations of the autonomous navigation capability, and it is utilised independently from the other subsystem autonomous capability's graphical represention.

States/Autopilot Functional Subsystem

Table 4.5 presented the descriptions of the three LOD of the *States/Autopilot* functional subsystem and their equivalent visualisations.

The autonomous capability of this subsystem, where the degree of the UAV's ability to control and manage the three crucial aspects of location - speed, altitude, and direction -

Table 4.5: Three LOD descriptions for the *States/Autopilot* functional susystem

LOD	Description
High	The position data of the UAV was represented in a form of a UAV icon over the map of the arena.
Medium	A graphical representation of the altitude, speed, and thrust setting of the UAV was available.
Low	The sensory reading of altitude, speed, and thrust setting of the UAV was also visually represented in a separate panel.

Table 4.6: Three LOA descriptions for the *States/Autopilot* functional susystem

LOA	Description
High	The UAV is able to autonomously manage all the state parameters to the desired waypoint and the waypoints' requested speed and altitude, is represented with no physical leader lines. The purpose is to avoid any unnecessary cluster on the interface.
Mid	The UAV is able to autonomously manage the direction of flight, the speed and altitude parameters must be manually managed by the human agent to match the desired waypoint speed/altitude requirement. At this LOA, the leader line is presented as a dashed line.
Low	The UAV is unable to autonomously manage any of the UAV's three state control parameters, that is to say, control management is the responsibility of the human agent, the leader line is presented as a solid line.

is represented by the UAV icon's leader line. Three types of leader line representations are used to depict high, mid, and low LOA, and the length of the line represents the distance the UAV will travel in one minute based on its current speed. The choice of interface representation is based on the Thales Group's EuroCat-X air traffic management system [216].

Table 4.6 presented the descriptions of the three LOA of the *States/Autopilot* functional subsystem and their equivalent visualisations. The visual representation of this autonomous capability is used independently from the other subsystems autonomy representation, hence the operator can determine a UAV's state autonomy in isolation from the other subsystem's representation.

Payload Functional Subsystem

The functional capability of the payload subsystem was different from the remaining three functional subsystems. The LOD corresponding to this subsystem was related to the functional LOA, hence, this subsystem's visualisation was described according to its LOA. For this reason, Table 4.7 presented the descriptions of the three LODs and three LOAs of the *Payload* functional subsystem and their equivalent visualisations.

This subsystem is involved the identification and differentiation between IOIs and decoys. The purpose of the agents was to identify and mark all IOIs and ignore the decoys. Three variations of the IOIs and decoys were included: Person-type icons were represented as IOIs (Figure 4.6a), and non-person type icons were represented as decoys (Figure 4.6b). The agent's (human/automation) responsibility of selecting the appropriate IOIs was governed by the amount of automation or LOA, designated by the experimenter and the LOA, which triggers the payload subsystem's LOD. It had three levels: high, medium, and low.

At the highest LOA, which means the UAV is capable of identifying and selecting all IOIs, the visual search scope would not automatically appear. However, the marking of

Table 4.7: Three LOD descriptions for the *Payload* functional susystem

LOD / LOA	Description
High	This LOD implies a highly autonomous searching capability where the UAV detects and decides on the appropriate Items Of Interests (IOIs) to select. No searching scope is visible by default. However, scope can be opened on request.
Medium / Mid	The payload search scope opens and closes automatically when a decision requires to be made by the human operator. UAV autonomously detects and identifies potential IOIs.
Low	The search scope's visibility is at the discretion of the human operator. The payload subsystem has no autonomy in terms of item search. All searches are done manually.

(a) The three Items Of Interest (IOIs) that were required to be found during the experiment

(b) The three decoys that were required to be ignored during the experiment

Figure 4.6: Possible items icons that appear in the scope during searches

the IOI was done autonomously.

At medium LOA, the UAV was capable of identifying and notifying the participant of potential IOIs by automatically making the search scope appear.

At low LOA, no assistance from the UAV was available. Hence, the scope would not only not automatically appear, but no identification or selection of the IOIs would be done. This was now the responsibility of the participants during the experiment.

The autonomous capability of this subsystem, where the degree of the UAV's ability to detect, identify, and select the IOIs, was represented by the search scope's popup behaviour:

A high LOA, where the UAV was able to autonomously detect any potential IOIs, identify the IOIs from any other items not of interest, and make a selection on the IOIs, was represented with a search scope which did not open up autonomously at any time. The autonomy could be verified with a selection marking indicator, where the IOIs had an orange circle overlayed.

Figure 4.7: Visualisation of the UAV with a high LOD of the states subsystem - illustrating the absence of any UAV state information

A partial LOA, where the UAV was able to autonomously detect and identify any potential IOIs, but where the verification that these IOIs were identified correctly, selecting the IOIs was the responsibility of the human agent. This autonomy was represented by a momentary pop-up and closing of the search scope. The pop-up behaviour was managed and initiated by the searching autonomy, where the machine agent detects and identifies potential IOIs. Once the potential IOIs have been set, the scope is popped-up for further verification and the next step - selection, to be carried out by the human agent. Once the machine agent determines either the identification was a false alarm or the IOIs no longer appears in the search scope field of view, the scope autonomously closes.

A low LOA, where the UAV was unable to autonomously detect, identify or select any IOIs, the responsibility falls to the human agent. At this LOA, similar to high LOA, the search scope will not pop up. However, the difference lies in that the IOIs. If untouched by the human agent, will not be identified or marked. Hence, to ensure that the IOIs are successfully selected, the human agent (participant) must manually open up the search scope, detect, identify and select the IOIs.

Example Configuration: States Subsystem

For example, as illustrated in Figure 4.7, at the highest LOD (High LOD), the least amount but the most aggregated information about the states subsystem was available as indicated by the position icon of the UAV. At the next LOD (Medium LOD), the abstraction decreased, hence the altitude, speed and thrust of the UAV was available (Figure 4.8). At the final LOD (Low LOD), another layer of less abstracted information was available to each of the three attributes (altitude, speed and thrust, Figure 4.9).

This layering of information applied to all four subsystems of each UAV asset.

The functional autonomy capability has also been visualised in the experiment.

4.2.3 Status Communication Feature

The Status Communication Feature, used in Experiment 3 (Chapter 7) enables the UAV to share live status and help-requests information to the participant. Upon reception of these information, the participants also had the ability to acknowledge help-requests sent by the UAVs. These were achieved by a rudimentary pane in a message box layout illustrated in Figure 4.10.

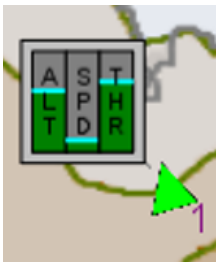


Figure 4.8: Visualisation of the UAV with a medium LOD of the states subsystem - illustrating the graphical aggregation of the UAV states information

The message box has four features:

1. Movability: The message box was capable of being relocated as preferred by the participants. Examples of when the participants moved this box would be to reposition it to work with their preferred workflow, such as relocating it to the bottom of the work area where it was convenient for them to have access to, or alternatively the participants would move the box into positions which they consider as less intrusive, i.e. where it would be out of the way of other on-screen information.

2. Minimisability: The message box was capable of being almost hidden completely from view, where only the upper dark gray pane would still be visible. This was to allow the participants to have greater freedom to minimise the amount of potential intrusion from the message box.

3. Scrollability: The messages inside the box were capableof being scrolled up/down, providing the participants with the capability to revisit message history. During

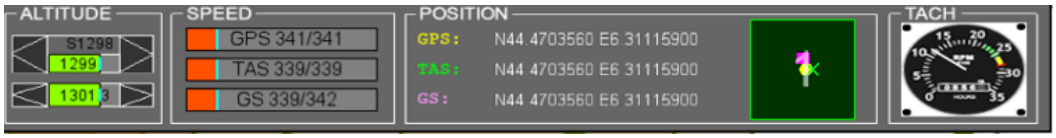


Figure 4.9: Visualisation of the UAV with a low LOD of the states subsystem - illustrating an raw level of UAV states information

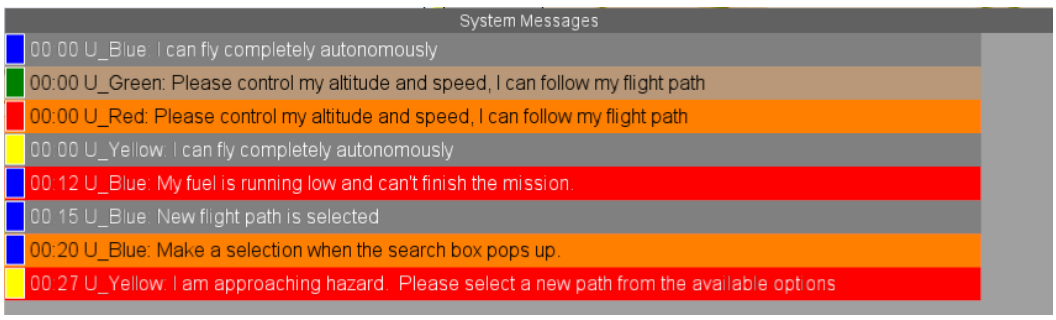


Figure 4.10: An example of the message box used during the experiment for the UAVs to share their status and help-request information to the participants during the experiment.

operation, the messages populated the visible areas of the message box until it overflowed, where arrows to the right of the message area would appear indicating new messages were accumulated. This provided the participants with the ability to scroll up and down through the visible messages, revisiting or addressing situations where the UAV required assistance.

4. Acknowledgeability: Some information presented with the messages involved requests for help from the UAVs. These message lines were usually highlighted and interactive. The participants were able to tap on the message once to cancel the highlight, indicating that the help request had been acknowledged and attended.

The format of the messages must be formulated in a way which is practical and easily comprehended by the human operators. For this reason, the structure of the messages is inspired by the Belief-Desire-Intention (BDI) software architecture [22, 181].

As discussed in Section 2.2.5 of Chapter 2, the messages do not require all three primary elements of the BDI architecture due to the purpose and the context of the messages, hence, the *Belief* element is omitted. Each message contains three parts: sender identifier, purpose of message, and intended (selected) action.

For example, if UAV_{Red} has the autonomy to re-plan its flight path to refuel when fuel level is becoming low, a message is constructed as:

U_Red: My fuel is running low, but I can go and refuel.

In this example, the *identifier* is *"U_Red"*, the purpose of message is *"My fuel is running low"*, and the intended action is *"but I can go and refuel"*.

However, in a situation where the UAV does not have the autonomy to perform the intended action, the message will be constructed as:

U_Red: My fuel is running low, please take me to refuel.

A generic message library that is customised to the experiment scenario is developed, and a suitable message will be used in synchronisation with the given situation.

4.2.4 Interaction Design

Each layer of information was individually accessible by a collection of gestures.

There were a few conventions that had been followed throughout the testing; long touch of a specific icon enabled the access of a lower LOD. For example, Figure 4.11 illustrates the transition from high LOD to low LOD by the long-touch gesture to access the medium LOD health subsystem information from the high LOD; long touch of the UAV icon was applied. Likewise, to access the low LOD from medium, a long touch of the medium LOD information menu is applied. In reverse, double tapping the menus can hide the lower LODs.

Figure 4.12 presented a summary of the gestures required to access different LODs for three different functional subsystems health, states, and payload, while the navigation

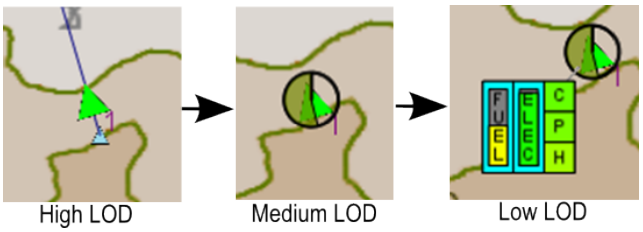


Figure 4.11: Health subsystem LOD gestures access example

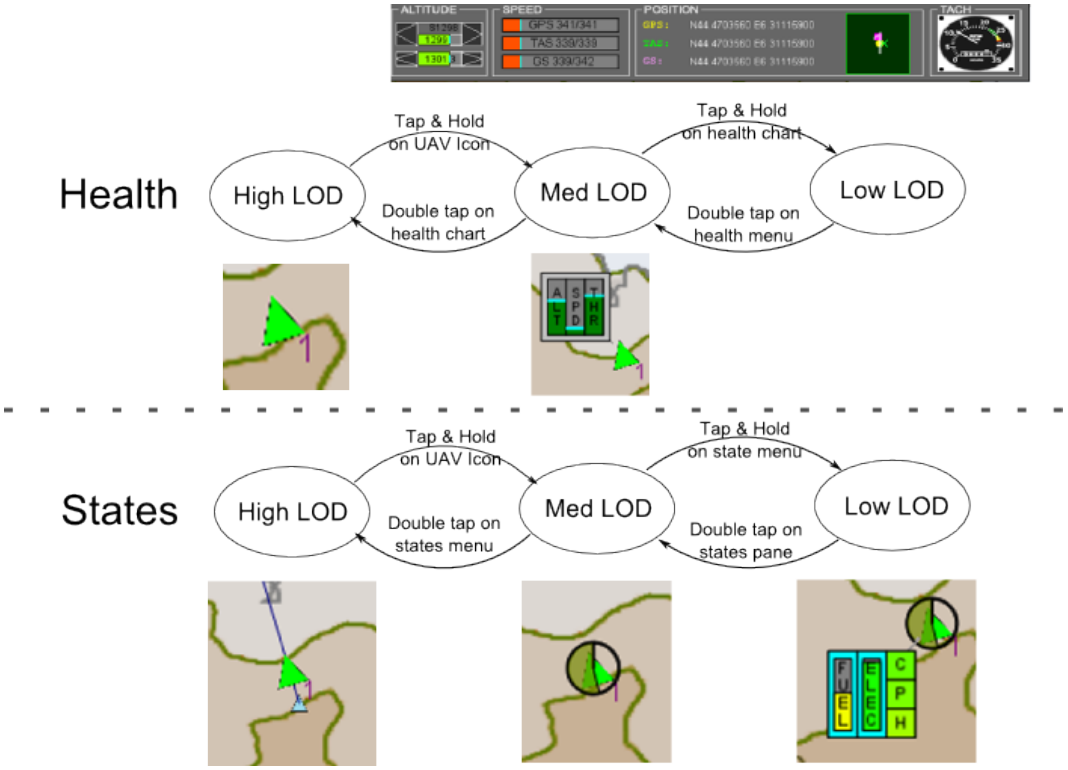


Figure 4.12: Summary of gestures required to transition between different LODs for different functional subsystems

subsystem was neglected, as it did not have the capability to be manipulated manually by the participants.

In addition, a set of gestures was designed for the interaction of the searching stage of the experiment where the agents (human or machine) monitored the search window, differentiated the IOIs from the decoys and made a marking selection over the IOIs.

For this feature, the human agents were able to relocate the search window (scope) around the projection area of the tabletop, resize the search scope, open, and close the scope. Once the scope was opened and IOIs were identified, they were able to mark or unmark the IOIs with a select, or deselect gesture. Table 4.8 presented a summarised list of gesture descriptions and examples.

Table 4.8: Table of gestures on the search payload (search scope)

Description	Gesture	Example
Open Scope	Touch & hold on UAV icon	Figure 4.13
Close Scope	Double tap on search scope	Figure 4.13
Resize Scope	One finger at each scope corner & drag in opposite direction	Figure 4.14
Move Scope	Single finger placed anywhere in the scope and drag	Figure 4.15
Select IOI	Touch & hold onto an item until star appears	Figure 4.16
Deselect IOI	Touch & hold onto an item until star disappears	Figure 4.17

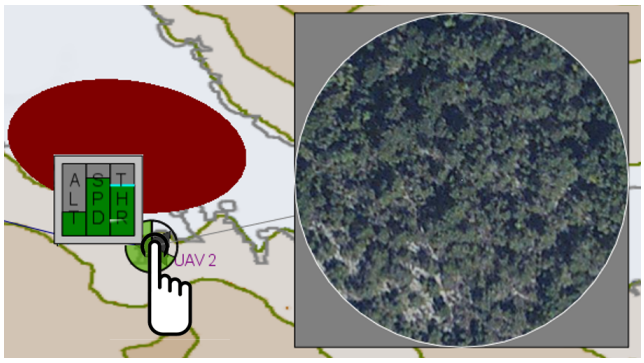


Figure 4.13: Open and close the search scope by touch and holding the UAV icon

4.3 Design and Apparatus

Experiment 1 was conducted in Lab-STICC, Departement Logiques des usages et l'information in Télécom Bretagne, France. Experiment 2 and 3 were conducted at ARCAA, an off-site research facility of QUT, Australia.

Even though the locations of the experiments were different, the testing environment was set up in a similar way. Figure 4.18 illustrates the typical arrangement of the testing facility. A square table was positioned on the floor with a DiamondTouch DT104 touch interactive tabletop positioned in the centre of the room on the table. A projector, which provided the projection of the experiment prototype hung over the tabletop. A chair was positioned at the bottom edge of the tabletop and the conductive mat (which was required by the device to correctly detect the hand and finger gestures when a participant interacts with the table) was positioned on the chair. Two physical cameras were positioned on the opposite edge of the tabletop to where the participant was situated. An additional table with a chair was situated offset from the main experiment zone where the host, or experimenter was situated. The general atmosphere of the laboratory was dim and isolated, which limited the interruption that could have an impact on the participants.

The experiment was captured using three video recording mechanisms; two cameras as illustrated in Figure 4.19 and a screen capture: an overhead camera capturing the hand gestures and the prototype projection on the tabletop, a facial camera to capture the expressions of the participants as they perform the experiment, and a software screen capture

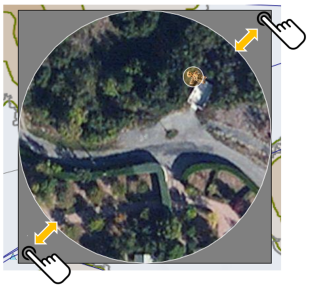


Figure 4.14: Resize the scope by dragging the opposite corner of the scope box in the opposite direction

of the experiment throughout the trial.

A Nikon D800 with a wide angle lens (Nikkor 14-24mm f2.8) was used as the overhead camera, as it provided a wide field of view to capture the participant's hand/finger gestures and the prototype projection. The video was exposed to the brightness of the projection illumination so that the images captured will be correctly exposed, leaving the hand/finger gestures to be in silhouette. The camera was mounted on a camera tripod, approximately 500mm overhung above the participant.

An off-the-shelf webcam was used to capture the facial expression of the participant (Face Cam). The webcam was situated on the same side as the overhead camera at the table height looking diagonally up towards the participant. The camera was set on auto-exposure which by default exposed to the participant's face brightness.

CamStudio (used in Experiment 1) and VLC player (used in Experiment 2 and 3) were used to capture the screen of the prototype throughout the experiment and was running on the gesture detection computer.

Experiment 1 used two separate desktop computers were used to run the experiment; one computer was used to run the software prototype and project the information onto the touch table, while the other computer was used to run DTIvy (the touch detection engine) and capture the facial and screen capture footage. These computers were linked via the local area network, while the image capturing was saved onto the hardrive internally (and not shared in any way). Experiments 2 and 3 used only one desktop and one laptop com-

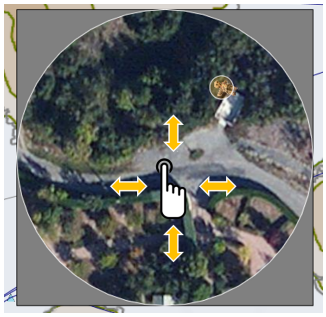


Figure 4.15: Move/relocate the scope by touch-and-drag anywhere within the scope box

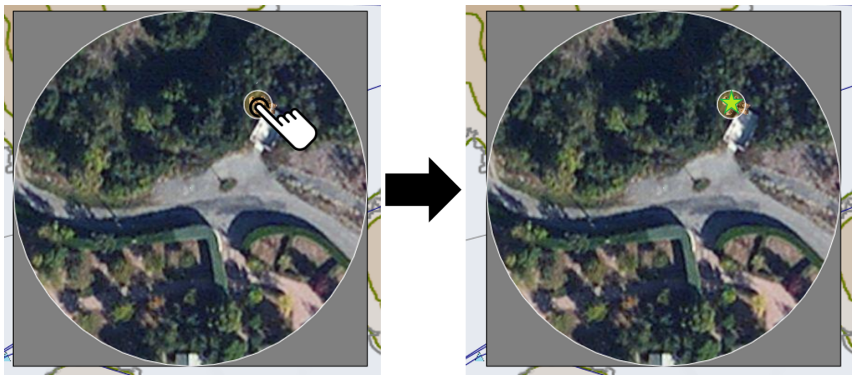


Figure 4.16: Select items by touch-and-hold on the icon until the item is marked by a star, denoting the item is selected

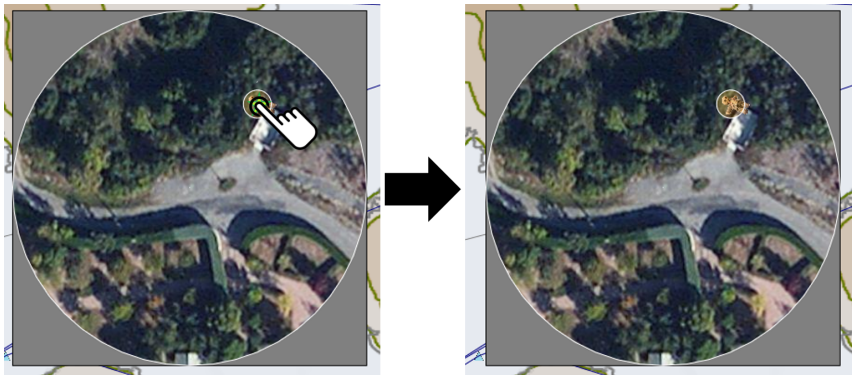


Figure 4.17: Deselect items by touch-and-hold on the marked icon until the star disappears

puter was used for the experiment; the desktop computer was used to run the experiment locally and captured the facial and screen capture footage. The laptop computer was used separately to provide the briefing presentation at the commencement of the experiment.

During experiments, a sign was placed at all entrances to the testing area to alert bypassers that an experiment was in progress and the zone should not be entered. However, permissions were given to other laboratory staff who needed to access other equipment unrelated to the experiment to enter the room. In these circumstances, the participants expressed no concerns of any distractions to their attempts at the experiment.

4.4 Procedure

The experimental procedure throughout this research is largely similar with some minor differences. However, the content and scenarios of each experiment are different. The details of the differences are outlined in Chapters 5, 6, and 7.

The steps in experiment preparation involved the planning of the experiment scenarios, specifying of the prototype requirements, implementing the prototype software, obtaining the ethics clearance, internal prototype functionality testing, internal prototype experiment testing, subject recruitment, session preparation, and finally, execution.

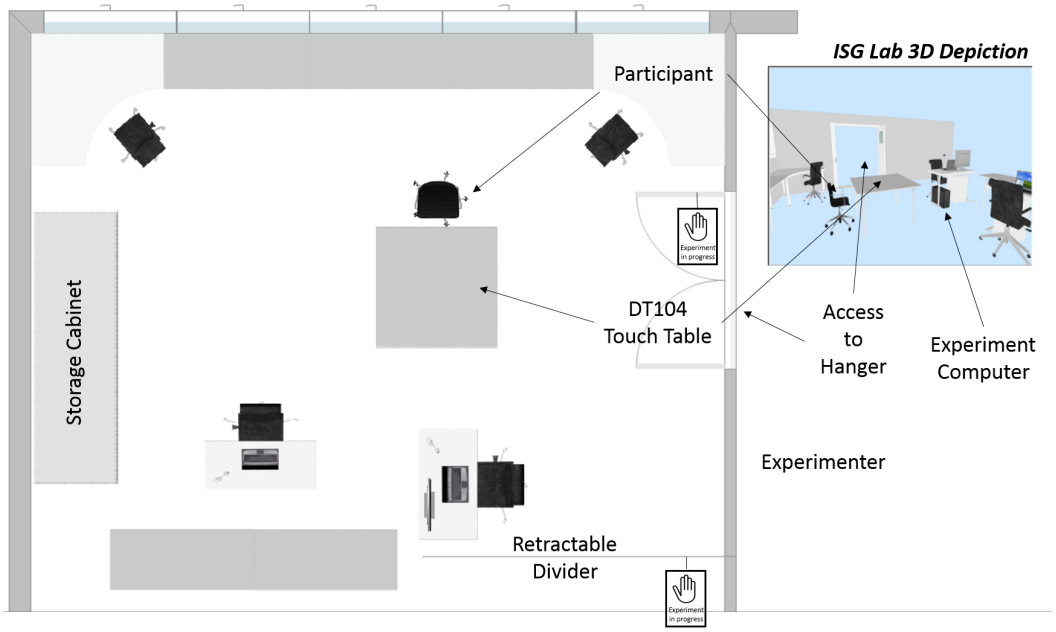


Figure 4.18: Testing facility set up floor plan at ISG Lab - ARCAA with an 3D depiction to the right of the image

Figure 4.19: Testing area setup, illustrating the placement of cameras in relation to the other hardware components (image used with the signed consent from the participant)

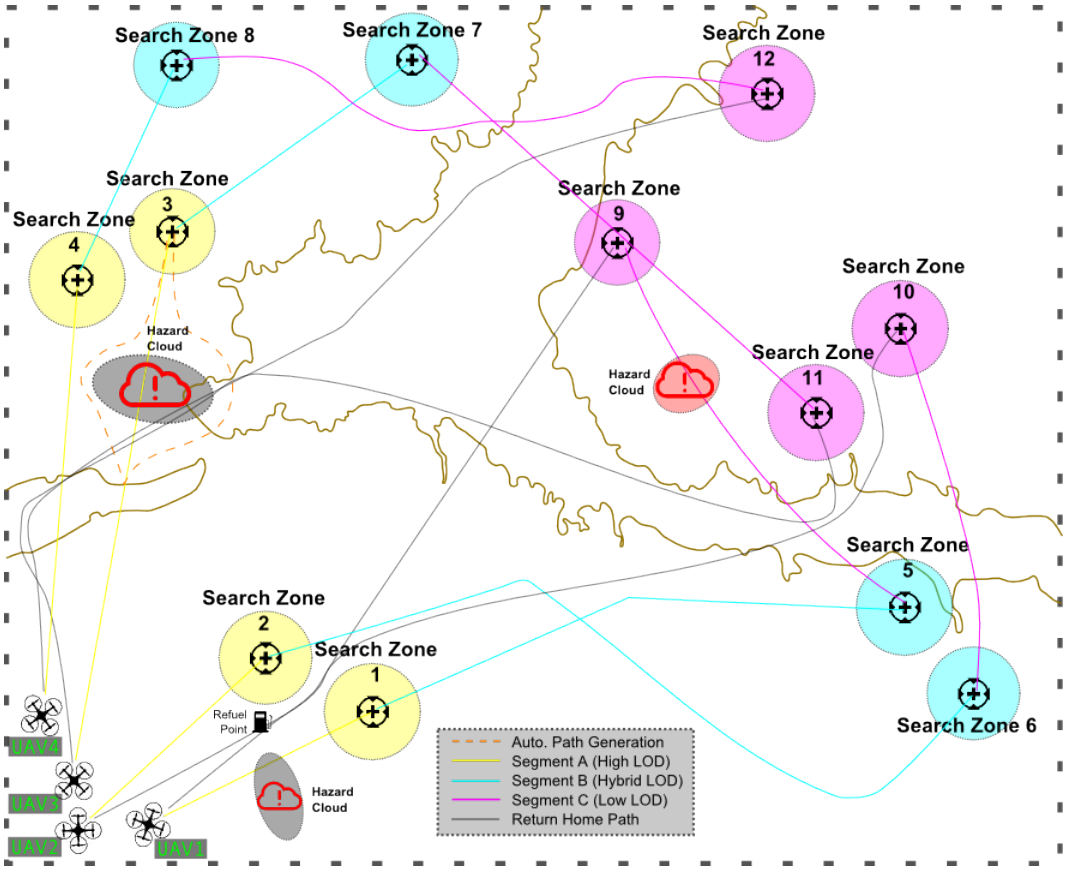


Figure 4.20: Three segments of scenario elements of the experiment's storyline: Segment A elements are illustrated in yellow, Segment B elements are illustrated in pink, Segment C elements are illustrated in cyan.

4.4.1 Experiment Preparation

The preparation of the experiment took place in three parts: theoretical grounding, software preparation, and ethical clearance. This section details the procedure of each part.

Theoretical Grounding

Theoretical grounding includes identifying the experiment aim, and the design of the scenarios to achieve the aim. The aims of the three experiments are presented in Section 4.1. Once the aim and its associated theory were defined, the experiment scenario planning began. The planning involved the designing of a story, initially stepping through sequences of events using a whiteboard and magnets, this is then translated onto a storyboard keyframed by significant changes in the scenario; such as a UAV encountering a hazardous cloud that required an alternate strategy of evasion. Figure 4.20 illustrates an example of the scenario elements of Experiment 1 for the three segments described in Section 5.1 in Chapter 5.

Upon the completion of the preparation, the experiment could commence.

Software preparation includes implementing the software prototype to accommodate

the experiment's specific needs (Section 4.2), and the implementation of new familiarisation scenarios.

All three experiments were designed to take 90 minutes. A detailed schedule is illustrated in Table 4.9.

Software Preparation

Software preparation includes identifying the prototype specifications and experiment-specific requirements. These specifications and requirements were derived from the theoretical grounding and the scenario specifics, which is then scripted and captured in the scenario files (described in Section 4.2.1). The details of the software prototype platform is presented in Section 4.2.

Ethical Clearance

Once the software prototypes implementation neared completion, internal functionality testing began, the experiment ethics application was compiled, submitted and reviewed to the QUT Ethics Committee (approval number 1300000530).

The final stages of the experiment preparation involved compiling a Power Point presentation to aid the experiment brief given to the participants during Experiment 1, and recording this presentation for Experiments 2 and 3. This presentation introduced the research objectives, experimental procedure, prototype usage, and the data collection process.

4.4.2 Subject Recruitment

During final stages of experiment preparation run-through, an email was sent out to seek expressions of interest to participate in the experiment. For a statistically significant pilot study, 23 or more sample points must be collected [44, 40]. Therefore, a batch email was sent to fellow students and academic staff of each home institution (Télécom Bretagne or QUT), and the neighbouring businesses to request for expression of interest to participate in the experiment. The invitation email included an online schedule on which the potential participants could indicate their availability, a medical requirement requiring interestees cannot have any diagnosis of colour-blindless, a language requirement that required the interestees to have a *reasonable* command of English, and an experiment consent form. From their responses, the experimenter was able to form a testing schedule.

The reason for the requirement that the participants must possess no prior diagnosis of colourblindness was set due to the fact that colour-coding was used throughout the experiment, particularly in interface colour selection to cue certain system or situational behaviour. Colourblindness might present situations where the participants were not able to fully take advantage of the nature of the colour code related meaning, hence handicapping the knowledge acquisition stage, producing contaminated results. The reason for the requirement that the participants must be fluent in comprehending oral and written

Table 4.9: Experiment schedule of each experiment

Time	Experiment 1	Experiment 2	Experiment 3
10 min	Introduction/briefing		
20 min	Prototype familiarisation (hands-on)		
10 min	Segment A	Scenario 1 (Base/Eval)	Scenario 1 (Base/Eval)
10 min	Data Collection	Data Collection	Data Collection
10 min	Segment B	Scenario 2 (Eval/Base)	
10 min	Data Collection	Data Collection	Scenario 2 (Eval/Base)
10 min	Segment C	Post-experiment interview	Data Collection
10 min	Data Collection		

English was due to the fact that English was the expected language of communication, and that the software prototype and the data collection process required the participants to comprehend data and queries in written English. A less-than-fluent level of English could cause misunderstanding of communicated intentions, hence producing contaminated results. The latter requirement was especially important for Experiment 1 as the experiment was conducted in France, therefore majority of the participants had English as a Second Language (ESL).

As such, 45 participants were recruited to perform this experiment.

The results from the outreach concluded with approximately 40 respondents for the first experiment, 45 from the second, and 39 from the third experiment who indicated interest. The participant age ranged from 20 to 60 years old, and the majority of the participants had very little to no prior experience in UAV operations or supervisory control. A large portion of these candidates eventually participated in the respective experiment.

4.4.3 Greet and Brief

After each participant was greeted by the experimenter, he/she read, agreed and signed the ethics participant information and consent form as a requirement prior to any further experimental procedures. Once the consent form was signed, the participant was invited to sit at the touch tabletop where the experiment was conducted.

At this time, the experiment briefing presentation, which consisted of either a verbal presentation (Experiment 1) or a scripted voice-over PowerPoint presentation (Experiments 2 and 3), was presented to the participant. The briefing included a simple introduction and research aim, the estimated timing of the respective experiments (Table 4.9 outlines the timing for each experiment), the participant's tasks and objectives, the potential unexpected/perturbation events, an overview of the software prototype that he/she would be using, and a prompt for questions which the participant might have.

Upon the conclusion of the briefing and when the participant felt satisfied with his/her level of understanding, *prototype familiarisation* would begin.

4.4.4 Prototype Familiarisation

There were several repeatable scenarios designed for each participant to see and become accustomed to the experiment platform and prototype prior to performing the experiment. Each scenario was designed to be completed within three or four minutes with commentary and explanation from the experimenter. However, the participants had a total of ten minutes to complete each scenario. The participants also had the opportunity to repeat any one of the scenarios if he/she felt more practice was required.

The three experiments had different levels of familiarisation content, as for example, Experiment 1 participants did not need to know the operations of the Experiment 3's message box. The following subsections describe the familiarisation scenarios for Experiment 1, 2, and 3.

Experiment 1 and 2

The general familiarisation process for all three experiments included only one UAV which flew directly to the designated search zone in the first scenario. The scenario layout is presented in Figure 4.21. During this scenario, the participant was introduced to the prototype interface, basic gestures, the icons used throughout the scenario and the methods to perform the tasks of supervision and search.

The second scenario built upon the knowledge and the practice of the first scenario, and further introduced multiple UAV operations, hazardous cloud visualisation, and dynamic path replanning. The scenario layout was presented in Figure 4.22. This scenario provided more opportunity for the participants to practically familiarise with the prototype.

Scenario three again built upon the two previous scenarios, with the addition of a refuel zone. The scenario layout was presented in Figure 4.23. The participants were introduced to the refuel zone's presentation and behaviour.

The final scenario is the most complete and complex scenario. This provided a sound opportunity for the participants to perform the trial with minimal assistance from the experimenter, which better prepared them for the main experiment.

Experiment 3

The third familiarisation scenario of Experiment 3 was different as the new feature of a message box was introduced. An extra fourth scenario was also added to ensure sufficient training is delivered prior to the commencement of the experiment.

The first two scenarios were similar to Experiments 1 and 2, while the third scenario was changed to involve all four UAVs. Further to the items previously covered in earlier scenarios, this scenario introduced the communication (text-based) dialog box where the UAVs communicate their status, autonomy and requests with the operators during the mission. The items related to the usage of the text-box that were introduced were:

- an overview of the message box, its main purpose, and the three message types:

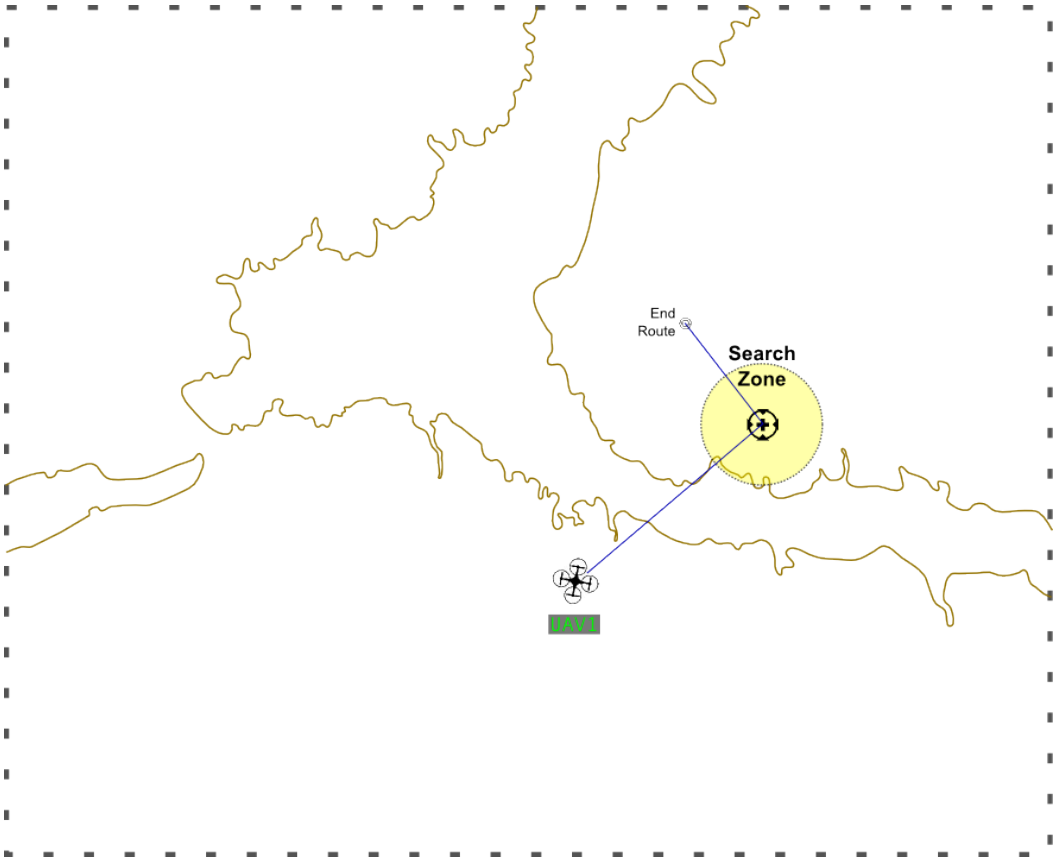


Figure 4.21: Layout of the first familiarisation scenario - basic prototype interface introduction

status/intention (gray), warning (orange), and alert (red)

- the reposition and minimisation feature of the message box
- the structure of the message in the message box: Colour, UAV callsign, content of message
- the acknowledgement feature and it's appropriate use

The final scenario also involved all four UAVs and all the content previously covered. This scenario provided an opportunity which was more complicated, demanding, and up-paced, and the purpose was to allow the participants to familiarise with a fully loaded/complicated scenario similar to ones they were expected to perform in the main trials of the experiment. The scenario was run under similar conditions as the formal experiment trials where limited assistance regarding the scenario situation was provided by the experimenter, only assistance that was related to the physical usage and interpretation of the prototype was provided.

The participant's competencies for the experiment and the tools were gauged subjectively by the participant. They were encouraged to reattempt the simple practice scenario as many times as necessary until they feel they are prepared to proceed with the first scenario.

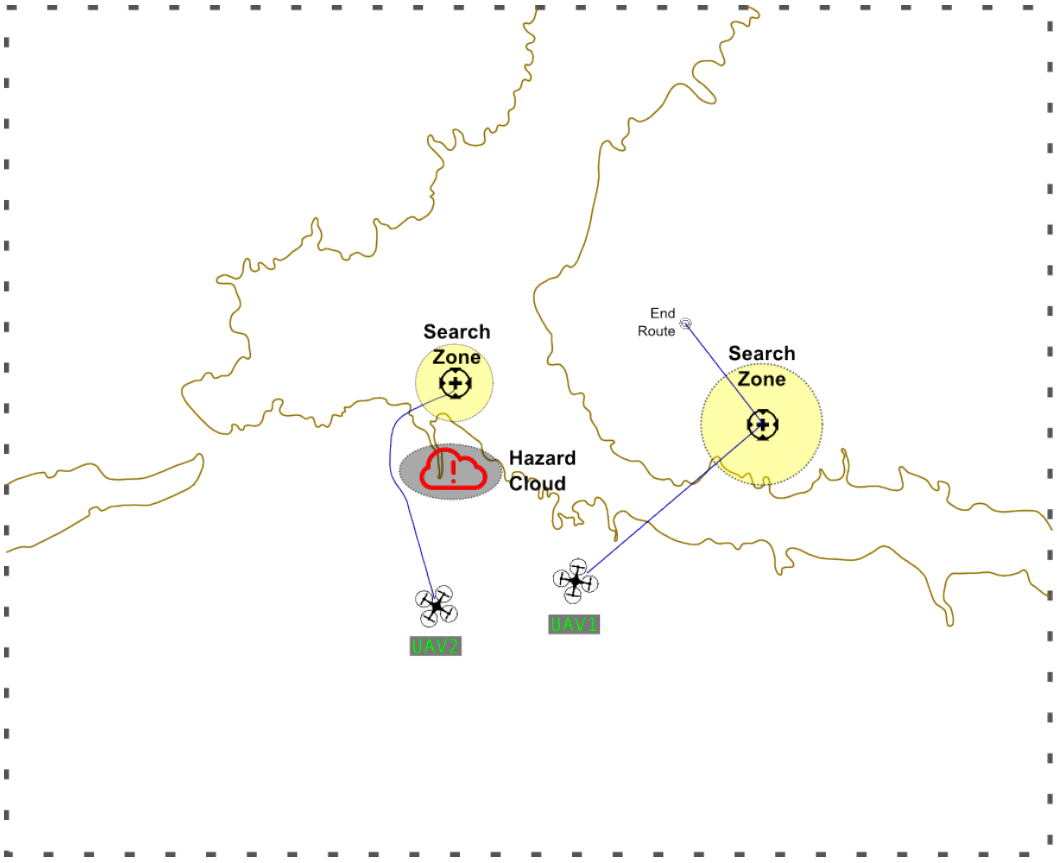


Figure 4.22: Layout of the second familiarisation scenario - hazard area and multiple UAV operation

4.4.5 Experiment Trial

The main experiment component consisted of three segments for Experiment 1 and two separate trials for Experiments 2 and 3. Each segment or trial consisted of a test scenario and a set of questionnaires. Experiment 1 commenced with Segment A, followed by two paper-based questionnaires (SA and CW), then Segment B and C would follow. Experiments 2 and 3 commenced with either the baseline or the evaluation scenario, and then a set of online SA questionnaire and a paper-based CW rating form, followed by the remaining scenario.

The order of these scenarios was interchangeable as it promoted a bias-free data collection effort; free from any possibilities of situation familiarisation, half of the participants commenced with the baseline scenario as the first stage, while the other half commenced with the evaluation scenario.

After the completion of the segments/scenarios, each participant was thanked for their time with a small token of appreciation (a small chocolate bar).

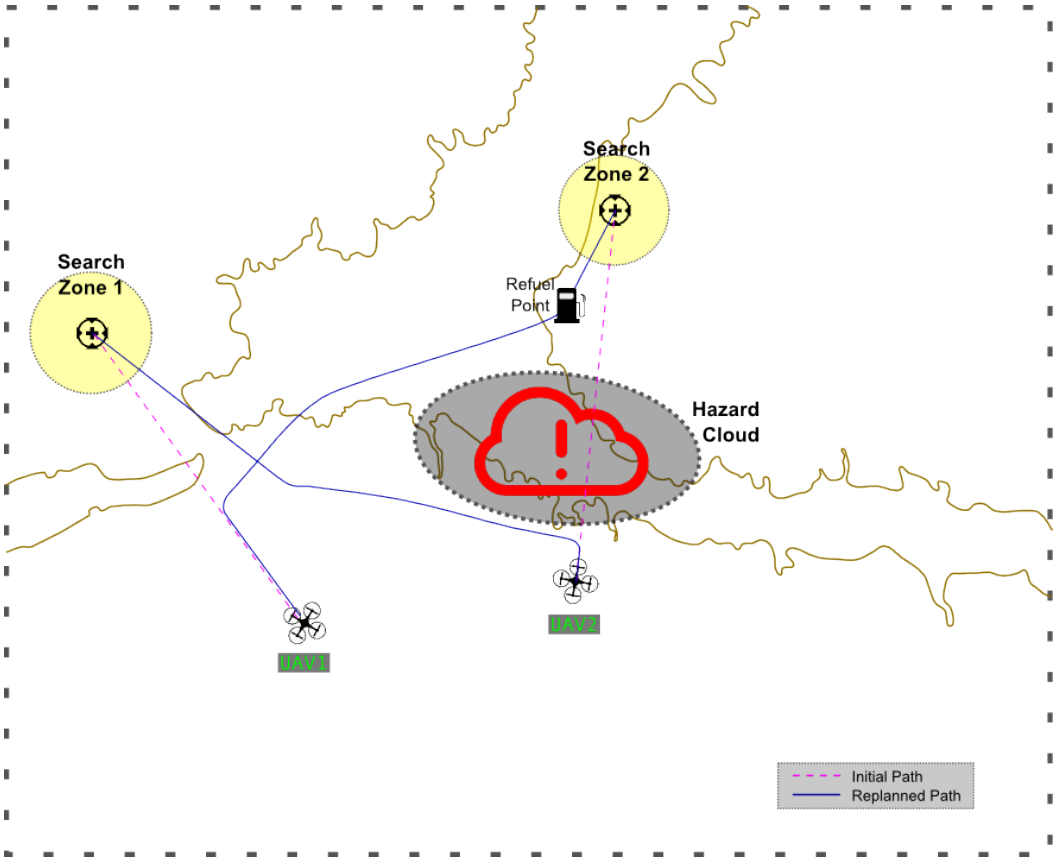


Figure 4.23: Layout of the third scenario - refuelling and multiple UAV operation

4.4.6 Data Collection

The data collection process involved capturing three types of information; the software attributes and status, the participant's interaction and facial expression through video recordings, and the participant's cognitive performance. The software attributes were captured during the experimentation as functionalities were inbuilt to record all software aspects of the experiment at a frequency of 2 Hz, that is, the information regarding the UAVs and the status of the environmental variables were written to a log file every 500 ms. This information was then used as an objective measure of the participant's mission performance. The participant's hand/arm gestures and their facial expressions were captured through video, which was used to further analyse specific reaction times in the third experiment. This meta-information was used to assist in confirming the expression associated with the interactions and prototype status recorded by the software. This experiment's primary instruments used to measure the participant's cognitive performance were a web or paper-based rating form which recorded the subjective feeling about the participant's CW, SA, and Trust in automation. These metrics records the participant's understanding of the situation presented during the corresponding segments of the experiment that they had just completed.

This section reviews the key aspects of each data collection instrument from their

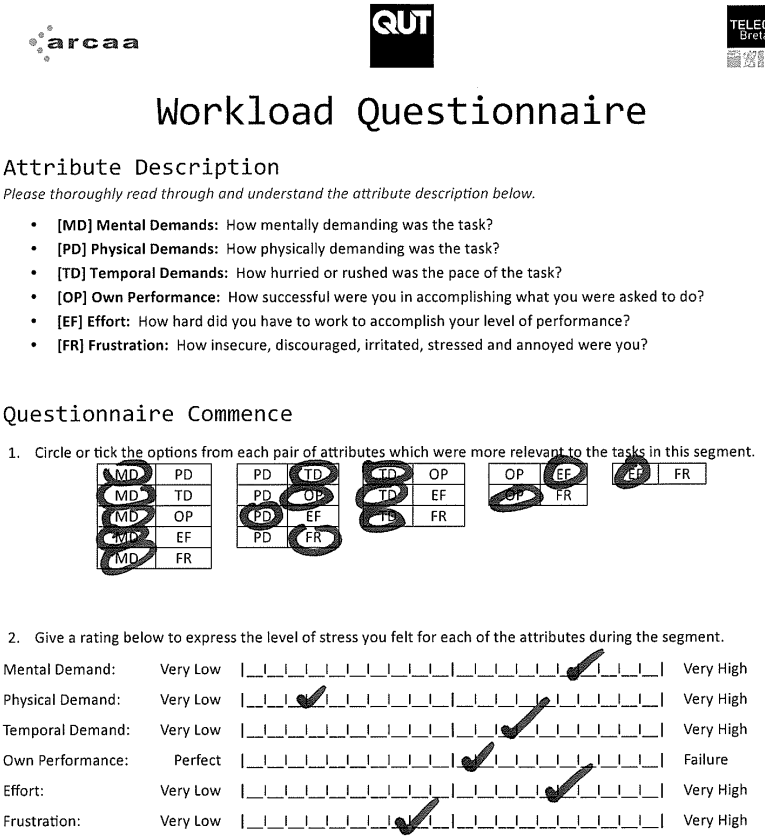


Figure 4.24: An sample of the assessment rating form completed anonymously

practical aspect as they were used in the experiments.

Cognitive Workload - NASA-TLX

Hart and Staveland defined CW in six dimensions [97], these include the operator's mental workload, physical workload, temporal workload, operator performance, effort, and frustration. A subjective assessment of these attributes were conducted using the NASA-TLX instrument. This form was simplified to capture only the ratings of each attribute, neglecting on the weighing of the attribute's importance and relevance, hence, losing resolution to the final result. As such, a modified assessment form (Figure 4.24), similar to the form initially proposed in [97], was developed to streamline the assessment process. Since not all the attributes were equal in terms of significance to the experiment, the attribute weigthing measure contributed to differentiate the attributes that had greater significance from ones that did not.

Upon the conclusion of each scenario, the participant received an unfilled NASA-TLX rating form to complete. In order to reduce the effort and time required to understand the form and the time, the experimenter offered assistance to explain the assessment process to the participants to help them complete it.

The example in Figure 4.24 illustrates a typical form when completed, details on the

results were presented.

During the analysis phase, the weighting and the rating of each attribute were tabulated. The influence of the weight on the significance of the attribute (as a factor) was determined using equation x as the weighting pair, with 15 combinations, while the percentage of the rating was normalised using equation 4.4.6 as the rating scale had 20 steps.

$$Sig_{attribute} = \frac{Weight_{attribute}}{15}$$

$$Rating_{percentage} = 5 \cdot Rating_{raw}$$

$$|Score_{attribute}| = Sig_{attribute} \cdot Rating_{percentage}$$

$$CW_{combined} = \sum |Score_{attribute}|$$

This process was carried out for all six attributes and an average combined percentage was established using equation 4.4.6.

Situation Awareness - SAGAT

SA referred to in this research adopted Endsley's definition, which was *a person's perception of the elements of the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future [75].* The instrument used to assess SA was the SA Global Assessment Technique (SAGAT) [76]; which was a set of custom-designed questionnaires tailored to the experiment details, to assess an operator's knowledge of the situation and mission environment.

This questionnaire was implemented by hand in Experiment 1, and online using QUT's KeySurvey system in Experiments 2 and 3. Given SA level 2 can only be satisfied if level 1 was satisfied [76], the system was implemented such that if a level 1 query for a specific task was incorrectly answered, its following level 2 SA query would not be presented to the participant to answer.

SA was divided into three levels; 1) Perception, 2) Comprehension, and 3) Projection. These three levels had been studied extensively in many areas including unmanned vehicle operation and teleoperations. From past research, a list of SA requirements was compiled for Unmanned Ground Vehicles (UGV)[192], which shared many similarities with UAVs. Further synthesis was undertaken to extract the core of the requirements that were applicable to UAVs in this experiment.

The requirements outlined in the refined list illustrated in Table 4.10 was administered to the operator as an SA questionnaire, and a relevant subset was scoped into questions with multiple choice answers.

Table 4.10: SA requirements motivated from the UGV requirements (Adapted from [192])

Level 1	Level 2	Level 3
Heading of vehicle	Area coverage	Projected control actions
Starting position/base	Vehicle capabilities	Projected ability to complete mission
Ground clearance	Vehicle limitations	Projected target identity
Partitioning of search area	Impact of weather on time to task completion	Projected coverage of area
Robotic vehicle assigned to area	Status of task/progress	Projected activities of other robots
Search strategy of vehicles(s)	Task priority	
Automation (on/off, LOA)	Robot's ability to perform mission	

Further to this list, Drury and Scott [70] had proposed a model of awareness in UAV operations. In their study, the proposed framework included four parts of the UAV-related SA decomposition for one operator to one UAV operation, and multiple operator to multiple UAV operations as well [69]. These four parts include [70]:

- Human operator's awareness of the UAV
- Human operator's awareness with other human operators
- UAV's awareness of the human operator
- UAV's awareness of other UAVs

The key relationship being investigated in this experiment was the operator's awareness of the UAVs . Several elements of awareness were identified by Drury and Scott [70]:

- 4-Dimensional Spatial Awareness: Geographic location, altitude, velocity, and time
- UAV capabilities: sensors, communication links, contingency logic (alternate plan) etc.
- Health: Fuel, electricity, and other consumables
- Other (non-health) statuses: Current/predicted flight parameters, current sensory status, LOA etc.
- Weather: Current weather, predicted weather
- Certainty: Accuracy and integrity of sensory data

However, many of these elements did not apply to the experiments, and they were not contributing factors to the dynamics of the testing scenarios.

Table 4.11: Number of SA questions in Experiment 2

Scenario	Level 1	Level 2	Combined
Baseline	4	5	9
Evaluation	5	4	9

The experiments captured the first two levels of SA - Perception and Comprehension.

Level 1 queries were designed to assess the participant's perception of the situation, that is, whether a change in the state of the UAVs or the environment was detected by the participant. As an example from Experiment 2, Level 1 SA questions included *Which UAVs had an obstacle along the flight path?* and *Which UAV deviate from its original path to refuel?*

Level 2 queries were designed to assess the participant's comprehension, or their understanding of the situation (perceived in the previous level) - whether the participants understood the implications of the UAV's change in status. Examples from Experiment 2 of Level 2 SA included *What did* UAV_4 *do when the level became critically low?* and *What sort of change (to the flight path) occurred on* UAV_2*?*

During the analysis, each participant's Level 1, 2 and combined SA results were normalised in an average percentage of the total number of questions related to that level of SA. Table 4.11 illustrated the number of Level 1, 2, and combined (Levels 1 and 2) questions in each scenario.

For example, sample A scored three correct level 1 SA questions and two correct Level 2 SA, the percentage of correct answers for sample A Level 1 SA became $\frac{3}{4} = 75\%$ and level 2 became $\frac{2}{5} = 40\%$, while the combined SA percentage for sample A became $\frac{SA_1+SA_2}{Q_{SA1}+Q_{SA2}} = \frac{3+2}{4+5} = \frac{5}{9} = 56\%$.

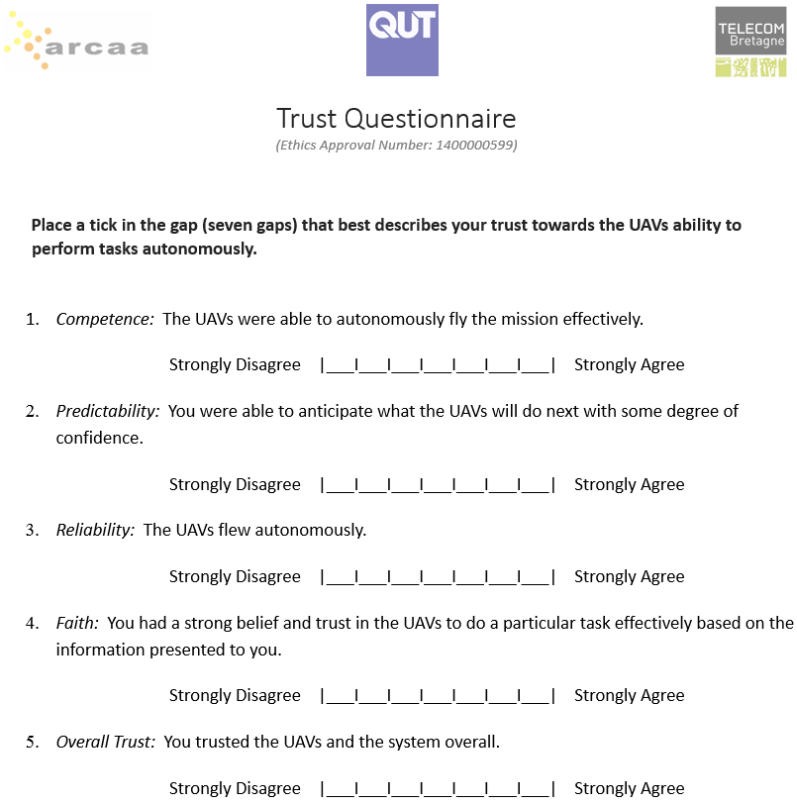
Hypothesis tests were then conducted on these percentages to compare the means.

Automation Trust

Experiment 3 investigates the impact of autonomy transparency. As such, a key human performance metric which is directly influenced by the degree of autonomy transparency is trust. Therefore to understand the impact of autonomy visualisation on human-machine trust, Jian *et al.*'s scale of trust questionnaire is used [111] as a measure of trust in Experiment 3.

Lee and See define trust as *the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability [126].* And *Jian et al.* described three form of trust:

1. General Trust,
2. Human-Human Trust, and
3. Human-Automation Trust.



Trust Questionnaire

(Ethics Approval Number: 1400000599)

Place a tick in the gap (seven gaps) that best describes your trust towards the UAVs ability to perform tasks autonomously.

1. *Competence:* The UAVs were able to autonomously fly the mission effectively.

   Strongly Disagree |___|___|___|___|___|___|___| Strongly Agree

2. *Predictability:* You were able to anticipate what the UAVs will do next with some degree of confidence.

   Strongly Disagree |___|___|___|___|___|___|___| Strongly Agree

3. *Reliability:* The UAVs flew autonomously.

   Strongly Disagree |___|___|___|___|___|___|___| Strongly Agree

4. *Faith:* You had a strong belief and trust in the UAVs to do a particular task effectively based on the information presented to you.

   Strongly Disagree |___|___|___|___|___|___|___| Strongly Agree

5. *Overall Trust:* You trusted the UAVs and the system overall.

   Strongly Disagree |___|___|___|___|___|___|___| Strongly Agree

Figure 4.25: An example trust rating form used during Experiment 3

The Human-Automation trust questionnaire consists of 12 questions, each is scaled from 1 (strongly disagree) to 7 (strongly agree). The questions start with a non-trustworthy impression, such as "The system is deceptive?", to the final question where the impression is completely trustworthy, such as "I am familiar with the system".

Participants received a rating form (Figure 4.25) along with the CW and SA questionnaires, which allowed them to give a rating based on the following questions:

1. *Competence:* The UAVs were able to autonomously fly the mission effectively.
2. *Predictability:* You were able to anticipate what the UAVs will do next with some degree of confidence.
3. *Reliability:* The UAVs flew autonomously.
4. *Faith:* You had a strong belief and trust in the UAVs to do a particular task effectively based on the information presented to you.
5. *Overall Trust:* You trusted the UAVs and the system overall.

The results were then manually recorded for all the participants for each scenario for the next step of the statistical analysis.

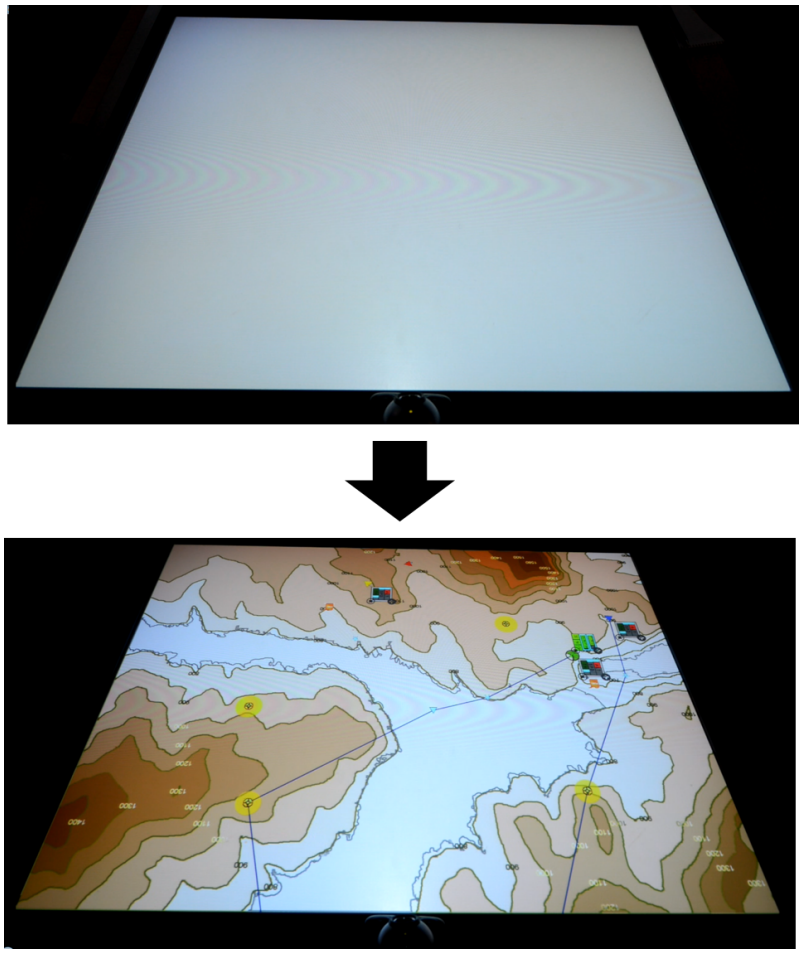


Figure 4.26: Experiment initial blank screen transitioning to scenario launch screen

Mission Performance

Three operator mission performance measures were extracted from the log files and video footages collected during the experiment to gauge the objective task performance differences of the participants between the two autonomy transparency configurations. Initial Response Time (IRT), Event Response Time (ERT), and the number of IOIs Found.

IRT referred to the time that the participants took to *correctly* conFigure the first UAV instance. This meant that t_0 (initial time) denoted the moment where the launch screen transitioned to the scenario screen (as illustrated in Figure 4.26) and t_{IRT} denoted the moment where the participant manipulated the controls of any of the UAVs that required operator assistance, such as UAVs that have a medium or low autopilot LOA, where human control was required to adjust the speed, altitude and/or direction of the UAV. The IRT was recorded in seconds.

ERT referred to the time that the participants took to *correctly* react to perturbation events, such as selecting a new flight path upon request by the UAV when the UAV was confronted by hazardous clouds. ERT's t_0 denoted the moment when a hazard was presented

or when alternative options to mitigate certain hazards were presented, and t_{ERT} denoted the moment when the participant *correctly* reacted to these events or options. The ERT was recorded in seconds.

IOIs Found referred to the total number of IOIs found throughout the scenario, where a percentage of the total IOIs deployed were recorded. For example, in a scenario, UAV_1 found 7 IOIs while 9 were deployed, UAV_2 found 8 while 9 were deployed, UAV_3 found 6 while 10 were deployed and UAV_4 found 9 while 9 were deployed. $IOI_{perc} = \frac{IOI_{UAV_1}+IOI_{UAV_2}+IOI_{UAV_3}+IOI_{UAV_4}}{\sum IOI} \cdot 100$, therefore, the total percentage of IOIs found in this example is $IOI_{perc} = \frac{7+8+6+9}{9+9+10+9} \cdot 100 = 81\%$.

4.4.7 Post-Experiment Interview

The post-experiment interview was formally carried out in Experiment 2 to gain some data point regarding the participant's perception of the system. At the final stages of Experiment 2, the participants were asked to answer a few casual questions regarding their experience of the experiment, and they were encouraged to answer more than what was asked. The aim of this interview was to capture the reasoning and justification of the participant's answer.

For example, if a participant scored a low SA score, a question *why did you think that your SA was lower in the first scenario than the second?* The participant's response might suggest that the reduction of the score was not directly due to a lack of information as initially anticipated, rather, a lack of hands-on participation during the trial.

The results collected in the interviews were considered qualitative, as oppose to the CW and SA measures.

4.4.8 Post Experiment

All three experiments concluded at the end of the 90-minute timeslot. After the participant had departed, the questionnaires were collected, the recorded data were moved to a portable storage location where the daily backup was carried out for safety, and the participant consent form was stored separately. A new set of consent form and questionnaires were prepared for the arrival of the following participant and the procedure was repeated.

4.5 Conclusion

This chapter had presented the core components which made up the majority of the three experiments: the aim, the software platform, the design and setup, and the experiment procedure.

Experiment 1 aims to validate the implementation of the FCF, Experiment 2 aims to evaluate the influence of autonomy transparency through interface graphical feedback on the operator's cognitive performance, and Experiment 3 aims to verify the effect of com-

municating the UAV agents' functional and autonomy capabilities to the human operators, in order to increase autonomy transparency.

These aims were tested through a software prototype which simulates a simplified hypothetical aerial search mission involving multiple heterogeneous UAVs. This software system was implemented for a touch interactive tabletop to enhance the participant's mission command experience.

Experiment 1 was carried out in France, while Experiments 2 and 3 were carried out in Australia. The procedure of the experiments was similar, where the experiment preparation started with a theoretical grounding, which is then developed into an experiment scenario that can be tested. Ethical clearance were obtained for the experiments and the cognitive performance, as well as objective performance data were collected throughout the experiment.

The details of the experiment scenarios and the analysis of the collected data were not presented in this chapter. However, their in-depth discussions were presented in Chapters 5, 6, and 7, or Experiment 1, 2 and 3 respectively.

CHAPTER 5

Experiment 1: Functional Capability Framework Validation

This experiment aims to validate the implementation of the Functional Capability Framework (FCF) proposed in chapter 4, and the experiment overview, design, and methodology is presented in Chapter 4. The rest of this chapter presents the specific testing scenario specifications. Furthermore, detailed statistical analysis of the collected Cognitive Workload (CW) and Situation Awareness (SA) data for each Segment using the Paired Sample T-Test or Wilcoxon Signed-Rank Test are also presented.

5.1 Scenario Description

The experiment was set in an arbitrary region of the French Alps (Figure 5.1). The scenario stated that a number of distressed signals were received from the region and a team of UAVs were required to deploy to that area and conduct identification searches of specific Items Of Interests (IOIs).

In this scenario, there were four discrete locations for each segment of the experiment, each UAV was assigned to one search location.

During transit, unexpected and hazardous events were present, where the UAVs randomly encountered these events which included three types as illustrated in Table 5.1.

There were two baseline scenarios to establish a participant's performance benchmark with a high LOD (where the highest level of aggregated information of the UAV and the environment was available in scenario Segment A) and a low LOD (where all the raw information of the UAV and the environment was available/scenario Segment C). The evaluation scenario (Segment A) enabled the participants to freely adapt the LOD of the functional capability information of the UAVs to their preferences.

During each of the scenarios, each UAV performed at a different functional LOA. Depending on the scenario, different LOD for each subsystem branch was necessary.

5.1.1 Segment A: High LOD/Min Information

Segment A presented the participant with a high LOD interface for the functional subsystems of the UAV where the participant was advised not to attempt to modify the information interface. This subsection presents the scenario breakdown structure in terms of UAV

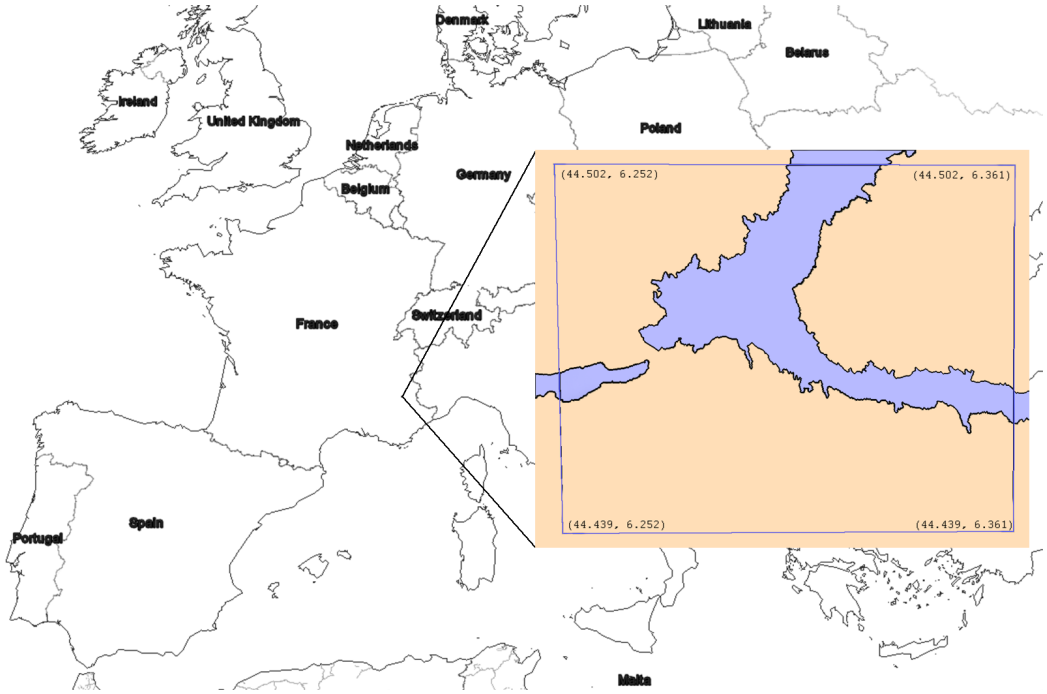


Figure 5.1: The search arena defined by the coordinates in the zoomed image

Table 5.1: Types of possible interruptions to the mission, their consequences if imappropriately handled, and their suggested method of mitigation

Event Type	Consequences	Mitigation Strategy
Hazardous Cloud/Region	Two levels of damage: Low - which enables normal UAV operations High - which causes a random lost of the asset	High LOA - UAV can plot alternate path to evade the event, Medium LOA - UAV generates possible solutions which the operator chooses the best option; Low LOA - UAV is under manual control, where the operator directs its flight path
Low Fuel Status/Level	Can cause the loss of a UAV when the fuel is completely empty	Refuelling is possible if a refuel point is available and the UAV's state subsystem can be manually controlled
Lost of Control Capability	Uncontrollable behaviour in low LOA situations	No mitigation strategy available. The situation can only be observed and observations acknowledged for possible questions in the data collection stage

Table 5.2: Experiment 1 Segment A LOA and LOD (at the commencement of the segment) configuration table for each UAV. The four columns under the LOA/LOD main column represents the four subsystems: Health, Navigation (Nav), States/Autopilot (States), and Payload

	LOA				**LOD**			
Subsystem	Health	Nav	States	Payload	Health	Nav	States	Payload
UAV_1	HA	HA	HA	PA	HL	HL	HL	HL
UAV_2	LA	HA	HA	PA	ML	HL	HL	HL
UAV_3	HA	PA	HA	PA	HL	HL	HL	HL
UAV_4	HA	HA	HA	PA	HL	HL	HL	HL

Table 5.3: Experiment 1 Segment A Search Area Specification

	Coordinates (Degress)	**Decoys**	**IOIs**	**Speed (KM/H)**
UAV_1	(44.478298, 6.259112)	4	8	400
UAV_2	(44.449976, 6.290541)	3	10	380
UAV_3	(44.477984, 6.266881)	3	8	350
UAV_4	(44.453706, 6.279397)	6	10	350

and environment configuration specifications.

Asset Functional Capability Specification

The LOA and LOD's configuration for each UAV were illustrated in Table 5.2. The functional subsystems of the UAVs included (from left to right of the table) health monitoring, navigation capability, autopilot/state autonomy, and payload operations. Three levels of LOA and LOD were applied: High LOA/LOD (HA/HL), Medium LOA/LOD (MA/ML), and Low LOA/LOD (LA/LL).

Designated Search Area Specification

Each UAV received a preassigned search area, and each search area contained a different amount of decoy items and IOIs to be found. Table 5.3 presents the central coordinates of the search areas, the amount of decoys deployed, the amount of IOIs to be found, and the required speed of the respective UAVs at arrival to the search areas.

Perturbation Event and Environmental Specification

A summary of the perturbation events seen in Segment A are presented in Table 5.4. In this segment, UAVs 1 and 4 did not experience any perturbation events, while UAV 2 encountered a low fuel status while confronting a hazardous cloud, which had a potential asset loss percentage of 60%. UAV 3 had only encountered a hazardous cloud, and it has a potential asset loss percentage of 30%.

The different functional autonomy capabilities of each UAV determined the impact of the hazards. UAV_2 had an HA LOA, hence the UAV was able to autonomously construct

Table 5.4: Experiment 1 Segment A Perturbation Event Specification

	Type	**Coordinate (Degrees)**	**Severity**	**Rotated Angle**
UAV_1	None	N/A	N/A	N/A
UAV_2	Cloud + Fuel	Cloud (44.444245, 6.281505) Fuel (44.449715, 6.280581)	60% Change of Loss, Inability to complete the first stage of searches	60° Clockwise
UAV_3	Cloud	(44.444245, 6.281505)	30% Chance of Loss	60° Clockwise
UAV_4	None	N/A	N/A	N/A

a new route which was able to both evade the hazard, as well as reach the refuel zone to refuel. UAV_3's navigation autonomy was under MBC (PA LOA), which meant that the UAV had the autonomy to generate two possible solutions to evade the hazard, while the final selection of the solution was the responsibility of the operator.

5.1.2 Segment B: Hybrid LOD/Mixed Information

Segment B presented the participant with a hybrid LOD interface for the functional subsystems of the UAV where the participant was encouraged to modify the information interface, that is to say, they were encouraged to configure the amount of details presented on the interface to their level of comfort. This subsection presents the scenario breakdown structure in terms of UAV and environment configuration specifications.

Asset Functional Capability Specification

The LOA and LOD's configuration for each UAV illustrated in Table 5.5. The functional subsystems of the UAVs include (from left to right of the table) health monitoring, navigation capability, autopilot/state autonomy, and payload operations. Three levels of LOA and LOD were applied: High LOA/LOD (HA/HL), Medium LOA/LOD (MA/ML), and Low LOA/LOD (LA/LL).

The LOD arrangement presented in Table 5.5 indicates only the configurations of each UAV at spawn. This configuration was defined by the experimenter through self-assessment where the experimenter performed the experiment under similar test conditions as the other participants several times, and the most preferred LOD arrangement was synthesised and used as a preset.

Designated Search Area Specification

Table 5.6 presents the central coordinates of the search areas, the amount of decoys deployed, the amount of IOIs to be found, and the required speed of the respective UAVs at arrival to the search areas for Segment B. The coordinates were visible regardless of whether all the UAVs was successful in completing the previous segment. That is to say,

Table 5.5: Experiment 1 Segment A LOA and LOD (at the commencement of the segment) configuration table for each UAV and its four subsystems

	LOA				LOD			
Subsystem	Health	Nav	States	Payload	Health	Nav	States	Payload
UAV_1	HA	PA	HA	PA	HL	ML	HL	ML
UAV_2	HA	HA	PA	PA	LL	ML	ML	ML
UAV_3	HA	HA	HA	HA	HL	ML	HL	ML
UAV_4	PA	HA	HA	HA	ML	HL	HL	HL

Table 5.6: Experiment 1 Segment A Search Area Specification

	Coordinates (Degrees)	**Decoys**	**IOIs**	**Speed**
UAV_1	(44.497688, 6.269603)	3	7	380
UAV_2	(44.452300, 6.329000)	5	8	350
UAV_3	(44.490324, 6.287855)	4	7	380
UAV_4	(44.452100, 6.319600)	2	8	280

all four designated search areas were visible even if its corresponding UAV was lost in the previous segment.

5.1.3 Segment C: Low LOD/Max Information

Segment C is the final segment of the experiment. It is a continued scenario from the previous two segments, which meant that any consequences from the previous two segments were carried over.

Asset Functional Capability Specification

The LOA and LOD configurations for each UAV is illustrated in Table 5.7. The three levels of LOA and LOD were applied: High LOA/LOD (HA/HL), Medium LOA/LOD (MA/ML), and Low LOA/LOD (LA/LL). During this segment, the participants were reminded not to change the LOD.

Table 5.7: Experiment 1 Segment C LOA and LOD configuration table for each UAVand its four subsystems

	LOA				LOD			
Subsystem	Health	Nav	States	Payload	Health	Nav	States	Payload
UAV_1	HA	HA	PA	PA	LL	LL	LL	LL
UAV_2	HA	HA	HA	PA	LL	LL	LL	LL
UAV_3	PA	PA	PA	PA	LL	LL	LL	LL
UAV_4	LA	HA	LA	PA	LL	LL	LL	LL

Table 5.8: Experiment 1 Segment C Search Area Specification

	Coordinates (lat, lon)	Decoys	IOIs	Speed
UAV_1	(44.495808, 6.331808)	5	9	321
UAV_2	(44.478085, 6.345580)	3	10	290
UAV_3	(44.471658, 6.337110)	2	10	387
UAV_4	(44.486652, 6.324928)	2	9	280

Table 5.9: Experiment 1 Segment C Perturbation Event Specification

	Type	Coordinate (Lat, Lon)	Severity	Rotated Angle
UAV_1	None	N/A	N/A	N/A
UAV_2	None	N/A	N/A	N/A
UAV_3	Cloud	(44.469765, 6.321810)	100% Chance of Loss	45° Anti-Clockwise
UAV_4	None	N/A	N/A	N/A

Designated Search Area Specification

Table 5.8 presented the central coordinates of the search areas, the amount of decoys deployed, the amount of IOIs to be found, and the required speed of the respective UAVs at arrival to the search areas for Segment C. The coordinates were visible regardless of whether all the UAVs were successful in completing the previous segment. That is to say, all four designated search areas were visible even if its corresponding UAV was lost in the previous segment.

Perturbation Event and Environmental Specification

A summary of the perturbation events seen in Segment C is presented in Table 5.9. In this segment, UAVs 1, 2, and 4 did not experience any perturbation events, while UAV 3 had encountered a hazardous cloud, and it has a potential asset loss percentage of 100%. This meant that the UAV would definitely be lost if it entered the hazard area.

Given the different functional autonomy capabilities of each UAV, UAV_3's navigation autonomy was under MBC (PA LOA). This meant that the UAV had the autonomy to generate two possible solutions to evade the hazard, while the final selection of the solution was the responsibility of the operator.

5.2 Result and Analysis: Cognitive Workload

The analysis method of CW (and SA)'s results had two candidates: the Repeated Measures ANOVA, and it's post hoc Paired Sample T-Test. The Repeated Measure ANOVA is applicable to statistical analyses that contains three or more distinct conditions for each test subject. The aim of the test is to understand the statistical difference as a collective sample, and in the case of this experiment, the three segments. However, this method is not applicable for this experiment as this experiment aims to perform paired comparisons

between Segment A and B, as well as B and C. For this reason, The Paired Sample T-Test is the chosen statistical test to compare the segment means.

The CW results obtained from 24 effective samples in the three segments had shown a statistical difference between Segment B and C, whereas no difference was visible between Segment A and B.

This section presents the analysis of the six dimensions of the NASA-TLX's CW and the combined CW. This is achieved by comparing the segments using the parametric Paired Sample T-Test when all the test assumptions was confidently satisfied. In some cases where parametric test assumptions are not met, the non-parametric Wilcoxon Signed-Rank Test was used. The assumptions used throughout this section yield:

- H_0: There is no significant statistical difference between the mean CW of the participants between two segments of comparison (A and B, or B and C)
- H_a: There is a significant statistical difference between the mean CW of the participants between two segments of comparison (A and B, or B and C)

The hypothesis for the *one-tailed* test states:

- H_0: The mean CW of Segment B is not less than Segment A, or segment C is not greater than Segment B ($\bar{x_A} \not> \bar{x_B}$ or $\bar{x_B} \not< \bar{x_C}$)
- H_a: The mean CW of Segment B is less than Segment A, or segment C is greater than Segment B ($\bar{x_A} > \bar{x_B}$ or $\bar{x_B} < \bar{x_C}$)

5.2.1 Mental Demand

Mental Demand (MD) describes the level of mental and perceptual activity that was required to perform the task, such as thinking, deciding, calculating, remembering, looking, searching. etc. The analysis of MD focused on the magnitude of load.

The aim was to compare the means of the MD rating collected in the three segments.

An analysis of the raw dataset revealed a negative skew was present, which influenced the normality of the data and the reliability of the outcomes produced by applying parametric tests. Due to the negative skew being only subtle, a square-transformation was used to correct the data, achieving a normal distribution.

This data was then analysed using the parametric T-Test. However, a non-parametric Wilcoxon Signed-Rank test was also applied to the original data, which supported the verdict.

The source-of-load data was intentionally left out as it did not provide much relevance when analysing temporal demand in isolation. However, detailed analysis was conducted and is presented in Section A.1.7.

Detailed data preparation and analysis is presented in appendix A. From this analysis, the means of the transformed data were compared using the Paired Sample T-Test. Table 5.10 presents a summary of the results and outcomes of the test.

Since the aim of the test was to determine the significance of the differences between each segment, a direct utilisation of the experiment's transformed data would not impact on the validity of the test.

Table 5.10: Summary table of paired sample T-Test comparing the participants' mental demand in the experiment

($\alpha = 0.05$, $n = 24$)	**Segment A & B**	**Segment B & C**
2-Tailed	$p = 0.746$, Accept H_0	$p = 0.184$, Accept H_0

The outcome of the test revealed that there were no significant differences in MD required by the participants between both segments A and B, and B and C. However, for completeness, a non-parametric Wilcoxon Signed-Rank test was conducted on the original/untransformed dataset.

In addition, the application of a non-parametric statistical test on the untransformed data was required, as it was not normally distributed. The Wilcoxon Signed-Rank test enables the testing of the means without the required assumption of the data's normal distribution; Table 5.11 illustrates the outcomes of this test based on the hypothesis defined in Appendix A.

Table 5.11: Summary table of two sample Wilcoxon Signed-Rank Test comparing the participants' mental demand in the experiment

($\alpha = 0.05$, $n = 24$)	**Segment A & B**	**Segment B & C**
2-Tailed	$p = 0.965746$, Accept H_0	$p = 0.161$, Accept H_0

The outcomes of the test revealed that there were no significant differences in MD required by the participants between both segments A and B, and B and C.

The MD data collected during the experiment had revealed to be not normal, hence, a direct application of parametric tests were not desirable, and the outcomes would not be considered reliable. Therefore, a transformation of the data was carried out.

The results illustrated in Table A.1 revealed a negative skew in its distributions, therefore a square root transformation was applied to the correct the skew for further parametric analysis of the means. Table A.2 illustrated the descriptive statistics of the transformed data which demonstrates the skewness had been corrected.

The transformed data, having passed the assumptions testing for the parametric T-Test, was analysed using the Paired-Sample T-Test. The results yielded that at Confidence Level (CL) of 95%, with a sample size of 24. One concluded that there were no significant statistical differences between the (transformed) means of Segment A and B, and B and C.

The untransformed data were analysed using the non-parametric Wilcoxon Signed-Rank test. The results also agreed with the parametric test that with a sample size of 24, at CL=95%, there were no significant statistical differences between the means of the segments.

Hence, one concluded that no significant differences of the operator's MD were experienced by the participants between the experiment's segments A and B, and segments B and C.

5.2.2 Physical Demand

The *Physical Demand (PD)* dimension describes the amount of physical effort required to perform the activity, such as pushing, pulling, turning, controlling, activating etc.. It also considers whether the task was considered easy or demanding, slow or brisk, slack or strenuous, restful or laborious.

The aim was to compare the means of the PD rating collected in the three segments and determine the statistical differences between them.

An analysis of the raw dataset revealed the distribution of the PD data in Segment A was not normal. Hence, a transformation was carried out on the data of both Segment A and B to enable paired comparison of their means using a parametric test. No transformation was required for Segments B and C as the parametric test assumptions were all satisfied.

The outcome of the tests revealed that there were no significant differences between the mean PD rating across the three segments.

The source of load data was intentionally left out as it did not provide much relevance when analysing temporal demand in isolation. However, detailed analysis was conducted and presented in Section A.1.7.

The parametric test assumptions testing were satisfied (detailed assumptions testing is presented in Appendix A) for the transformed data pair of Segment A and B. The untransformed/raw data pair of Segment B and C. Hence, the Paired Sample T-Test could be carried out. The test outcome was summarised in Table 5.12.

Table 5.12: Summary table of paired sample test comparing the means of the participants' PD in the experiment

($\alpha = 0.05$, $n = 24$)	**Segment A & B (transformed)**	**Segment B & C (original)**
2-Tailed	$p = 0.055$, Accept H_0	$p = 0.040$, Reject H_0
1-Tailed	-	$\bar{x_B} < \bar{x_C}$: Reject H_0

Table 5.12 presents the results of the hypothesis testing of the sample means. As evident from the testing outcomes, at a CL = 95% and a type-one error $\alpha = 0.05$, at a significance value $p_{A|B_{transformed}} = 0.055$, one could conclude that there was no significant differences between the PD rating of the transformed data of Segment A and B, and so H_0 was rejected. However, at the same CL and α, at a significance value $p_{B|C} = 0.040$, one could conclude that there was a significant difference between the original PD rating of Segment B and C. Furthermore, the one-tailed test revealed that the mean PD rating of Segment C was higher than that of Segment B. Therefore, H_0 was rejected. Finally, one concluded that the mean PD experienced by the participants in Segment C is significantly different and higher than that of Segment B.

5.2.3 Temporal Demand

The *temporal demand* (TD) dimension describes the time pressure felt by the participants due to the rate or pace at which the tasks or task elements occurred. The analysis of TD focused on the magnitude of load.

The aim was to compare the means of the TD rating collected in the three segments.

An initial set of the raw data revealed that there were several outliers which influenced the overall distribution of the data as well as the outcome of the analysis, therefore, another set of analysis were carried out with the removal of the outliers. The final outcome revealed a mix of results from the assumptions testing, hence, a parametric T-Test was used to evaluate the datasets with acceptable results from the assumptions testing, and non-parametric Wilcoxon Signed-Rank test to evaluate the datasets with less-than acceptable assumption results.

The source of load data was intentionally left out as it did not provide much relevance when analysing temporal demand in isolation. However, detailed analysis was conducted and presented in Section A.1.7.

The parametric test assumptions testing detailed in Appendix A were not satisfied for the data pair (Segment A and B) suggested the non-parametric/comparing of the means test was used. While the assumptions testing was satisfied for the data pair (Segment B and C). Hence, the Paired Sample T-Test could be carried out. The test outcome was summarised in Table 5.13.

Table 5.13: Summary table of paired sample test comparing the means of the participants' temporal demand in the experiment, with the outliers removed

($\alpha = 0.05$, $n = 24$)	**Segment A & B**[1]	**Segment B & C**[2]
2-Tailed	$p = 0.159$, Accept H_0	$p = 0.016$, Reject H_0
1-Tailed	-	$\bar{x_B} < \bar{x_C}$: Reject H_0

Given the assumptions of the parametric test were able to be validated for pair of comparison between Segment B and C, while the Segment A and B pair was not satisfied. Two separate comparisons-of-the-means tests were carried out to analyse the magnitude of load of the participant's TD: Non-parametric Wilcoxon Signed-Rank Test was used to compare the mean TD of Segment A and B, and the Paired Sample T-Test was used to compare the mean TD of Segment B and C.

The results as presented in Table 5.13 illustrates that with a type-one error ($\alpha = 0.05$), there was no significant statistical evidence to reject the null hypothesis for paired comparison of the means between Segment A and B. That is to say, there was an no significant difference between the mean TD experienced by the participants in Segments A when compared with Segment B.

Similarly, with a type-one error ($\alpha = 0.05$), there was a significant statistical evidence to reject the null hypothesis for paired comparison of the means between Segment B and

Table 5.14: Summary table of paired sample test comparing the participants' performance in the experiment

($\alpha = 0.05$, $n = 24$)	**Segment A & B**	**Segment B & C**
2-Tailed	$p = 0.226$, Accept H_0	$p = 0.948$, Accept H_0

C. That is to say, there was a significant difference between the mean TD experienced by the participants in Segment B when compared with Segment C. Furthermore, the one-tailed test revealed that the mean TD was greater in segment C than it was in Segment B. Therefore, H_0 was rejected.

5.2.4 Performance

The *performance* describes the participant's perceived success or self-satisfaction in accomplishing the goals of the task set by the experimenter. The analysis of *performance* focused on the magnitude of load.

The aim was to compare the means of the performance rating collected in the three segments. Given the assumptions were statistically satisfactory, a parametric T-Test was used to evaluate the magnitude of load.

The source of load data was intentionally left out as it did not provide much relevance when analysing performance in isolation. However, detailed analysis was conducted and presented in Section A.1.7.

Two pairs of comparisons of the means were conducted, and the summarised outcome is presented in Table 5.14.

Given the assumptions of the parametric T-Test were able to be validated (presented in detail in Appendix A), the analysis of the magnitude of load of the participant's *performance* was able to be carried out with a high reliability using the parametric test.

The results as presented in Table 5.14 illustrated that with a type-one error ($\alpha = 0.05$), there were no significant statistical evidence to reject the null hypothesis That is to say, there were no significant statistical evidence to suggest that there was a difference in the performance perceived by the participants through the segments of the experiment.

5.2.5 Effort

The *effort* describes how hard the participant worked (both mentally and physically) to accomplish their level of performance. The analysis of *effort* focused on the magnitude of load.

The aim of was to compare the means of the amount of effort asserted by the participants through the three segments. Given the assumptions were statistically satisfactory, a parametric T-Test was used to evaluate the magnitude of load.

The source of load data was intentionally left out as it did not provide much relevance when analysing frustration in isolation. However, detailed analysis was conducted and presented in Section A.1.7.

Two pairs of comparisons of the means were conducted, and the summarised outcome is presented in Table 5.15.

Table 5.15: Summary table of paired sample test comparing the *effort* asserted by the participant throughout the experiment

($\alpha = 0.05$, $n = 24$)	**Segment A & B**	**Segment B & C**
2-Tailed	$p = 0.349$, Accept H_0	$p = 0.133$, Accept H_0

Given the assumptions of the parametric T-Test were able to be validated in Appendix A, the analysis of the magnitude of load of the *effort* asserted was able to be carried out with a high reliability using the parametric testing of the means (T-Test).

The results as presented in table 5.15 illustrated that with a type-one error $(\alpha = 0.05)$, there was insignificant statistical evidence to reject the null hypothesis. That is to say, there was no significant statistical evidence to suggest that there was a significant difference in the *effort* asserted by the participants through the segments of the experiment.

5.2.6 Frustration

The *frustration* level describes the level of insecurity, discouragement, irritation, stress and annoyance the participant felt during the task. The analysis of frustration focused on the magnitude of load of this dimension.

The aim of was to compare the means of the level of frustration felt by the participants through the three segments, and given the assumptions were statistically satisfactory, a parametric T-Test was used to evaluate the magnitude of load using SPSS.

The source of load data was intentionally left out as it did not provide much relevance when analysing frustration in isolation. However, detailed analysis was conducted and presented in Section A.1.7.

Two pairs of comparisons of the means were conducted, and the summarised outcome were presented in table 5.16.

Table 5.16: Summary table of paired sample test comparing the level of frustration felt by the participants during the experiment

($\alpha = 0.05$, $n = 24$)	**Segment A & B**	**Segment B & C**
2-Tailed	$p = 0.852$, Accept H_0	$p = 0.113$, Accept H_0

Given the assumptions of the parametric T-Test were able to be validated (detailed analysis presented in Appendix A), the analysis of the magnitude of load of the level of frustration was able to be carried out with a high reliability using the parametric test.

The results as presented in Table 5.16 illustrated that with a type-one error $(\alpha = 0.05)$, there was no significant statistical evidence to reject the null hypothesis. This means that there was no significant statistical evidence to suggest that there was a significant difference in the level of frustration felt by the participants through the segments of the experiment.

5.2.7 Combined Cognitive Workload

The combined CW defines the overall CW score by applying formula 5.1; where SL denoted the source of load, or the weight of a CW dimension (indicated by the subscript, i.e. MD, PD etc.), which had a maximum weight of 5, ML denoted the magnitude of load, or the rating of the same dimension, which had a maximum value of 20, and $CW_{Combined}$ denoted the combined CW score in percentage. Since all the assumptions of the parametric test was satisfied with a CL of 95%, the Paired Sample T-Test was conducted to understand the impact of LOD on the human operator's CW associated with this goal.

$$CW_{Combined} = \frac{5}{15} \cdot \sum_{MD \rightarrow FR}^{attr.} (SL \cdot ML) \tag{5.1}$$

Two pairs of comparisons of the means were conducted, and the summarised outcome is presented in table 5.17.

Table 5.17: Summary table of the paired sample test comparing the combined CW of the participants in Segments A & B, and the Wilcoxon Signed Rank Test comparing the combined CW in Segments B & C

($\alpha = 0.05$, $n = 24$)	**Segment A & B**	**Segment B & C**
2-Tailed	$p = 0.428$, Accept H_0	$p = 0.036$, Reject H_0
1-Tailed	-	$\bar{x_B} < \bar{x}_C$, Reject H_0

Given the assumptions of the parametric T-Test were able to be validated (detailed process presented in Appendix A), the analysis of the combined CW was able to be carried out with a high reliability using the parametric testing of the means (T-Test) to compare Segment A and B, and the non-parametric testing of the means (Wilcoxon Signed Rank Test) to compare Segment B and C.

The results as presented in Table 5.17 illustrate that with a type-one error ($\alpha = 0.05$), there was insignificant statistical evidence to reject the null hypothesis. That is to say, there was no significant statistical evidence to suggest that there was a significant difference in the combined CW in Segment A compared to Segment B. However, there was in fact a significant statistical difference in operator CW between Segment B and C.

Through further analysis using the one-tailed test, the null hypothesis (H_0) was rejected; the CW experienced by the operator in Segment B was lower than that experienced in Segment C.

5.2.8 Analysis Summary

This section presents the in depth analysis of the six attributes that form CW as defined by Hart *et al.* [97]. The six attributes were analysed individually as well as jointly to illustrate the differences in the participant's CW when different LODs were presented.

Table 5.18: A summary of the results obtained from the testing of the mean CW ($\sigma = 0.472$) between the Segment A (high LOD) and Segment B (hybrid LOD).

CL=**95%**,$\alpha = 0.05$	**Outcome**	$\bar{x}_A$	$\bar{x}_B$
Mental Demand	No Different ($p = 0.746$)	66.45%	66.25%
Physical Demand	No Different ($p = 0.055$)	25.4%	35%
Temporal Demand	No Different ($p = 0.159$)	70.2%	63.35%
Performance	No Different ($p = 0.226$)	46.65%	52.90%
Effort	No Different ($p = 0.349$)	52.10%	55.65%
Frustration	No Different ($p = 0.852$)	42.5%	43.4%
Combined CW	No Different ($p = 0.428$)	58.50%	59.42%

Table 5.19: A summary of the results obtained from the testing of the mean CW ($\sigma = 0.472$) between the Segment B (hybrid LOD) and Segment C (low LOD).

$CL = 95\%$, $\alpha = 0.05$	**Outcome**	$\bar{x}_B$	$\bar{x}_C$
Mental Demand	No Different ($p = 0.184$)	66.25%	70.85%
Physical Demand	SA Lower in Segment B ($p = 0.040$)	35%	42.5%
Temporal Demand	SA Lower in Segment B ($p = 0.016$)	63.35%	70.85%
Performance	No Different ($p = 0.948$)	52.90%	52.70%
Effort	No Different ($p = 0.133$)	55.65%	61.25%
Frustration	No Different ($p = 0.113$)	43.4%	49.6%
Combined CW	***SA Lower in Segment B*** ($p = 0.036$)	59.42%	66.63%

Table 5.18 presents a summarised table of all the statistics along with specific details and outcomes in the comparative pair of Segment A and B, where *desirable* denotes that the magnitude of load was greater during the baseline experiment.

Evidently from Table 5.18, the ability for the participant to adjust the elements in the FCF visualisation did not show significant statistical differences in reducing CW levels. However, on close inspection of the results, although the remaining attributes were not tested to have a significant statistical evidence, the direction of the mean magnitude of loads indicated that the loads were mostly higher in Segment A. This is because a high LOD of system information was visually represented to the participants, while the loads were lower in the configuration where the participants were able to adjust the LOD for each of the subsystem branches.

Table 5.19 presents a summarised table of all the statistics along with specific details and outcomes in the comparative pair of Segment B and C.

As presented in Table 5.19, the ability for the participant to adjust the elements in the FCF visualisation showed significant statistical differences in reducing CW levels when comparing to an interface with a large volume of FCF information (low LOD). Furthermore, on close inspection of the results, although the remaining attributes were not tested to have significant statistical evidence. The direction of the mean magnitude of loads indicated that the loads were mostly higher in Segment C, where a low LOD of system information was

visually represented to the participants, while the loads were lower in the configuration where the participants were able to adjust the LOD for each of the subsystem branches.

From the CW result collected during this experiment, the segment with the interface configured to enable an adjustable LOD of the FCF subsystem branches had shown to be lower when compared to a low LOD configuration. No significant statistical reduction in CW were observed when comparing to Segment A (high LOD).

5.3 Result and Analysis: Situation Awareness

The SA results obtained from the three segments were shown to be in-conclusive. That is, the results were not significantly different between the LOD configurations between the segments.

This section presents the SA scoring methods and results from the experiment. The general testing of the means assumptions testing (detail presented in Appendix A), and the outcome from the statistical tests. The results were analysed using the software statistical package SPSS. The hypothesis related to SA are:

- H_0: There is no significant statistical difference between the mean SA of the participants between two segments of comparison (A and B, or B and C)
- H_a: There is a significant statistical difference between the mean SA of the participants between two segments of comparison (A and B, or B and C)

The hypothesis for the *one-tailed* test states:

- H_0: The mean SA of Segment B is not greater than Segment A, or segment C is not greater than Segment A ($\bar{x_A} \ngtr \bar{x_B}$ or $\bar{x_B} \nless \bar{x_C}$)
- H_a: The mean SA of Segment B is greater than Segment A, or segment B is greater than Segment C ($\bar{x_A} < \bar{x_B}$ or $\bar{x_B} > \bar{x_C}$)

5.3.1 Scoring Method

Each SA questionnaire incorporated seven questions. The number of questions were decided due to the intra-segment time constraint. Each question contains a maximum of one point, where a full mark is awarded to responses that contained all correct answers, half mark (0.5) is awarded to responses that contained partial answers.

The SA of each participant is scored in a percentage, where a "100%" SA score is achieved when all answers on their SA questionnaire is correct. The percentage is determined by the total number of marks awarded from their responses as a portion of seven (the total score). From this, SA performance analysis is further conducted.

Table 5.20: Summary table of paired sample T-Test (Segment A & B) and the Wilcoxon Signed Rank Test (Segment B & C) comparing the respective combined SA

($\alpha = 0.05$, $n = 23$)	**Segment A & B**	**Segment B & C**
2-Tailed	$p = 0.922$, Accept H_0	$p = 0.417$, Accept H_0

5.3.2 Analysis

The overall SA analysis provided an indication of the SA experienced by the participants during the three segments. Two forms of statistical tests were conducted to understand the impact of LOD on the human operator's SA: The non-parametric Wilcoxon Signed-Rank Test for the high LOD segment (Segment A) and the hybrid segment (Segment B) pair, and the parametric T-Test for hybrid LOD segment (Segment B) and the low LOD segment (Segment C) pair.

The two pairs of mean SA comparisons are presented in Table 5.20.

The outcomes of the respective tests revealed that the operator SA across all three LOD configurations did not exhibit any significantly negative impact. Hence, the results are inconclusive.

5.4 Discussion and Conclusion

This chapter presented the aim of Experiment 1, the detailed scenario design and the analysis of the collected results of an operator's cognitive workload and situation awareness performance.

A Paired Sample T-Test was conducted to compare the CW and SA of a human operator managing multiple UAVs in high LOD and hybrid LOD configurations, as well as in low LOD and hybrid LOD configurations. There was a significant reduction in operator CW in the hybrid LOD ($\bar{x}_B = 59.42\%$, $\sigma = 0.472$) configurations and low LOD ($\bar{x}_B = 66.63\%$, $\sigma = 0.472$) configurations. There was no significant difference in the CW of the hybrid LOD ($\bar{x}_B = 59.42\%$, $\sigma = 0.472$) configurations and high LOD ($\bar{x}_A = 58.5\%$, $\sigma = 0.472$) configurations: $N = 23$, $p = 0.036$. Also, there was no significant differences in the operator SA throughout the three LOD configurations: $N = 23, p_{A|B} = 0.922$ and $p_{B|C} = 0.417$.

These results suggested that a positive cognitive performance was demonstrated in the hybrid application of the FCF in the multiple heterogeneous UAV management context. As a result, Contribution 1 is made and Research Question 1 is answered.

CHAPTER 6

Experiment 2: Partial Autonomy Transparency

This experiment aims to evaluate the influence of autonomy transparency through interface graphical feedback on the operator's cognitive performance. Details of this experiment overview, design, and methodology is presented in Chapter 4. The rest of this chapter presents the specific testing scenario specifications and a detailed analysis of the collected Cognitive Workload (CW) and Situation Awareness (SA) data for each scenario, as well as some general qualitative feedback about the participant experience and preferences between the scenarios.

6.1 Scenario Description

This experiment was set in the same location as the first experiment. The scenarios include four UAVs which participate in a search mission. However, one of the primary differences is that the evaluation scenario in this experiment was not a continuation of the baseline scenario like the first experiment, they were two discrete scenarios which had similar difficulties.

The baseline scenario evaluated the cognitive performance of each individual participant compared against the evaluation scenario. This scenario had a very minimal amount of transparency of the navigation autonomy communicated (visually) to the participants, which simulates a limited transparency of the UAVs' autonomy capabilities, setting the benchmark for the evaluation trial to be compared against.

The evaluation scenario benchmarks the cognitive performance of each individual participant for comparison against the baseline results. This scenario increases the navigation autonomy's transparency through the display flight plans and paths of each UAV. Hence, it establishes a communication channel of the UAV's autonomy capability and intentions for future actions to the participant.

The differences in the number of events and IOIs between the two scenarios were intentional to promote realism. Since the participants were made aware the potential range of the number of perturbations and IOIs, the ambiguity is expected.

Table 6.1: Types of possible interruptions to the mission, their consequences if inappropriately handled, and their suggested method of mitigation

Event Type	Consequences	Mitigation Strategy
Hazardous Cloud/Region	Causes the loss of the UAV that comes in contact with the hatched region	High LOA - UAV can plot alternate path to evade the event, Medium LOA - UAV generates possible solutions which the operator chooses the best option; Low LOA - UAV is under manual control, where the operator directs it's flight path
Low Fuel Status/Level	Can cause the loss of a UAV when the fuel is completely empty	Refuelling is possible if a refuel point is available and the UAV's state subsystem can be manually controlled

6.1.1 Tasks and Objectives

Each experimental trial was broken down and viewed in two stages, the transit stage and the search stage. The transit stage was defined from the initial moment where the trial launched to the moment where the UAVs entered the search area and the search operation started. This stage was were the perturbation events were scripted to take place. There were two types of events that the participants expected: The hazardous cloud/region, and the low fuel level status. Table 6.1 presents the consequences of the events if they were unattended to, and the mitigation strategy required to overcome the events. During this stage, the participants were required to ensure the safe flight of the assets from the initial launch position to the next stage (the outer bounds of the search zone).

The search stage was defined from the UAVs' outer boundary of the designated search zone to the conclusion of the trial. The participant's task was to ensure that all IOIs were identified and found, and all decoys were left untouched. For each of the experiments, there were different scripted events and flight paths for each UAV.

6.1.2 Baseline Scenario (Opaque Autonomy Transparency)

The baseline scenario set a comparative benchmark for the cognitive performance measures described in Section 4.4.6. The baseline scenario presented the participant with an opaque autonomy transparency of the UAVs' navigation capability. This meant that the flight path and trajectory were hidden from the participant's view throughout the trial as illustrated in Figure 6.1.

Asset Functional Capability Specification

The LOA and LOD configurations for each UAV were illustrated in Table 6.2. The functional subsystems of the UAVs included (from left to right of the table) health monitoring,

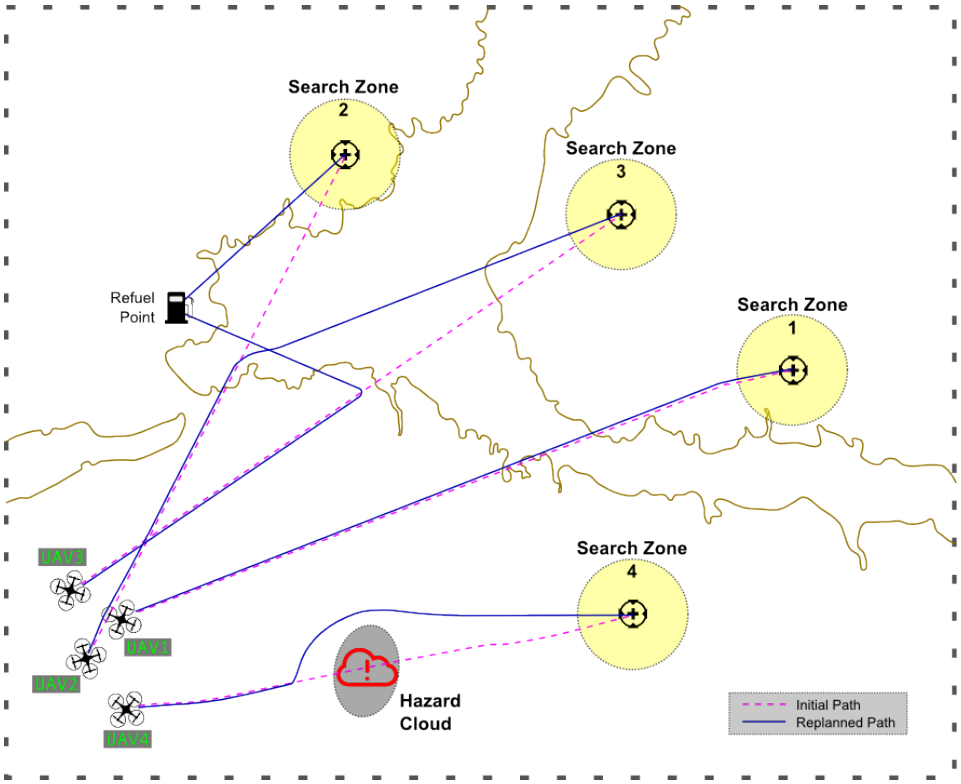


Figure 6.1: Planned layout of the baseline scenario (not to scale)

Table 6.2: Experiment 2 Baseline Scenario LOA and LOD (at the commencement of the trial) configuration table for each UAV and its four subsystems

	LOA				LOD			
Subsystem	Health	Nav	States	Payload	Health	Nav	States	Payload
UAV_1	HA	HA	PA	HA	HL	HL	ML	LL
UAV_2	LA	HA	HA	LA	LL	HL	ML	ML
UAV_3	HA	PA	HA	PA	HL	HL	HL	ML
UAV_4	HA	LA	HA	PA	ML	HL	ML	LL

navigation capability, autopilot/state autonomy, and payload operations. Three levels of LOA and LOD were applied: High LOA/LOD (HA/HL), Medium LOA/LOD (MA/ML), and Low LOA/LOD (LA/LL). The designated LOAs were not applicable to change by the participant, while the LOD could be changed as desired by the participant.

Designated Search Area Specification

Each UAV received a designated search area, and each search area contained a different number of decoy items and IOIs to be found. Table 6.3 presents the central coordinates of these search areas, the amount of decoys deployed, the amount of IOIs to be found, and the required speed of the respective UAVs upon arrival to the search areas.

Table 6.3: Experiment 2 Baseline Scenario Search Area Specification

	Coordinates (Degrees)	Decoys	IOIs	Speed (KM/H)
UAV_1	(44.472111, 6.343017)	4	8	300
UAV_2	(44.485097, 6.322871)	3	8	290
UAV_3	(44.490383, 6.290100)	3	10	340
UAV_4	(44.452123, 6.323987)	6	10	340

Table 6.4: Experiment 2 Baseline Scenario Perturbation Event Specification

	Type	Coordinate (Degrees)	Severity	Rotated Angle
UAV_1	None	N/A	N/A	N/A
UAV_2	None	N/A	N/A	N/A
UAV_3	Fuel	Fuel Zone (44.457524, 6.337058)	100% Chance of Loss on fuel starvation	60° Clockwise
UAV_4	Cloud	(44.469409, 6.274197)	100% Chance of Loss on Contact	60° Clockwise

Perturbation Event and Environmental Specification

A summary of the perturbation events seen in the baseline scenario are presented in Table 6.4. In this scenario, UAVs 1 and 2 did not experience any perturbation events, while UAV 3 encountered a low fuel status upon launch and UAV 4 confronted a hazardous cloud, which had a potential asset loss percentage of 100%. This meant that any asset that came in contact with the cloud will be lost.

Given that UAVs 3 and 4 (the UAVs which encountered perturbation events) had a high LOA, they were capable of autonomously replanning their flight path to avoid catastrophe. UAV 3 rerouted to the fuel station and redirected to a new search zone that was closer to its location. UAV 4 rerouted around the threat. However, due to the restricted transparency of the autonomy, these behaviours were not directly displayed to the participant. Even though the UAVs autonomously reacted to the threat, and the participants were able to observe a change in the UAVs' behaviour.

6.1.3 Evaluation Scenario (Transparent Autonomy Transparency)

The evaluation scenario produced a set of cognitive performance results for comparison with the measures collected in the baseline scenario. The evaluation scenario presented the participant with a transparent autonomy capability of the UAVs' navigation capability, this meant that the flight path and trajectory were displayed to the participant's view throughout the trial as illustrated in Figure 6.2.

Asset Functional Capability Specification

The LOA and LOD configurations for each UAV are illustrated in Table 6.5. The functional subsystems of the UAVs include (from left to right of the table) health monitoring, navig-

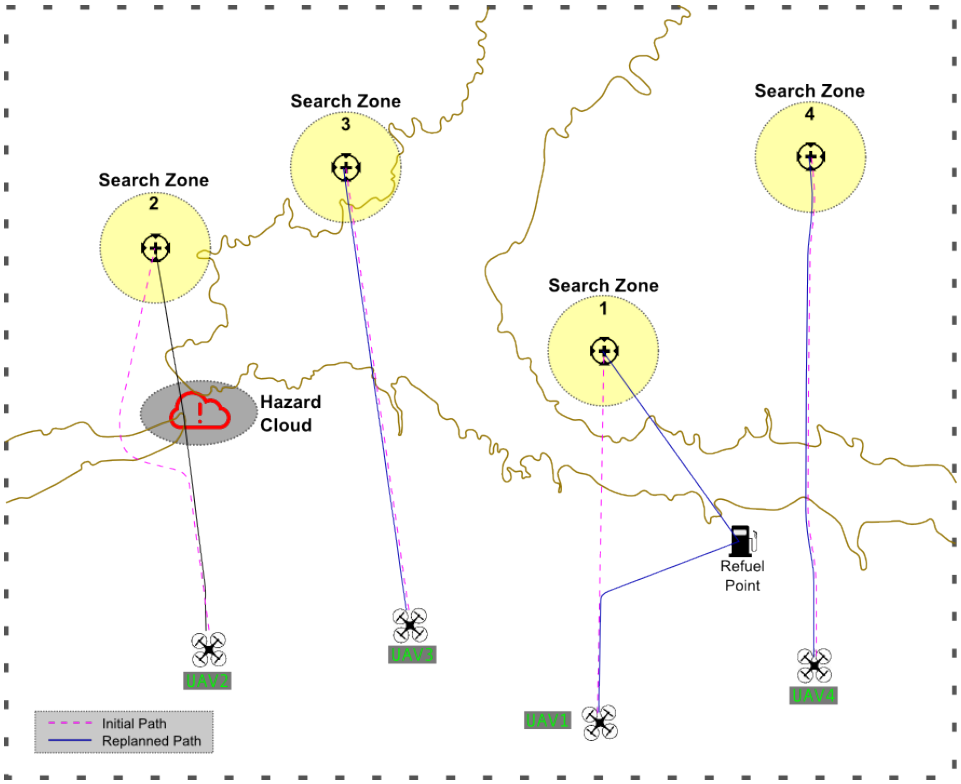


Figure 6.2: Planned layout of the evaluation scenario (not to scale)

Table 6.5: Experiment 2 Evaluation Scenario LOA and LOD (at the commencement of the trial) configuration table for each UAV and its four subsystems

	LOA				LOD			
Subsystem	Health	Nav	States	Payload	Health	Nav	States	Payload
UAV_1	HA	HA	HA	HA	HL	LL	HL	LL
UAV_2	LA	HA	HA	LA	LL	LL	HL	ML
UAV_3	HA	PA	HA	PA	HL	LL	HL	ML
UAV_4	HA	LA	HA	LA	ML	LL	HL	LL

ation capability, autopilot/state autonomy, and payload operations. The designated LOAs, although very similar to the configuration for the baseline scenario presented in Section 6.1.2, were not able to be altered by the participant, while the LOD could be changed as desired by the participant.

Designated Search Area Specification

Each UAV received a designated search area, and each search area contained a different number of decoy items and IOIs to be found. Table 6.6 presents the central coordinates of these search areas, the amount of decoys deployed, the amount of IOIs to be found, and the required speed of the respective UAVs upon arrival to the search areas.

Table 6.6: Experiment 2 Evaluation Scenario Search Area Specification

	Coordinates (Degrees)	Decoys	IOIs	Speed (KM/H)
UAV_1	(44.474277, 6.321760)	5	7	400
UAV_2	(44.483293, 6.267751)	3	8	380
UAV_3	(44.489976, 6.290896)	3	9	350
UAV_4	(44.490330, 6.345762)	2	7	350

Table 6.7: Experiment 2 Evaluation Scenario Perturbation Event Specification

	Type	Coordinate (Degrees)	Severity	Rotated Angle
UAV_1	Cloud	(44.469409, 6.274197)	100% Chance of Loss on Contact	60° Clockwise
UAV_2	Fuel	Fuel Zone (44.457524, 6.337058)	100% Chance of Loss on fuel starvation	N/A
UAV_3	None	N/A	N/A	N/A
UAV_4	Fuel	N/A	100% Chance of Loss on fuel starvation	N/A

Perturbation Event and Environmental Specification

A summary of the perturbation events seen in the evaluation scenario are presented in Table 6.7. In this scenario, UAV 3 did not experience any perturbation events. UAV 1 encountered a low fuel status upon launch. UAV 2 confronted a hazardous cloud, which had a potential asset loss percentage of 100%. And UAV 4 also encountered a low fuel status but could not be saved and a forced landing was the result.

UAVs 1, 2, and 4 (the UAVs which encountered perturbation events) had a high LOA, which meant that they were capable of autonomously replanning their flight path to avoid or minimise catastrophe. UAV 1 rerouted to the nearest refuel zone. UAV 2 rerouted around the threat. UAV 4 was not able to safely reach a refuel zone, and was rerouted to a safe location to be ditched. Since the autonomy was transparent in this scenario, all the UAVs' flight paths were visually presented to the participants throughout the trial, enabling the participants to acquire the most up-to-date information related to the UAVs.

6.2 Result & Analysis: Cognitive Workload

The CW results obtained from the two scenarios had shown evidence of alignment with the initial hypothesis of this experiment presented in Section 4.1.

This section presented the analysis of the six dimensions of the NASA-TLX's CW individually and the combined CW. This is achieved comparing the baseline and the evaluation scenario results using the parametric Paired Sample T-Test when all the test assumptions was confidently satisfied. In some cases where parametric test assumptions are not met,

the non-parametric Wilcoxon Signed Rank Test was used. The hypothesis used in this section follows the follow sets:

- H_0: There is no significant statistical difference between the mean CW of the participants between the two scenarios (baseline and evaluation)
- H_a: There is a significant statistical difference between the mean CW of the participants between the two scenarios (baseline and evaluation)

The hypothesis for the *one-tailed* test states:

- H_0: The mean CW of the Evaluation Scenario is not less than the mean CW of the Baseline Scenario ($\bar{x_{Baseline}} \not< \bar{x_{Evaluation}}$)
- H_a: The mean CW of the Evaluation Scenario is less than the mean CW of the Baseline Scenario ($\bar{x_{Baseline}} < \bar{x_{Evaluation}}$)

6.2.1 Mental Demand

An analysis of the raw *Mental Demand (MD)* dataset from both scenarios revealed a minor negative skew (Table B.1 Appendix B), which was corrected using a square transformation. The transformed data presented a set of analysis which satisfied the requirements of the assumptions testing. Hence, a T-Test was used.

The outcome from the test revealed that there was insignificant difference between the mental demand of the baseline and the evaluation experiment.

A paired comparison of the means was conducted on the transform of the original data (the process of the assumptions testing is presented in Appendix B). The application of the square transform of the data enabled the data set to satisfy all the assumptions of the T-Test.

Table 6.8: Summary table of paired sample test comparing the magnitude of mental demand/load felt by the participants during the experiment

($\alpha = 0.05$, $n = 43$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.263$, Accept H_0

Table 6.8 presents the test summary of the parametric T-Test. From the table, with a Confidence Level (CL) = 95% (type-one error $\alpha = 0.05$), there was no significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was no significant difference between the mean mental demand of the baseline and evaluation scenario.

The magnitude of mental demand data presented to be not normally distributed (Appendix B), as such, a parametric test could not be carried out. The test descriptives revealed that the baseline scenario's dataset was negatively skewed, resulting in a non-normal distribution. A square transform was applied to both sets of data (Table B.2 Appendix B) which was retested for normality and skewness.

Post transformation revealed that the new data set had satisfied the normality tests numerically and graphically, as well as being able to satisfy the remaining assumptions to enable an accurate parametric testing of the means. Therefore, applying the parametric T-Test revealed that, with a CL = 95%, the magnitude of the mental demand results of the baseline scenario was not significantly different from the evaluation scenario.

6.2.2 Physical Demand

From the *Physical Demand (PD)* dataset's assumptions testing presented in Appendix B, a paired comparison of the means was conducted on the transform of the original data. The application of the square transform of the data enabled the data set to satisfy all the assumptions of the T-Test.

Table 6.9: Summary table of paired sample test comparing the magnitude of physical demand felt by the participants during the experiment

($\alpha = 0.05$, $n = 43$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.931$, Accept H_0

Table 6.9 presents the test summary of the parametric T-Test. From the table, with a CL = 95% (type-one error $\alpha = 0.05$), there was no significant statistical evidence to reject the null hypothesis (H_0), meaning that there was no significant difference between the mean physical demand of the baseline and evaluation scenario.

The original magnitude of physical demand data appeared to not be normally distributed (Table B.3). Hence, a parametric test could not be carried out. As the test descriptives revealed, the data set presented to be positively skewed, resulting in a non-normal distribution of data. A square-root transform was applied to repair the dataset (Table B.4), which was again retested for normality and skewness.

Post transformation revealed that the new dataset had satisfied the normality test,and satisfied the remaining assumptions to enable an accurate parametric testing of the means. Therefore, applying the parametric T-Test revealed that, with a CL = 95%, the magnitude of the physical demand results of the baseline scenario was not significantly different from the evaluation scenario.

6.2.3 Temporal Demand

An initial set of the raw *Temporal Demand (TD)* data appeared to be negatively skewed. As such, a square transform was applied. After the transform, the distribution of the transformed data was still not normally distributed. Hence, a non-parametric test was used to evaluate the data set.

The source of load data was intentionally left out as it did not provide sufficient relevance when analysing temporal demand in isolation. The full assumptions testing process is presented in Appendix B.

From the assumptions testing (Appendix B), the non-parametric paired comparison of the means was used on the original and the transformed data.

Table 6.10: Summary table of non-parametric Wilcoxon Signed-Rank test comparing the original and the transformed temporal demand data set

$(\alpha = 0.05, n = 42)$	**Baseline & Evaluation Scenario**
Original Data Set	$p = 0.113$, Accept H_0
Transformed Data Set	$p = 0.171$, Accept H_0

Table 6.10 presents the test summary of both datasets. This confirmed the test verdict. From the table, with a CL = 95% (type-one error $\alpha = 0.05$), there was no significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was no significant difference between the mean TD rating of the baseline and evaluation scenario.

As the results showed, with a CL = 95%, the TD magnitude of load of the baseline scenario was not significantly different from that of the evaluation experiment. This suggested that the TD required to perform the tasks was not significantly different when there was a low level of autonomy transparency in the system as compared to a system with a high level of autonomy transparency, and in the context of this experiment, the autonomy was referred to as the UAV's autonomous navigation capability.

6.2.4 Performance

The *performance* (or perceived performance) data was tested to be normally distributed, and along with the positive outcomes of the other assumptions for a parametric to be performed.

The source of load data was intentionally left out in the analysis of this attribute as it did not provide much relevance when analysing temporal demand in isolation. From the outcomes of the assumptions testing presented in Appendix B, a parametric test to compare the means was used.

Table 6.11: Summary table of parametric paired-sample test comparing the mean perceived performance rating of the baseline and the evaluation experiment

$(\alpha = 0.05, n = 42)$	**Baseline & Evaluation Scenario**
Two-Tailed	$p = 0.276$, Accept H_0

Table 6.11 presents the test summary of the dataset. At CL = 95% (type-one error $\alpha = 0.05$), there was no significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was no significant difference between the mean perceived performance rating of the baseline and evaluation scenario.

As the results showed, with a CL = 95% and a significance value ($p - value$) of 0.276, the perceived performance magnitude of load of the baseline scenario was not signific-

antly different from that of the evaluation experiment. This suggested that the perceived performance required to perform the tasks was not significantly different when there was a low level of autonomy transparency in the system as compared to a system with a high level of autonomy transparency. In the context of this experiment, the autonomy was referred to as the UAV's autonomous navigation capability.

6.2.5 Effort

An analysis of the raw *effort* data set from the baseline scenario revealed that the dataset was not normally distributed, while the data set collected from the evaluation experiment was. Furthermore, a negative skew was presented which had contributed to the non-normal distribution of the data. Hence, a square transformation was applied to correct the data.. The transformed data presented a set of analysis which met the assumptions needed to perform a parametric test of the means. As such, a Paired Sample T-Test was used.

The outcome from the test revealed that there was a significant difference between the effort required to perform the tasks in the baseline scenario and that of the evaluation experiment. From the outcomes assumptions testing presented in Appendix B, a paired comparison of the means was conducted on the transform of the original data. The application of the square transform of the data landed the data set to satisfy all the assumptions of the T-Test.

Table 6.12: Summary table of paired sample test comparing the magnitude of effort/load felt by the participants during the experiment

($\alpha = 0.05$, $n = 42$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.032$, Reject H_0
1-Tailed	$\bar{x}_{baseline} > \bar{x}_{evaluation}$, Reject H_0

Table 6.12 presents the test summary of the parametric T-Test. From the table, with a CL = 95% (type-one error $\alpha = 0.05$), there was significant statistical evidence to reject the null hypothesis (H_0). This means that there was a significant difference between the mean effort required to perform the tasks in the baseline and the evaluation scenario. Furthermore, the one-tailed test revealed that there were significant statistical evidence to reject the null hypothesis (H_0), and accept that the effort required to perform the baseline scenario tasks were significantly higher than the effort required to perform the evaluation scenario tasks.

The original data of the magnitude of effort required to perform the tasks presented to be not normally distributed, hence a parametric test could not be carried out. As the test descriptives revealed, the baseline scenario's data set presented to be negatively skewed, hence resulting in a non-normal distribution of data. A square transform was applied to both sets of data to repair the data set, which was again retested for normality and skewness.

Post transformation revealed that the new dataset had satisfied the normality tests nu-

merically and graphically, as well as able to meet the remaining assumptions to enable an accurate parametric testing of the means. Applying the parametric T-Test revealed that the magnitude of the required effort in the baseline scenario was significantly higher than the effort required to perform the tasks in the evaluation scenario. And in the context of this experiment, the transparency in the UAVs' navigation autonomy contributed positively to the operator's effort to perform tasks.

6.2.6 Frustration

An initial set of the raw *frustration* level data revealed to be not normal with an acceptable skewness. Due to the wide distribution and randomness of the data, no transforms could be applied to sanitise the data set such that all the assumptions for conducting a parametric test could be met. However, the assumptions of the non-parametric test of the means were met, hence, the non-parametric test was used to evaluate the dataset.

The source of load data was intentionally left out as it did not provide much relevance when analysing temporal demand in isolation. From the outcomes assumptions testing presented in Appendix B, a non-parametric paired comparison of the means was conducted and the conclusion was reached.

Table 6.13: Summary table of non-parametric Wilcoxon Signed-Rank test of the frustration level dataset

$(\alpha = 0.05, n = 42)$	**Baseline & Evaluation Scenario**
Original Data Set	$p = 0.172$, Accept H_0

Table 6.13 presents the test summary of the frustration dataset. This confirmed the test verdict. From the table, with a CL = 95% (type-one error $\alpha = 0.05$), there was no significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was no significant difference between the mean frustration level experienced by the participant in the baseline and evaluation scenario.

As the results reflected, the frustration magnitude of load of the baseline scenario was not significantly different from that of the evaluation scenario. This suggested that the frustration felt during the scenarios did not vary significantly when there was a low level of autonomy transparency in the system as compared to a system with a high level of autonomy transparency. In the context of this experiment, the autonomy was referred to as the UAV's autonomous navigation capability.

6.2.7 Combined Cognitive Workload

The combined CW defines the overall CW score by applying formula 5.1. The aim was to compare the means of the combined CW score between the baseline and the evaluation scenario.

An analysis of the raw dataset from both scenarios revealed a negative skew (Table B.11), which was corrected using a square transformation. The set of transformed data

was analysed and revealed a set of outlier data points (Table B.12), which was removed for further analysis. From the outcomes assumptions testing presented in Appendix B, a paired comparison of the means was conducted on the transform of the original data, which presented to be negative skewed. The application of the square transform of the data followed by the removal of a pair of outliers landed the dataset to satisfy all the assumptions of the T-Test.

Table 6.14: Summary table of paired sample test comparing the combined CW felt by the participants during the experiment

$(\alpha = 0.05, n = 42)$	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.033$, Reject H_0
1-Tailed	$\bar{x}_{baseline} > \bar{x}_{evaluation}$, Reject H_0

Table 6.14 presents the test summary of the parametric T-Test. From the table, with a CL = 95% (type-one error $\alpha = 0.05$), there was a significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference between the mean combined CW rating of the baseline and evaluation scenario.

The further investigation using a one-tailed test revealed that at CL = 95%, there was significant statistical evidence to reject H_0. That is to say, the combined CW of the baseline scenario was greater than that of the evaluation scenario.

As the results showed, the combined CW of the baseline scenario was significantly higher than the evaluation experiment. This suggested that the combined CW was significantly higher when there was a low level of autonomy transparency in the system when compared to a system with a high level of autonomy transparency. In the context of this experiment, the autonomy was referred to as the UAV's autonomous navigation capability.

6.2.8 Analysis Summary

This section presents the in depth analysis of the six attributes that form CW as defined by Hart *et al.*. The six attributes were analysed individually and as a group to illustrate the differences in the participant CW when the interface configuration included the UAV's autonomy status versus the configuration where they were absent.

Table 6.15 presents a summarised table of all the statistics along with specific details and outcomes, where *desirable* denotes that the magnitude of load was greater during the baseline experiment.

Table 6.15: A summary of the results obtained from the testing of the mean CW ($\sigma = 0.365$) between the opaque navigation autonomy and the transparent navigation autonomy interface configurations.

$\alpha = 0.05$	**Outcome**	$\bar{x}_{baseline}$	$\bar{x}_{evaluation}$
Mental Demand	No Different ($p = 0.263$)	13.31 ($\sigma = 3.5$)	12.69 ($\sigma = 3.9$)
Physical Demand	No Different ($p = 0.931$)	4.548 ($\sigma = 4.0$)	4.768 ($\sigma = 4.1$)
Temporal Demand	No Different ($p = 0.113$)	11.33 ($\sigma = 4.6$)	10.4 ($\sigma = 5.1$)
Performance	No Different ($p = 0.276$)	9.0 ($\sigma = 2.8$)	8.33 ($\sigma = 3.8$)
Effort	*Desirable ($p = 0.032$)*	13.31 ($\sigma = 4.0$)	12.69 ($\sigma = 4.4$)
Frustration	No Different ($p = 0.172$)	7.83 ($\sigma = 4.8$)	7.1 ($\sigma = 5.0$)
Combined CW	**Desirable ($p = 0.033$)**	55.86 ($\sigma = 14$)	51.95 ($\sigma = 15$)

Evidently from Table 6.15, the ability for the participant to visualise the UAV's navigation autonomy, presented in a graphical form, is significantly reduced the operator's effort requirement and the overall CW levels. On close inspection of the results, although the remaining attributes were not tested to have significant statistical evidence, the direction of the mean magnitude of loads indicated that the loads were mostly higher in interface configuration where the UAVs' navigation autonomy were not graphically illustrated to the participants.

6.3 Result & Analysis: Situation Awareness

The SA results obtained from the experiment showed positive outcome towards the scenario with a higher degree of automation transparency. This reflected the scenario which the participants had greater information about their assets' navigation autonomy resulted in a greater operator SA.

This section presented the statistical tests from the experiment in three directions; the Level 1 SA, the Level 2 SA, and the overall SA. The results were recorded as a percentage of the correct responses. The hypothesis for the *two-tailed* test states:

- H_0: There was no significant statistical difference between the mean SA between the scenarios
- H_a: There was a significant statistical difference between the mean SA between the scenarios

The hypothesis for the *one-tailed* test states:

- H_0: Mean SA in the baseline scenario is higher than evaluation scenario
- H_a: Mean SA in the baseline scenario is not higher than evaluation scenario

6.3.1 Level 1 SA

Level 1 SA was described as the *perception of environmental elements with respect to time or space* [76], or the subject's perception of the situation. In the context of this experiment,

the perception was the participant's awareness of the mission situation and the asset's navigation capability.

Analysis focused on the responses to the questions that were designed to be assessing the participant's perception of the situation in percentage.

The assumptions testing process is presented in Appendix B, and the outcome suggests that a Wilcoxon Signed-Rank Test is to be used for analysis. The assumptions of the test are the same as the parametric test assumptions with the removal of the normality distribution requirement. Hence, H_0 of the variance was accepted and the remaining assumptions were also met, the paired comparison of the means was conducted. The summarised outcome is presented in Table 6.16.

Table 6.16: Summary table of paired sample comparing the Level 1 SA of the baseline and the evaluation scenario using the Wilcoxon Signed-Rank Test

($\alpha = 0.05, n = 42$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.086$, Accept H_0

As presented in Table6.16, there was no significant difference in the participant's mean Level 1 SA between the baseline and evaluation scenario. The participants experienced similar levels of situational perception when there were a greater level of autonomy transparency compared to lower levels of transparency.

The SA data collected for the participants Level 1 SA could not be analysed using a parametric test due to the assumption of normality which could not be met. Hence, the non-parametric hypothesis testing of means, the Wilcoxon Signed-Rank test, was used. The collected level 1 SA data revealed that at CL = 95%, there was no significant difference between the level of perception in a scenario where the UAVs' autonomy transparency were greater or lesser.

6.3.2 Level 2 SA

Level 2 SA was described as the *comprehension of environmental elements with respect to time or space* [76], or the subject's understanding of the situation after they had perceived the event. In the context of this experiment, the comprehension was the participant's understanding of the mission situation and the asset's navigation capability.

Analysis focused on the responses of the questions that were designed to be assessing the participant's comprehension of the situation in percentage.

The assumptions testing process is presented in Appendix B, and the outcome suggested that the Wilcoxon Signed-Rank Test is to be used to perform the hypothesis testing of the means. The assumptions of the test was the same as the parametric test assumptions with the removal of the normality distribution requirement. Hence, the H_0 of the variance was accepted and the remaining assumptions were also met, the paired comparison of the means was conducted, and the summarised outcome was presented in Table 6.17.

Table 6.17: Summary table of paired sample comparing the (square-root) transformed the Level 2 SA data of the baseline and the evaluation scenario using the Wilcoxon Signed-Rank Test

$(\alpha = 0.05, n = 42)$	**Baseline & Evaluation Scenario**
2-Tailed	$p < 0.001$, Reject H_0
1-Tailed	$\bar{x}_{baseline} < \bar{x}_{evaluation}$, Reject H_0

As presented in Table6.17, there was a significant statistical difference in the participant's mean SA between the baseline and evaluation scenario. The one-tailed test also revealed that the subject's mean Level 2 SA was lower in the baseline scenario than the evaluation scenario. That is to say, the participants experienced a higher level of situation comprehension when there were a greater level of autonomy transparency. This result confirmed the initial testing hypothesis.

The SA data collected for the participant's Level 2 SA could not be analysed using a parametric test due to the assumption of normality and homogeneity of variances could not be met. Hence, a variance-stabilising, square-root transform was performed on the data set, resulting the homogeneity of variance assumption to be met, and was able to perform a non-parametric hypothesis testing of the means on the transformed data.

Through the application of the the non-parametric Wilcoxon Signed-Rank test, the transformed Level 2 SA data revealed that at CL = 95%, there was a significant improvement in the participant's level of comprehension of the scenario where the UAVs' autonomy transparency was higher when compared with the scenario where the transparency was lower.

6.3.3 Combined SA

The hypothesis was that the overall SA increased along with an increase in the autonomy transparency. This was observed through the analysis of the combined SA of Level 1 and 2.

The raw data presented initially was not able to satisfy the assumptions of the parametric test, where normality in the distribution was not present with baseline scenario's result being slightly negatively skewed. The application of a square transform was used to rectify the skewness, which consequently addressed the distribution, on the other side, the similar variance assumption was sabotaged. Hence, a parametric and a non-parametric test was used to compare the sample means; the non-transformed data was used for the parametric test, while the square-transformed data was used for the non-parametric test.

The results of both tests revealed that the combined SA possessed by the participant during the scenario with a greater autonomy transparency was significantly higher than that of the lower autonomy. This section discussed the process and details of the analysis.

Given that only the homogeneity of variance assumption (presented in Appendix B) was not met, a parametric test was still carried out. Two pairs of comparisons of the means were conducted, and the summarised outcome is presented in Table 6.18 based on

the following two sets of hypotheses.

Table 6.18: Summary table of paired sample test comparing the combined SA of the baseline and the evaluation scenario

$(\alpha = 0.05, n = 42)$	**Baseline & Evaluation Scenario**
2-Tailed	$p < 0.001$, Reject H_0
1-Tailed	$\bar{x}_{baseline} < \bar{x}_{evaluation}$, Reject H_0

As presented in Table6.18, there was a significant difference in the participant's mean SA between the baseline and evaluation scenario. The one-tailed test also revealed that the subject's mean SA was lower in the baseline scenario than the evaluation scenario. That is to say, the participants experienced a higher level of SA when there were a greater level of autonomy transparency. This result confirmed the initial testing hypothesis. Furthermore, a non-parametric paired comparison test was also conducted.

The application of the non-parametric Wilcoxon Signed-Rank test assumed: The dependent variables of the data set must be continuous, The independent variables of the data set must be categorically collected, and The data set must have similar variances. The square root transformed data were able to satisfied all three assumptions. Hence, this test could be carried out reliably.

Two pairs of comparison of the means were conducted. The summarised outcome is presented in Table 6.19.

Table 6.19: Summary table of Wilcoxon Signed-Rank test comparing the square-root transformed combined SA of the baseline and the evaluation scenario

$(\alpha = 0.05, n = 42)$	**Baseline & Evaluation Scenario**
2-Tailed	$p < 0.001$, Reject H_0
1-Tailed	$\bar{x}_{baseline} < \bar{x}_{evaluation}$, Reject H_0

As presented in Table 6.19, there was a significant difference in the participant's mean SA between the baseline and evaluation scenario, and the one-tailed test also revealed that the subject's mean SA was lower in the baseline scenario than the evaluation scenario. That is to say, the participants experienced a higher level of SA when there were a greater level of autonomy transparency. This result, along with the parametric test result, both confirmed the initial testing hypothesis.

Two statistical tests were used to analyse the differences in the mean SA between the baseline and the evaluation experiment: a parametric, and a non-parametric. The parametric test was used after its assumptions were assessed and determined that only three of the four assumptions were met with the removal of an outlier. Although the reliability of the test results were reduced, it still gave an acceptable indication of the comparison. Furthermore, by the application of the square-root transformation on the data set, the assumptions of the non-parametric test were met, hence, the transformed mean SA data were also analysed with a greater reliability.

As presented in Tables 6.18 and 6.19, the significance values ($p - value$) of less than 0.001 revealed that the mean SA were lower in the baseline scenario when compared to the mean SA in the evaluation experiment. This meant that the participants experienced a higher level of SA in the scenario with a greater level of autonomy transparency. This also confirmed with the author's initial hypothesis for testing.

6.4 Discussion & Conclusion

The cognitive measures (CW and SA) were hypothesised to increase with an increase in autonomy transparency. This hypothesis was tested through two comparative scenarios: The baseline scenario established a cognitive performance benchmark where autonomy was opaque, and the evaluation scenario, where autonomy was transparent.

The functional subsystem that was used for testing was the UAVs' navigation capability, which was responsible for the reactive replanning of the flight path to mitigate threats detected by the UAV. Two forms of threats were scripted: A low UAV fuel level, and a hazardous cloud. The low fuel level warning alerts and informs the operator that the UAV was no longer capable of completing the current task with the current amount. Hence, refuelling was required. The hazardous cloud could cause the UAV to be destroyed and removed from the mission upon intersection.

During the scenario with the UAV's functional autonomy opaque to the participants, the UAVs' flight paths were not visually presented. During the scenario with the UAV's functional autonomy set to transparent, the flight path were presented as illustrated in Figure 4.2. The differences in cognitive performance of the pair of scenarios were analysed at CL = 95%. CW was analysed in six separate dimensions and one combined dimension. The six dimensions include: Mental demand, physical demand, temporal demand, performance, effort and frustration. The combined dimension included the results from the individual dimensions with each specific weight used to normalise the final score. SA was defined with three levels: Perception, comprehension, and projection. This experiment captured the perception and comprehension stages of an operator's SA and the analysis included a comparison of the participants' perceptional awareness, and their comprehensional awareness. A combined SA was also analysed where both levels of awareness were combined.

The results showed an improvement in the CW and SA scores as illustrated in Table 6.20.

Paired Sample T-Tests were conducted to compare the CW and SA of a human operator managing multiple UAVs in opaque navigation autonomy transparency and transparent navigation autonomy information conditions. There was a significant reduction in operator CW ($\bar{x} = 55.86\%$, $\sigma = 15$) and increased SA ($\bar{x} = 0.518$, $\sigma = 0.299$) in the transparent navigation autonomy information configurations and opaque transparency ($x_{\bar{CW}} = 59.95\%$, $\sigma_{CW} = 16, x_{\bar{SA}} = 0.171$, $\sigma_{SA} = 0.157$) conditions: $N = 42$, $p_{CW} = 0.033, p_{SA} < 0.001$.

Table 6.20: Summary of the cognitive performances

		$p - value$	**Verdict**
CW	Mental Demand	$p = 0.263$	No Different
	Physical Demand	$p = 0.931$	No Different
	Temporal Demand	$p = 0.171$	No Different
	Performace	$p = 0.276$	No Different
	Effort	$p = 0.032$	Improved
	Frustration	$p = 0.172$	No Different
	Combined	$p = 0.033$	Improved
SA	Level 1	$p = 0.086$	No Different
	Level 2	$p < 0.001$	Improved
	Combined	$p < 0.001$	Improved

The results and analysis presented in this chapter suggest that by visually representing the machine agent's objectives in confrontational circumstances; which implicitly suggests its functional autonomy capability, the human agent teammate was able to experience a reduction in CW and an increase in SA. Hence, Contribution 2 and Research Question 2 is addressed. Chapter 7 further explores the effect of increasing autonomy transparency in all functional subsystem's autonomy capabilities of the machine agents - applying the basic principle of natural language dialogue, and representing autonomy status through a text-based dialogue exchange system.

CHAPTER 7

Experiment 3: Complete Autonomy Transparency

The aim of this experiment is to verify the effect of communicating the UAV agents' functional and autonomy capabilities to the human operators, in order to increase autonomy transparency. Details of this experiment overview, design, and methodology is presented in Chapter 4. The rest of this chapter presents the specific testing scenario specifications and a detailed analysis of the collected Cognitive Workload (CW), Situation Awareness (SA), Automation Trust (Trust), and the three objective performances measures - Initial Reaction Time (IRT), the Event Reaction Time (ERT) and the number of Items Of Interest (IOIs) found.

7.1 Scenario Description

The scenario describes that a number of distress signals were received from the region and a team of four rotory-winged UAVs were required to deploy to that area and conduct identification searches of the expected targets. Each was differentiated by a separate colour: (B)lue, (G)reen, (R)ed, and (Y)ellow. Research had shown that the use of colour can emphasize the logical information organisation [80], and is very useful to differentiate components in an information-rich interface [87].

There were four discrete locations during each scenario, each UAV was assigned to search one location.

There was a baseline scenario which established the human operator's performance with the absence of autonomy information transparency. An evaluation scenario to assess the operator's performance with autonomy transparency in a form of text-based mixed-initiative dialogue followed.

During the scenarios, each UAV performed at different LOA for each mode-of - operation. They were deployed in transit mode initially, and depending on the scenario's specific events, it could enter different modes-of-operation with a shift in LOA. Furthermore, the different number of IOIs and decoys between each scenario is intentional in order to promote mission realism and unpredictability.

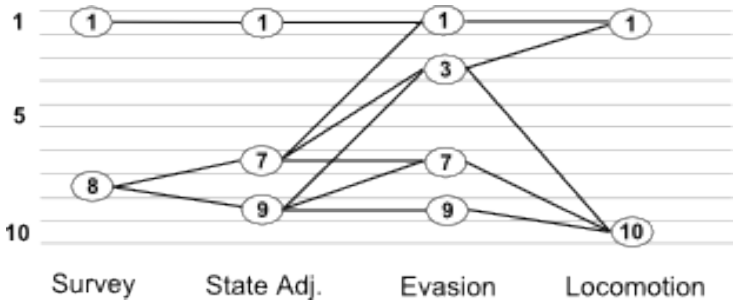


Figure 7.1: Transit Mode Autonomy Spectrum

7.1.1 Tasks, Objectives and Modes-Of-Operations

The primary objectives for the participants were to supervise the UAVs to their designated search location and to perform the search activity to "rescue" as many IOIs as possible. To achieve this, three mode-of-operations were defined: Mode 1 Transit, Mode 2 Hazard Avoidance, and Mode 3 Search.

Mode 1: Transit

The *Transit Mode* is the most commonly used throughout the experiment. The UAVs were required to move from one location to the next. This mode included several scenarios that the UAVs may experience and are required to address.

- **Survey:** The agent (human or UAV) constantly observed for any events that may interrupt the transiting safety of the UAVs. This included potential collisions with other UAVs or terrains, or other expected state profiles (altitude and speed)
- **State-Adjustment:** Determined the nature of the hazard and its effects
- **Evade:** The type of evasive action that was to be used in this situation was decided and prepared for execution
- **Locomotion:** The evasive action was carried out by an agent in a form of locomotion, where the UAV should be able to be separated from the hazard

Several autonomy levels were assigned to each of the steps as illustrated in Figure 7.1.

Mode 2: Hazard Avoidance

The hazard avoidance mode would trigger when the UAV encountered some perturbation events such as hazardous clouds which could have dire consequences. The steps below describe the progression of this mode.

- **Identify:** Identifies the location and its relation to the UAV
- **Consequence:** Determines the consequences of the UAV as it confronts the hazard

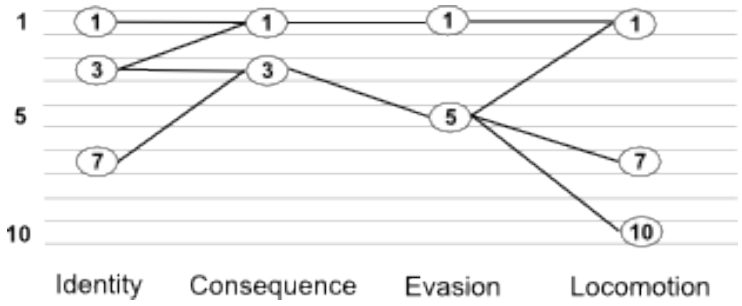


Figure 7.2: Complex Hazard Autonomy Spectrum

Figure 7.3: Item of Interest Searching Mode Autonomy Spectrum

- **Evasion:** Decides on the evasive measure for the hazard, i.e. how to best overcome the problem
- **Locomotion:** Commanded the UAV to execute the decided evasive action for the particular hazard.

Figure 7.2 illustrates the autonomy spectrum layout for the complex hazard mode

Mode 3: Search

The search mode took place when the UAV entered its specific search area. When the area was reached, each UAV entered this mode with the following steps.

- **Scan:** When specified search area was entered, the agent scans for items of interest (IOI) to pick up on items from the background
- **Identify:** Mentally identifies and distinguishes IOIs from decoy objects, as IOI were items that need to be selected and confirmed
- **Selection:** The agent initiates a selection of the IOI identified in the previous step
- **Confirmation:** Checks/confirms that the selection was indeed correctly initiated

Figure 7.3 illustrates the autonomy spectrum layout for the search mode.

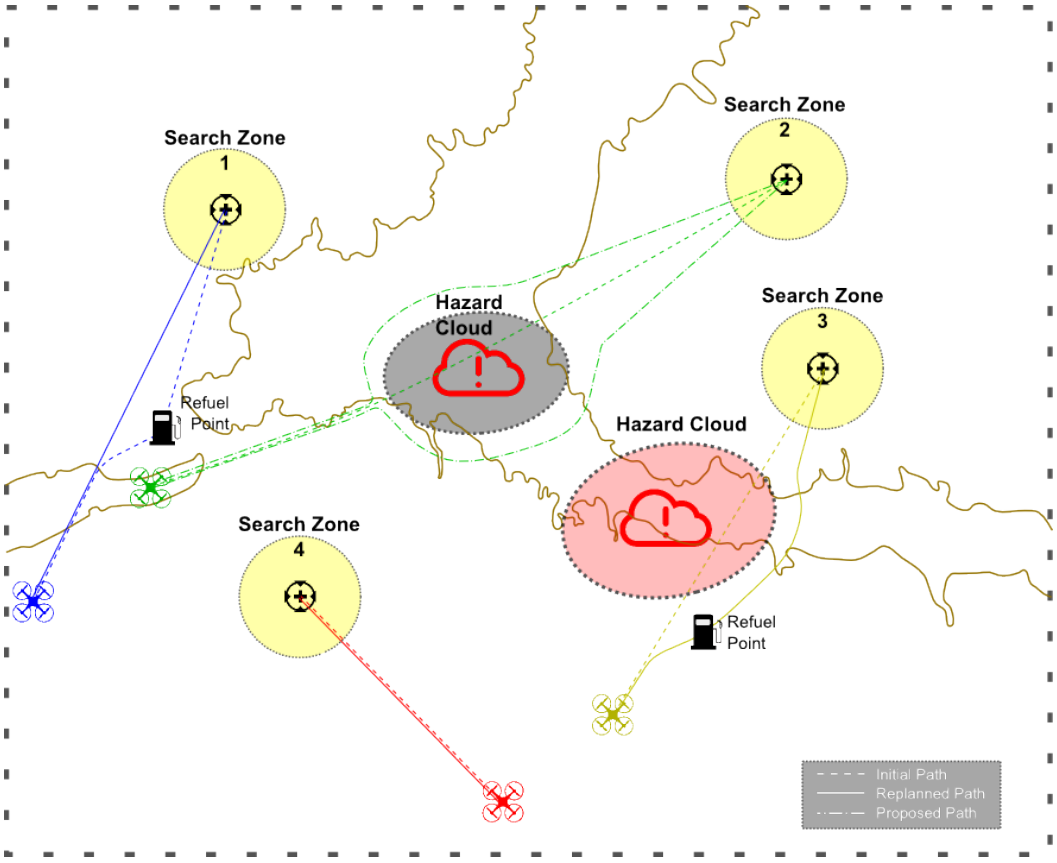


Figure 7.4: Graphical illustration of the baseline scenario layout (planned and produced through Google Earth)

7.1.2 Baseline Scenario (Opaque Autonomy Capability)

The purpose of the baseline scenario was to establish a benchmark result of a participant's performance with no autonomy transparency. This meant that regardless of the configurations of the autonomy capabilities in each mode-of-operation, the participant received no visual or textual cue about the UAV's capabilities.

The approximate experiment layout was illustrated in Figure 7.4. This figure illustrates four search zones, denoted by groups of yellow concentric circles. Each coloured aircraft icon denoted its corresponding UAVs. Two solid regions in the centre denote hazardous zones.

UAV Autonomy Spectrum

The UAVs' autonomy spectrum and associated events are illustrated in Table 7.1. The autonomy spectrum for each UAV was designed according to the combination of possible configurations for each mode as illustrated in Section 7.1.1.

Table 7.1: Autonomy spectrum configuration and associated perturbation events for the baseline scenario

Modes of Operation	UAV_{Blue}	UAV_{Green}	UAV_{Red}	UAV_{Yellow}
Transit Mode	❽-❼-❸-❿	❽-❾-❾-❿	❽-❼-❶-❶	❽-❼-❸-❿
Hazard Avoidance	❼-❸-❺-❶	❸-❸-❺-❼	❼-❸-❺-❿	❸-❶-❶-❶
Search Mode	❿-❶-❶-❶	❶-❶-❶-❶	❿-❸-❻-❶	❿-❸-❶-❶
Pert. Events	Low Fuel	Hazard Cloud	-	Hazard Cloud

Table 7.2: Experiment 3 Baseline scenario LOA and LOD (at the commencement of the trial) configuration table for each UAV to achieve each UAV's equivalent mode-of-operation

	LOA				**LOD**			
Subsystem	Health	Nav	States	Payload	Health	Nav	States	Payload
UAV_{Blue}	HA	PA	PA	LA	HL	LL	ML	LL
UAV_{Green}	HA	HA	PA	LA	LL	LL	ML	ML
UAV_{Red}	HA	LA	HA	PA	HL	ML	HL	ML
UAV_{Yellow}	HA	LA	LA	LA	ML	ML	ML	LL

UAV Functional Capability Specification

The LOA and LOD configuration to achieve the desired modes of operation for each UAV are illustrated in Table 7.2. The functional subsystems of the UAVs include (from left to right of the table) health monitoring, navigation capability, autopilot/state autonomy, and payload operations. Three levels of LOA and LOD were applied: High LOA/LOD (HA/HL), Medium LOA/LOD (MA/ML), and Low LOA/LOD (LA/LL). The designated LOAs were not available for change by the participant, while the LOD could be changed as desired by the participant.

Designated Search Area Specification

Each UAV received a designated search area, and each search area contained a different number of decoy items and IOIs to be found. Table 7.3 presents the central coordinates of these search areas, the number of decoys deployed, the amount of IOIs to be found, and the required speed of the respective UAVs upon arrival to the search areas.

Table 7.3: Experiment 3 Baseline Scenario Search Area Specification

	Coordinates (lat, lon)	**Decoys**	**IOIs**	**Speed**
UAV_{Blue}	Z1 (44.487313, 6.273086)	4	8	300
UAV_{Green}	Z2 (44.489409, 6.334134)	3	8	290
UAV_{Red}	Z4 (44.474535, 6.337805)	3	10	340
UAV_{Yellow}	Z3 (44.457197, 6.282938)	6	10	340

Table 7.4: Experiment 3 Baseline Scenario Perturbation Event Specification

	Type	Coordinate (Lat, Lon)	Severity	Rotated Angle
UAV_{Blue}	Fuel	Fuel Zone (44.469491, 6.266316) Fuel Zone (44.453889, 6.324341)	100% Loss on fuel starvation	N/A
UAV_{Green}	Cloud	Cloud (44.474799, 6.300654)	50% Chance of Loss on Contact	50° Clockwise
UAV_{Red}	None	N/A	N/A	N/A
UAV_{Yellow}	Cloud	Cloud (44.463347, 6.321352)	100% Chance of Loss on Contact	100° Clockwise

Perturbation Event and Environmental Specification

A summary of the perturbation events seen in the baseline scenario are presented in Table 7.4. In this scenario, UAV_{Red} did not experience any perturbation events, while UAV_{Blue} encountered a low fuel status and UAV_{Green} and UAV_{Yellow} confronted a hazardous cloud, which had a potential asset loss percentage of 50% and 100% respectively.

There were three perturbation events outlined in Table 7.4 which applied to UAV_{Blue}, UAV_{Green} and UAV_{Yellow}.

UAV_{Blue} had a low fuel status early in the session and a moderately high autonomy in its transit and hazard avoidance modes-of-operation. This was reflected through an autonomous flight path regeneration which operated under the *Management By Consent* (MBC) paradigm (flight path options were available for the human agent's manual selection), as opposed to the *Management By Exception* paradigm which required the human agent to interject the machine agent's decision and execution if necessary, usually seen in systems with higher LOAs.

UAV_{Green} experienced a hazardous cloud which had a 50% possibility to cause UAV failure if it attempted to go through the cloud. UAV_{Green} also had a moderate LOA for its hazardous avoidance mode-of-operation. This also operated under the MBC paradigm. The UAV was able to propose new flight paths nearing encounter with the hazardous cloud which required the participant to notice the proposals and make a suitable selection.

UAV_{Yellow} also experienced a hazardous cloud. However, this cloud had a 100% UAV failure rate. This meant that the UAV would be rendered inoperable upon encounter with the cloud. UAV_{Yellow} had a low autonomy in its hazard avoidance mode of operation, which meant that a high level of manual intervention by the human agent was required. In the context of the situation, the participant was required to manually *steer* the UAV away from the hazard and proceeded en-course manually.

In this (baseline) scenario, a lack of any form of communication was available to notify or alert the participant of the UAVs' autonomy or capability, all participant interactions and management were initiated based on the participant's observation of the system's

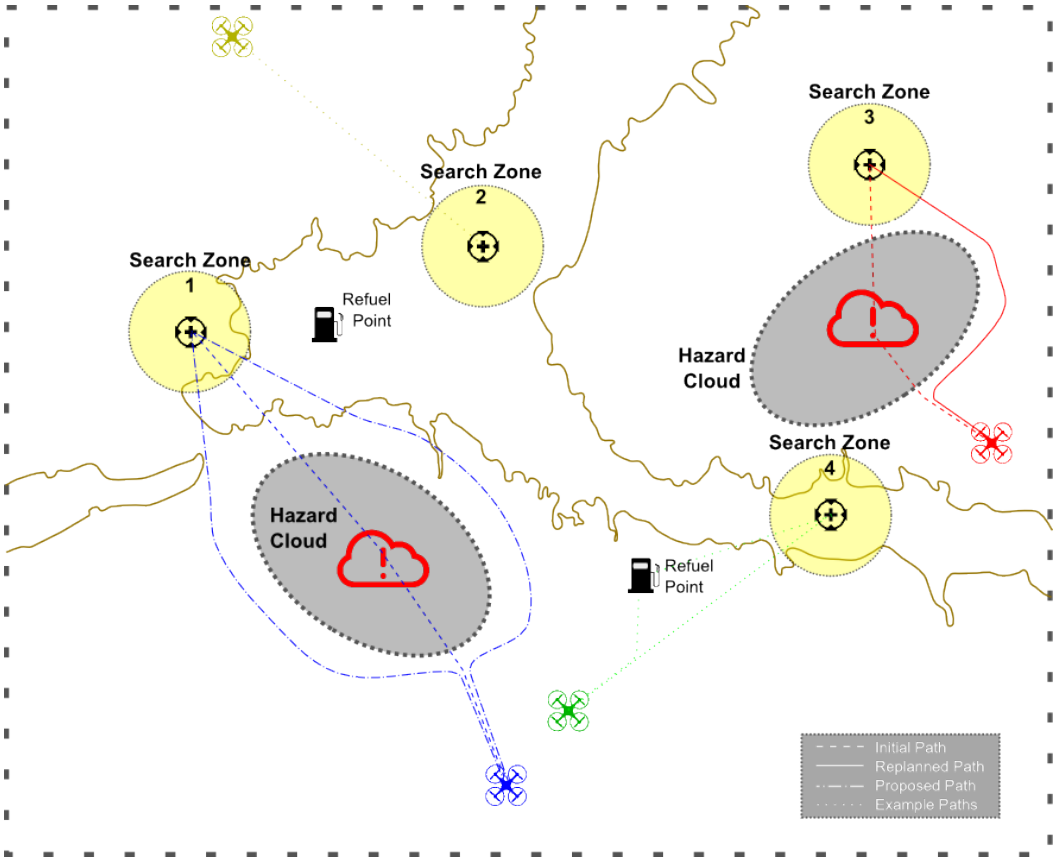


Figure 7.5: Graphical illustration of the evaluation scenario layout (planned and produced through Google Earth)

behaviour.

7.1.3 Evaluation Scenario (Transparent Autonomy Capability)

The purpose of the evaluation trial was to assess the mission operator's performance in comparison with the baseline trial results. The varying factor was the automation transparency feedback.

The increase of the system and UAVs' autonomy transparency was implemented through the communication of the system's autonomy capability and situational status to the participants through a simple textual form of mixed-initiative dialogue.

The approximate experiment layout was illustrated in Figure 7.5. The type of elements inside this configuration were the same as the baseline scenario; where the differences lay in the the positions of these elements. The coordinates of the search zones (Z1 to Z2), as well as the detail specifications of the perturbation events are described in this section.

UAV Autonomy Spectrum

Table 7.5 presents the autonomy spectrum configuration of the evaluation scenario. The autonomy spectrum for each UAV was designed according to the combination of possible

Table 7.5: Autonomy spectrum configuration and associated perturbation events for the evaluation scenario

Modes of Operation	UAV_{Blue}	UAV_{Green}	UAV_{Red}	UAV_{Yellow}
Transit Mode	❽-❾-❸-❶	❽-❼-❸-❿	❽-❾-❼-❿	❽-❼-❸-❶
Hazard Avoidance	❼-❸-❺-❼	❼-❸-❺-❿	❸-❸-❺-❿	❸-❸-❺-❼
Search Mode	❿-❸-❻-❶	❿-❸-❶-❶	❿-❸-❻-❶	❿-❸-❻-❿
Pert. Events	(S)(C)	-	(C)	(S)(F)

Table 7.6: Experiment 3 Evaluation scenario LOA and LOD (at the commencement of the trial) configuration table for each UAV to achieve each UAV's equivalent mode-of-operation

	LOA				**LOD**			
Subsystem	Health	Nav	States	Payload	Health	Nav	States	Payload
UAV_{Blue}	HA	PA	HA	LA	HL	LL	ML	LL
UAV_{Green}	LA	HA	HA	LA	LL	LL	ML	ML
UAV_{Red}	HA	PA	PA	PA	HL	LL	ML	ML
UAV_{Yellow}	HA	PA	LA	HA	ML	HL	ML	LL

configurations for each mode as illustrated in Section 7.1.1.

UAV Functional Capability Specification

The LOA and LOD configuration to achieve the desired modes of operation for each UAV are illustrated in Table 7.6. The functional subsystems of the UAVs included (from left to right of the table) health monitoring, navigation capability, autopilot/state autonomy, and payload operations. Three levels of LOA and LOD were applied: High LOA/LOD (HA/HL), Medium LOA/LOD (MA/ML), and Low LOA/LOD (LA/LL).

The designated LOAs were not available for change by the participant, while the LOD could be changed as desired by the participant.

Designated Search Area Specification

Each UAV received a designated search area, and each search area contained a different number of decoy items and IOIs to be found. Table 7.7 presents the central coordinates of these search areas, the amount of decoys deployed, the amount of IOIs to be found, and the required speed of the respective UAVs upon arrival to the search areas.

Table 7.7: Experiment 3 Evaluation Scenario Search Area Specification

	Coordinates (lat, lon)	**Decoys**	**IOIs**	**Speed**
UAV_{Blue}	Z1 (44.477578, 6.270476)	5	7	500
UAV_{Green}	Z4 (44.484674, 6.301303)	2	7	450
UAV_{Red}	Z3 (44.489693, 6.342926)	3	9	420
UAV_{Yellow}	Z2 (44.463575, 6.338638)	3	8	480

Table 7.8: Experiment 3 Baseline Scenario Perturbation Event Specification

	Type	Coordinate (Lat, Lon)	Severity	Rotated Angle
UAV_{Blue}	Cloud	Cloud (44.477666, 6.342978)	100% of Loss on Contact	45° Clockwise
UAV_{Green}	N/A	N/A	N/A	N/A
UAV_{Red}	Cloud	Cloud (44.461244, 6.291291)	100% of Loss on Contact	60° Clockwise
UAV_{Yellow}	Fuel	Refuel Zone (44.457524, 6.337058)	100% Loss on fuel starvation	N/A

Perturbation Event and Environmental Specification

A summary of the perturbation events seen in the evaluation scenario are presented in Table 7.8. In this scenario, UAV_{Green} did not experience any perturbation events, while UAV_{Yellow} encountered a low fuel status and UAV_{Blue} and UAV_{Red} confronted a hazardous cloud, which had a potential asset loss percentage of 100%, that is, UAV failure was imminent if the UAV had flown into the cloud.

There are three perturbation events outlined in Table 7.8 which applied to UAV_{Blue}, UAV_{Red} and UAV_{Yellow}.

UAV_{Yellow} experienced a low fuel event in this scenario. This UAV had a moderately high autonomy in its transit and hazard avoidance modes-of-operation. In the experiment, this was reflected through an autonomous flight path regeneration which operated under the MBC paradigm.

UAV_{Blue} experienced a hazardous cloud which had a 100% UAV failure rate if it attempted to go through the cloud. This UAV also had a moderate to high LOA for its hazard avoidance mode of operation, which meant that it operated under the MBC paradigm. The UAV was able to propose new flight paths nearing the encounter with the hazardous cloud which required the participant to notice the proposals and make a suitable selection.

UAV_{Red} had also experienced a hazardous cloud with 100% UAV failure rate. This UAV had a moderate decision-making autonomy and a high post-action capability as illustrated in Table 7.5. This represented an MBC management paradigm during the stage where a decision was required to be made on the evasive action to the hazard (decision selection was expected to be initiated by the human agent/participant), and a high autonomy to carry out the action.

The interface of this evaluation scenario introduced the text-based communication dialogue box (as illustrated in Figure 4.10), which represented each UAV's autonomous status model to the participants managing the mission. This representation provided an increased transparency in the system's autonomy capability, enabling the participants to be able to perform the experiment without extensive needs to make assumptions about the system's behaviour.

7.2 Result & Analysis: Cognitive Workload

The analysis of the CW for this experiment was similar to the previous experiments where the magnitude of load of the six dimensions of CW were analysed separately. A combined CW drew a synthesized outcome of the combination of the source of load (the weighing of each attribute) and the magnitude of load (the rating of each attribute).

The analysis was done using the software package SPSS. All conclusions were drawn based on a Confidence Level (CL) = 95% ($\alpha = 0.05$). The aim of the comparison was to determine the differences between the scenarios. Hence, statistical hypothesis tests were used. The Paired Sample T-Test was used for the parametric testing while the Wilcoxon Signed-Rank test was used as its non-parametric counterpart. The hypothese are defined in Section 6.2, Chapter 6.

7.2.1 Mental Demand

An analysis of the raw *Mental Demand (MD)* dataset from both scenarios revealed a non-normal distribution (Table C.1 Appendix C) in the baseline data and a normal distribution in the evaluation data. The baseline data could not satisfy the assumptions for the use of a parametric comparison of the means test, while the evaluation data could. Furthermore, there were no appropriate transformations that could be simultaneously applied to both data sets while ensuring that they both were able to still meet the assumptions, hence, a non-parametric test was used.

The outcome from the test revealed that the participants felt significantly greater MD in the baseline scenario, where there were limited autonomy transparencies about the UAVs' modes-of-operation, as compared to the evaluation scenario where the autonomy were communicated to the participants in real time. From the assumptions testing presented in Appendix C, a paired comparison of the means was conducted on the original data using the non-parametric test.

Table 7.9: Summary table of the Wilcoxon Signed-Rank test comparing the magnitude of mental demand/load felt by the participants during the experiment

($\alpha = 0.05$, $n = 43$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.043$, Reject H_0
1-Tailed	$\bar{x_{Baseline}} > \bar{x}_{Evaluation}$, Reject H_0

Table 7.9 presents the test summary of the comparison. From the table, there was significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference between the mean MD of the baseline and evaluation scenario. On further evaluation using the one-tailed test, H_0 was again rejected as the participants exhibited greater reported MD in the baseline scenario where minimal UAVs' autonomy was communicated to the human agent.

The descriptive statistics of the magnitude of MD of the baseline and the evaluation scenario presented in Table C.1 Appendix C reveal that the baseline MD data was not normally distributed while the evaluation MD data was, therefore, a transformation could not be carried out while ensuring that both sets of data still presented a normal distribution, hence, the parametric test assumptions could not all be satisfied. While the normality assumption could not be satisfied, the remaining assumptions could. Therefore, a non-parametric test was used to analyse the differences in the means.

Applying the non-parametric test revealed that the magnitude of the MD results of the baseline scenario was significantly higher than the evaluation scenario.

7.2.2 Physical Demand

An analysis of the raw *Physical Demand (PD)* data set from both scenarios revealed a positive skew, which would normally be corrected using a logarithmic or root transformation. However, due to a number of sample points being zero, no transformation could be applied. Since the normality assumption was not met, the non-parametric test was used.

The outcome of the non-parametric test revealed that there were insignificant statistical differences between the PD experienced by the participants when comparing the evaluation scenario to the baseline scenario.

Given the assumptions testing presented in Appendix C, a paired comparison of the means was conducted on the PD data using the non-parametric test.

Table 7.10: Summary table of the Wilcoxon Signed-Rank test comparing the magnitude of physical load felt by the participants during the experiment

($\alpha = 0.05$, $n = 36$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.208$, Accept H_0

Table 7.10 presented the test summary of the comparison. From the table, there were no significant statistical evidence to reject the null hypothesis (H_0), meaning that there was no significant difference between the mean PD experienced by the participants in the baseline and evaluation scenarios.

The descriptive statistics of the magnitude of PD of the baseline and the evaluation scenario presented in Table C.2 Appendix C reveal that the PD data of both scenarios were was not normally distributed with a positive skew, and that a logrithmic/root transformation would be necessary to adjust the data. This was not possible because as Figures C.2 and C.3 (Appendix C) illustrate, there were a number of sample points that had the value of zero (0), hence, correct calculation could not be carried out on those data. Since the normality assumption could not be satisfied but the remaining assumptions could, a non-parametric test was used to analyse the differences in the means.

Applying the non-parametric test revealed that the magnitude of the PD results between the scenarios did not differ. This suggested that the participants did not ex-

perience significant increased or decreased PD with an increase transparency of the UAVs' autonomy.

7.2.3 Temporal Demand

An initial set of the raw *Temporal Demand (TD)* data revealed to be not normal with a negative skew, to which a square transform was applied. After analysis, the distribution of the transformed data still was not normally distributed. Hence, a non-parametric test was used to evaluate the dataset. The source of load data was intentionally left out as it did not provide much relevance when analysing temporal demand in isolation.

From the assumptions testing presented in Appendix C, a hypothesis testing of the means was used and test was conducted on the transformed PD data using the parametric test.

Table 7.11: Summary table of the paired-sample test comparing the magnitude of temporal load (square-transformed) felt by the participants during the experiment

($\alpha = 0.05$, $n = 36$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.518$, Accept H_0

Table 7.11 presented the test summary of the comparison. From the table, there was no significant statistical evidence to reject the null hypothesis (H_0), meaning that there was no significant difference between the mean TD experienced by the participants in the baseline and evaluation scenario.

Table C.3 presented the descriptive statistics of the magnitude of TD of the baseline and the evaluation scenario. The statistics revealed that the PD data of both scenarios were was not normally distributed with a negative skew. Hence, a squared transform was applied.

The transformed datasets were able to satisfy the assumptions of the parametric test as presented in Table C.4. Hence, the Paired Sample T-Test was used.

The T-Test revealed that the magnitude of the TD between the scenarios did not differ significantly. This suggested that the participants did not experience significant increased or decreased TD with an increase transparency of the UAVs' autonomy.

7.2.4 Performance

The *performance* (or perceived performance) data was tested to be normally distributed, and along with the positive outcomes of the other assumptions for a parametric, Paired Sample T-Test was used to analyse the means. The source of load data was intentionally left out as it did not provide much relevance when analysing temporal demand in isolation.

From the assumptions testing presented in Appendix C, a paired-sample parametric test was conducted on the operators' perceived performance data.

Table 7.12: Summary table of the paired-sample test comparing the participants' perceived performance results

($\alpha = 0.05$, $n = 36$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.06$, Accept H_0

Table 7.12 presents the test summary of the comparison. From the table, there was no significant statistical evidence to reject the null hypothesis (H_0). Hence, there was no significant difference between the mean performance perceived by the participants during the baseline and evaluation scenario and hence no performance degradation could be determined.

The descriptive statistics of the participant's perceived performance during the baseline and the evaluation scenario presented in Table C.5 Appendix C illustrate that the raw data of both scenarios were sterile. Therefore, no manipulation was required to meet the assumptions for the parametric test. For this reason, a Paired Sampel T-Test was used for analysis. Applying the test revealed that there were no significant differences in the participants' perceived performances between the opaque and the transparency autonomy configurations.

7.2.5 Effort

The outcome of the test revealed that there was a significant difference between the *effort* required to perform the tasks in the baseline scenario than that of the evaluation experiment. From the assumptions testing presented in Appendix C, a parametric T-Test was conducted on the original data.

Table 7.13: Summary table of the T-Test comparing the magnitude of effort felt by the participants during the experiment

($\alpha = 0.05$, $n = 35$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.046$, Reject H_0
1-Tailed	$\bar{x_{Baseline}} > \bar{x}_{Evaluation}$, Reject H_0

Table 7.13 presents the test summary of the comparison. From the table, there was a significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference between the participant's perceived effort required to perform the mission described by the baseline and evaluation scenario. On further evaluation using the one-tailed test, $H_0$was again rejected as the participants exhibited greater reported effort to perform the test was required in the baseline scenario where minimal UAVs' autonomy was communicated to the human operator.

7.2.6 Frustration

An initial set of the raw *frustration* level data revealed to be not normal with an acceptable skewness. Due to the wide distribution and randomness of the data, no transforms could

be applied to sanitise the data set such that all the assumptions for conducting a parametric test could be met. However, the assumptions of the non-parametric test of the means were met, hence, the non-parametric test was used to evaluate the data set.

The source of load data was intentionally left out as it did not provide much relevance when analysing temporal demand in isolation.

From the assumptions testing presented in Appendix C, a non-parametric test of the means was conducted on the original data.

Table 7.14: Summary table of the Wilcoxon Signed-Rank test comparing the magnitude of frustration felt by the participants during the experiment

($\alpha = 0.05$, $n = 36$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.704$, Accept H_0

Table 7.14 presented the test summary of the comparison. From the table, there was no significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was not much statistical difference between the mean participant frustration level from a test configuration where there were limited autonomy transparency to one there was an increased transparency.

7.2.7 Combined Cognitive Workload

The combined CW defined the overall CW score by applying formula 5.1. The aim was to compare the means of the combined CW score between the baseline and the evaluation scenario.

An analysis of the raw combined CW data from both scenarios revealed negative skews (Table C.9 Appendix C) and three significant outliers. The outliers were removed for the further analysis. The set of transformed data was analysed which revealed a set of outlier data points (Table C.10 Appendix C), which was removed for further analysis.

From the assumptions testing presented in Appendix C, a paired comparison of the means was conducted on the datasets where the three outliers were removed.

Table 7.15: Summary table of paired sample test comparing the combined CW felt by the participants during the experiment

($\alpha = 0.05$, $n = 42$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.022$, Reject H_0
1-Tailed	$\bar{x_{Baseline}} > \bar{x}_{Evaluation}$, Reject H_0

Table 7.15 presented the test summary of the parametric T-Test. From the table, there was significant statistical evidence to reject the null hypothesis (H_0). This means that there was a significant difference between the mean combined CW rating of the baseline and evaluation scenarios.

The further investigation using a one-tailed test revealed that at CI = 95%, there was significant statistical evidence to reject H_0. That is to say, the combined CW of the baseline scenario was greater than that of the evaluation scenario.

This suggested that the combined CW was significantly higher in a system there the UAVs' autonomy capability were opaque from the participants, compared to a system where the autonomy capabilities were communicated to the participants. Hence, being transparent.

7.2.8 Analysis Summary

This section presented an in depth analysis of the six attributes that form CW as defined by Hart *et al.*. The six attributes were analysed individually as well as together to illustrate the differences in the participant's CW when the interface configuration included the UAV's autonomy status versus the configuration where they were absent. Table 7.16 presents a summarised table of all the statistics along with specific details and outcomes, where *desirable* denoted that the magnitude of load was greater during the baseline experiment.

Table 7.16: A summary of the results obtained from the testing of the mean CW ($\sigma = 0.365$) between the opaque autonomy and the transparent autonomy interface configurations.

CL=95%,$\alpha = 0.05$	**Outcome**	$\bar{x}_{baseline}$	$\bar{x}_{evaluation}$
Mental Demand	*Desirable* ($p = 0.043$)	16.65 ($\sigma = 9.0$)	13.64 ($\sigma = 7.8$)
Physical Demand	No Different ($p = 0.208$)	5.361 ($\sigma = 5.2$)	6.222 ($\sigma = 5.2$)
Temporal Demand	No Different ($p = 0.518$)	13.17 ($\sigma = 4.2$)	12.72 ($\sigma = 4.4$)
Performance	No Different ($p = 0.06$)	10.97 ($\sigma = 3.8$)	9.89 ($\sigma = 4.0$)
Effort	*Desirable* ($p = 0.046$)	13.17 ($\sigma = 4.0$)	12.19 ($\sigma = 4.4$)
Frustration	No Different ($p = 0.591$)	10.53 ($\sigma = 5.5$)	9.86 ($\sigma = 6.2$)
Combined CW	**Desirable** ($p = 0.022$)	67.06 ($\sigma = 16.5$)	61.94 ($\sigma = 16.1$)

Evidently from Table 7.16, the feature of autonomy transparency, presented in a style of a natural language dialogue message-box, significantly reduced the operator's mental demand levels, effort required, and the overall CW levels. On close inspection of the results, although the remaining attributes were not tested to have significant statistical evidence, the direction of the mean magnitude of loads indicated that the loads were generally higher in interface configuration where the UAVs' autonomy capabilities were not communicated to the participants, while the loads were lower in the configuration where they were communicated.

7.3 Result & Analysis: Situation Awareness

SA was captured using an online questionnaire which captured the participants' Level 1 and Level 2 SA. The results were analysed at CL = 95% ($\alpha = 0.05$) and from three perspectives; Level 1, Level 2 and a combined SA.

Similar to the CW result analysis, a parametric Paired Sample T-Test was preferred given that all the assumptions were satisfied. However, if this was not possible even after attempts to reorganise the results, a non-parametric Wilcoxon Signed Rank test was used. The hypothesis for these tests remained the same as previously stated in Section 6.3, Chapter 6.

7.3.1 Level 1 SA

Level 1 SA represents the subject's perception of the situation. In the context of this experiment, the perception was the participant's awareness of the mission situation and the UAVs' autonomy capabilities.

Analysis focused on the responses of the questions that were designed to be assessing the participant's perception of the situation in percentage. The detailed assumptions testing (presented in Appendix C) suggested that the parametric Paired Sample T-Test was to be used to perform the hypothesis testing of the means. Even though the actual mean values have been changed due to the transformation, its significance and relational relevance to each other did not change.

Table 7.17: Summary table of the paired samples comparing the level 1 SA of the baseline and the evaluation scenario using the paired-sample test

$(\alpha = 0.05, n = 33)$	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.002$, Reject H_0
1-Tailed	$\bar{x}_{Baseline} < \bar{x}_{Evaluation}$, Reject H_0

As presented in Table7.17, with CL = 95%, there was a significant difference in the participant's mean Level 1 SA between the baseline and evaluation scenario. The participants possessed high levels of SA during the evaluation scenario. This suggests that the participants experienced statistically significantly higher levels of situational perception when there was a greater level of autonomy transparency compared to lower levels.

7.3.2 Level 2 SA

Level 2 SA represents the subject's understanding of the situation after they had perceived the event. In the context of this experiment, the comprehension was the participant's understanding of the mission situation and the asset's autonomous capabilities.

Analysis focused on the responses of the questions that were designed to be assessing the participant's comprehension of the situation in percentage. The outcomes from the assumptions testing presented in detail in Appendix C yielded that the non-parametric Wilcoxon Signed-Rank test was to be used to perform the hypothesis testing of the means based on the data set where the outlier sample was removed from both scenario data sets. Subsequently, Table 7.18 presents the test outcomes of the mean comparisons.

Table 7.18: Summary table of the non-parametric test comparing the Level 2 SA of the baseline and the evaluation scenario

$(\alpha = 0.05, n = 33)$	**Baseline & Evaluation Scenario**
2-Tailed	Accept H_0, $p = 0.308$

As presented in Table 7.18, there was no significant difference in the participant's mean Level 2 SA between the baseline and evaluation scenario. The participants did not exhibit a greater level of comprehensible SA in the interface configuration where autonomy transparency was greater. This suggests that the participants did not experience significantly different levels of situational comprehension when there was a greater level of autonomy transparency compared to lower levels.

7.3.3 Combined SA

The hypothesis was that the overall combined SA increased along with an increase in the autonomy transparency in the evaluation scenario.

The raw combined SA data initially did not satisfy the assumptions for a parametric test, while an outlier in the baseline data was detected, causing a negative skew to the data. As a result, the outlier sample was removed from both sets of scenario data for fairness, and the data sets were analysed again to verify the assumptions.

The dataset with the outlier removed enabled the parametric test assumptions to be met. The application of the parametric test revealed that the combined SA of the baseline scenario configuration was significantly less than that of the evaluation scenario.

The outcomes from the assumptions testing presented in detail in Appendix C yielded that a paired sample comparison of the means using a parametric T-Test was conducted on the data sets where the three outliers were removed.

Table 7.19: Summary table of paired sample test comparing the combined SA felt by the participants during the experiment

$(\alpha = 0.05, n = 33)$	**Baseline & Evaluation Scenario**
2-Tailed	$p < 0.001$, Reject H_0
1-Tailed	$\bar{x}_{Baseline} < \bar{x}_{Evaluation}$, Reject H_0

Table 7.19 presented the test summary of the parametric T-Test. From the table, there was significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference between the mean combined SA score rating of the baseline and evaluation scenario.

The further investigation using a one-tailed test revealed that at CI = 95%, there was significant statistical evidence to reject H_0. This suggests that the combined SA of the baseline scenario was lower than that of the evaluation scenario. This suggested that the combined SA was significantly higher in a system where the UAVs' autonomy capability were well communicated to the participants, hence, being transparent, compared to the

configuration where the autonomy was opaque.

7.4 Result & Analysis: Trust in Automation

Five subjective measures were recorded to gauge the operators' trust in automation; competence, predictability, reliability, faith and overall trust. These measures were rated from 1 to 7 on a paper-based rating form. The *two-tailed* test hypothesis states:

The analysis of these parameters were conducted with hypothesis testing with the hypothesis below. The two-tailed hypothesis stated:

- H_0: There were significant differences in the reported score of the respective attribute between the opaque autonomy scenario configuration and the transparent autonomy scenario configuration
- H_a: There were insignificant differences in the reported score of the respective attribute between the opaque autonomy scenario configuration and the transparent autonomy scenario configuration

The hypothesis for the *one-tailed* test comparing the attributes stated:

- H_0: The reported score of the respective attribute from the opaque autonomy scenario configuration was lower than the transparent autonomy scenario configuration
- H_a: The reported score of the respective attribute from the opaque autonomy scenario configuration was higher than the transparent autonomy scenario configuration

Two forms of tests were primarily used for the analysis: the parametric Paired Sample T-Test, and the non-parametric Wilcoxon Signed-Rank test. In order to determine the appropriate statistic test, certain assumptions, which were introduced and described in Section C.1 must be met. All conclusions were drawn with CL = 95% ($\alpha = 0.05$).

7.4.1 Competence

Competence denoted the extent to which the system perform a given task effectively. In the scope of this experiment, the *competence* referred to the participant's trust in that the UAVs in the experiment worked competently without failures.

The competence measure was analysed using the non-parametric test of the means as the data set of the baseline and the evaluation scenario was not normally distributed, even after an attempt was made to transform the distribution to normal. The outcomes from the assumptions testing presented in detail in Appendix C yielded that a non-parametric hypothesis statistical test was to be used for analysis.

Table 7.20: Summary table of Wilcoxon Signed-Rank test comparing the participants' perceived competence about the system during the scenario with opaque autonomy information and the transparent autonomy information

$(\alpha = 0.05, n = 34)$	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.033$, Reject H_0
1-Tailed	$\bar{x}_{Baseline} < \bar{x}_{Evaluation}$, Reject H_0

Table 7.20 presents the test summary of the non-parametric Wilcoxon Signed-Rank test. From the table, there was significant statistical evidence to reject the null hypothesis (H_0). This suggests that there was a significant difference between the participant's perceived competence of the system in baseline scenario and that of the evaluation scenario.

On further analyses using the one-tailed test, one concluded that the competence the participants reported were indeed significantly higher when they were using the experiment configuration with a higher level of autonomy transparency when compared to the configuration with a lower level of transparency.

7.4.2 Predictability

Predictability determined the extent to which the participant was able to anticipate the system's behaviour with some degree of confidence. In the context of this experiment, the *predictability* referred to the participants' self trust that they were able to predict the UAVs' behaviour with a reasonable confidence.

Comparing the mean predictability was analysed using the non-parametric test as normal distribution was not observed in both data sets, and transformation to correct the data to a normal distribution was not feasible. The outcomes from the assumptions testing presented in detail in Appendix C yielded that a non-parametric hypothesis statistical test was to be used for analysis.

Table 7.21: Summary table of Wilcoxon Signed-Rank test comparing the participants' perceived predictability about the system during the scenario with opaque autonomy information and the transparent autonomy information

$(\alpha = 0.05, n = 34)$	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.046$, Reject H_0
1-Tailed	$\bar{x}_{Baseline} < \bar{x}_{Evaluation}$, Reject H_0

Table 7.21 presents the test summary of the non-parametric Wilcoxon Signed-Rank test. From the table, there was a significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference between the participant's perceived predictability of the system in baseline scenario and that of the evaluation scenario.

On further analyses using the one-tailed test, one concluded that the predictability the participants reported were indeed significantly higher when they were using the experi-

ment configuration with a higher level of autonomy transparency when compared to the configuration with a lower level of transparency.

7.4.3 Reliability

Reliability referred to the extent that the system was free of errors. In the context of this experiment, the amount of trust that the participants had for the UAVs to be able to make error-free judgements and decisions about its situation.

Comparing the mean reliability was analysed using the non-parametric test as normal distribution was not observed in both data sets, and transformation to correct the data to a normal distribution was not feasible.

The outcomes from the assumptions testing presented in detail in Appendix C yielded that a non-parametric hypothesis statistical test was to be used for analysis.

Table 7.22: Summary table of Wilcoxon Signed-Rank test comparing the participants' perceived reliability about the system during the scenario with opaque autonomy information and the transparent autonomy information

$(\alpha = 0.05, n = 34)$	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.025$, Reject H_0
1-Tailed	$\bar{x}_{Baseline} < \bar{x}_{Evaluation}$, Reject H_0

Table 7.22 presents the test summary of the non-parametric Wilcoxon Signed-Rank test. From the table, there was a significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference between the participants' perceived reliability of the system in baseline scenario and that of the evaluation scenario.

On further analyses using the one-tailed test, one concluded that the reliability the participants reported were indeed significantly higher when they were using the experiment configuration with a higher level of autonomy transparency when compared to the configuration with a lower level of transparency.

7.4.4 Faith

Faith was the extent to which the participant has a strong belief and trust that the system performs a task effectively when there might not be proof. In the context of this experiment, the *faith* referred to the trust that the participants had towards the UAVs' ability to perform its modes of operations or task effectively based on their behaviours and observations, rather than the hard facts displayed to them during the trials.

The outcomes of this measure was non-definitive due to the reason that *faith* was not comparable due to the nature of the tests. The baseline scenario configuration certainly required a level of faith exhibited by the participants due to the lack of UAV capability information exchange, and that the information exchange was available in the evaluation scenario, hence, a need for faith was reduced.

The comparison of the mean faith was analysed using the non-parametric test as normal distribution was not observed in both data sets, and transformation to correct the data to a normal distribution was not feasible.

The outcomes from the assumptions testing presented in detail in Appendix C yielded that a non-parametric hypothesis statistical test was to be used for analysis.

Table 7.23: Summary table of Wilcoxon Signed-Rank test comparing the participants' perceived faith about the system during the scenario with opaque autonomy information and the transparent autonomy information

$(\alpha = 0.05, n = 34)$	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.728$, Accept H_0

Table 7.23 presents the test summary of the non-parametric Wilcoxon Signed-Rank test. From the table, there was no significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was no significant difference between the participants' perceived faith of the system in baseline scenario and that of the evaluation scenario.

Due to the comparison of the faith with a modifying variable of system capability information, there was not a definitive condition for comparisons, and this was reflected in the inconclusive outcomes presented in Table 7.23.

7.4.5 Overall Trust

Overall Trust denoted the extent which the participants trust the system overall. In the context of this experiment, the *overall trust* referred to the participant's trust of the UAVs and the reporting system.

The outcomes of this measure was non-definitive due to the reason that *overall trust* was not comparable due to the nature of the tests. The baseline scenario configuration required a level of trust to be exhibited by the participants due to the lack of UAV capability information exchange, and that the information exchange was available in the evaluation scenario. Hence, a need for trust was reduced, and that the experiment scenarios were not a progressive trial where the aim was to evaluate the trust based on previous situations of the participant's loss-in-trust.

Comparing the mean overall trust was analysed using the non-parametric test as normal distribution was not observed in both data sets, and transformation to correct the data to a normal distribution was not feasible.

The outcomes from the assumptions testing presented in detail in Appendix C yielded that a non-parametric hypothesis statistical test was to be used for analysis.

Table 7.24: Summary table of Wilcoxon Signed-Rank test comparing the participants' overall trust about the system during the scenario with opaque autonomy information and the transparent autonomy information

($\alpha = 0.05, n = 34$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.082$, Accept H_0

Table 7.24 presented the test summary of the non-parametric Wilcoxon Signed-Rank test. From the table, there was no significant statistical evidence to reject the null hypothesis (H_0), meaning there was no significant difference between the participants' overall trust of the system in baseline scenario and that of the evaluation scenario.

Due to the comparison of the overall trust with an indepedent variable of system's capability information transparency, there was not a definitive condition for comparisons. This was reflected in the inconclusive outcomes presented in table 7.24.

7.5 Result & Analysis: Operator Performance

Three performance parameters were recorded to gauge the operator's objective performance for the opaque autonomy configuration and the transparent autonomy configuration. The parameters were the participant's Initial Response Time (IRT), Event Response Time (ERT) and IOIs found.

This section presented the results and analysis of these three performance measurements through initially testing the assumptions and selecting the appropriate statistical tests for paired comparison of the baseline and the evaluation scenario results. The assumptions and the associated hypothesis were described in Section C.1.

The hypothesis used for the *two-tailed* hypothesis testing stated:

- H_0: There were significant differences between the opaque autonomy scenario configuration and the transparent autonomy scenario configuration
- H_a: There were insignificant differences between the opaque autonomy scenario configuration and the transparent autonomy scenario configuration

The hypothesis for the *one-tailed test comparing the response times* stated:

- H_0: The respective participant's response time in the opaque autonomy scenario configuration was less than the transparent autonomy scenario configuration
- H_a: The respective participant's response time in the opaque autonomy scenario configuration was greater than the transparent autonomy scenario configuration

And the hypothesis for the *one-tailed test comparing the number of IOIs found* stated:

- H_0: The number of IOIs that the participants found was lesser in the opaque autonomy scenario configuration than the transparent autonomy scenario configuration

- H_a: The number of IOIs that the participants found was greater in the opaque autonomy scenario configuration than the transparent autonomy scenario configuration

The analysis of operator performance in this section also utilised the two forms of testing described in Section C.1; the parametric and the non-parametric tests. The parametric used for testing was the Paired Sample T-Test, where the data in both scenario sets must first be tested to be in normal distribution (tested using the Shapiro-Wilk method), and that it must have similar variances (tested using the Levene's method). The non-parametric used for testing was the T-Test*'s* counterpart - the Wilcoxon Signed-Rank test, which did not assume the data to be in normal distribution, however, the data must still have similar variances.

Furthermore, all analysis was conducted with a confidence interval of 95% ($\alpha = 0.05$).

7.5.1 Initial Response Time

The IRT, recorded in seconds, was determined by subtracting the time where each participant initially and correctly responded to any UAVs that required interaction from the epoch when the participant first saw the scenario screen. An example being that UAV_{red} had low autopilot capability, hence, human intervention was required to enable the UAV to move forward. The IRT in this instance is determined by the time when the scenario was first launched to the time where the participant correctly interacted and adjusted the UAV_{red}'s altitude controls. The purpose of this measure was to establish a comparison to assess whether the improved autonomy transparency would impact on the operator's reaction to the normal operations of the UAVs.

The outcomes from the assumptions testing presented in detail in Appendix C yielded that a paired-sample comparison of the means was to be conducted on the data sets that were transformed with the natural log transformation.

Table 7.25: Summary table of paired sample test comparing the participants' IRT during the scenario with opaque autonomy information and the transparent autonomy information

($\alpha = 0.05, n = 33$)	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.082$, Accept H_0

Table 7.25 presented the test summary of the parametric T-Test. From the table, there were insignificant statistical evidence to reject the null hypothesis (H_0). That is to say, there was no significant difference between the IRT of the baseline scenario and that of the evaluation scenario.

Based on the test results, there were insignificant evidence to suggest that the configuration of the amount of autonomy transparency available impacted on the IRT performance of the participants.

7.5.2 Event Response Time

The ERT, recorded in seconds, was determined by measuring the time when each participant interacted with the necessary on-screen component of the UAV that would address the encountered event, from the time when the event was initially presented. An example being that UAV_{green} had a moderate level of flight path replanning autonomy, and the operator was required to select a new flight path generated by the UAV. The ERT in this instance is the time where the possible new path options were presented by the UAV subtracted from the time that the operator made a new path selection. The purpose of this measure was to establish a comparison to assess whether the improved autonomy transparency would impact on the operator's reaction time to mission-critical situations, improving the possibility for greater mission success.

The outcomes from the assumptions testing presented in detail in Appendix C yielded that a paired-sample comparison of the means was to be conducted on the datasets. The outlier was removed and the logrithmic transform was applied to both datasets prior to further analysis.

Table 7.26: Summary table of paired sample test comparing the participants' ERT during the scenario with opaque autonomy information and the transparent autonomy information

$(\alpha = 0.05, n = 33)$	**Baseline & Evaluation Scenario**
2-Tailed	$p = 0.001$, Reject H_0
1-Tailed	$\bar{x}_{Baseline} > \bar{x}_{Evaluation}$, Reject H_0

Table 7.26 presented the test summary of the parametric T-Test. From the table, there was significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference between the ERT of the baseline scenario and that of the evaluation scenario.

The further investigation using a one-tailed test revealed that at CI = 95%, there was significant statistical evidence to reject H_0, meaning that the ERT of the baseline scenario was higher than that of the evaluation scenario.

This suggested that the difference between the mean reaction time of the participants as they confronted perturbation events where the interface configuration included minimal amount of autonomy transparency, was indeed higher than the configuration which included large amount of autonomy transparency.

7.5.3 Items-Of-Interest Found

The IOIs found, recorded as a percentage, was determined by the number of IOIs found by the participant (not including highly autonomous searching capable UAVs) divided by the total number of IOIs required to be found by the participant in all search zones. An example is that the participant found 20 IOIs during the baseline scenario, and there were a total of 40 IOIs deployed. However, 10 of those were designed to be found autonomously, which

means that 30 IOIs were designed to be found by the participant. The IOIs found in this example would be determined and recorded as 20/30 = 66.7%. The purpose of this measure was to establish a comparison to assess whether the improved autonomy transparency would improve on the operator's mission capability beyond the management and controls of the UAVs.

The outcomes from the assumptions testing presented in detail in Appendix C yielded that a paired-comparison of the means was to be conducted on the data sets where the three outliers were removed.

Table 7.27: Summary table of paired sample test comparing the percentage of IOIs found by the participants during the experiment

$(\alpha = 0.05, n = 33)$	**Baseline & Evaluation Scenario**
2-Tailed	$p < 0.001$, Reject H_0
1-Tailed	$\bar{x}_{Baseline} < \bar{x}_{Evaluation}$, Reject H_0

Table 7.27 presents the test summary of the parametric T-Test. From the table, there was significant statistical evidence to reject the null hypothesis (H_0). That is to say, there was a significant difference in the number of IOIs found between the baseline and evaluation scenario configurations.

The further investigation using a one-tailed test revealed that at CI = 95%, there was significant statistical evidence to reject H_0. This means that the number of IOIs found by the operator increased in the evaluation scenario over the baseline scenario.

This suggested that the participant's searching performance was significantly increased in a system where the UAVs' autonomy capabilities were well communicated to the participants - being transparent, compared to the configuration where the autonomy was opaque.

7.6 Discussion & Conclusion

The aim of this chapter was to present the implications and effects of increasing the UAV's functional autonomy transparency by communicating its functional status through a text-based dialogue system. This system utilised the natural-language dialogue principle, which enabled the machine agents (UAVs) to effectively and directly "talk" to the human agents (participants), exchanging functional and autonomous capability information with the human agents throughout the scenario and requesting human input/assistance where applicable. Through this system, a rudimentary dialogue was carried out between the human and the machine agents when the machine agent requested assistance, the human agent physically acknowledged the message line and manually took over the jurisdiction of the operative mode.

The success was gauged through three groups cognitive metrics and one group of objective performance metric. The three groups of cognitive metrics included the participant's CW, SA and the trust in the system, the one group of objective performance

Table 7.28: A summary of the participants' cognitive performance results and statistical significance outcome

Metrics	Dimensions	$p-value$	Verdict
CW	Mental Demand	$p = 0.043$	Improved
	Physical Demand	$p = 0.208$	No Different
	Temporal Demand	$p = 0.518$	No Different
	Performance	$p = 0.06$	No Different
	Effort	$p = 0.06$	Improved
	Frustration	$p = 0.704$	No Different
	Combined	$p = 0.022$	Improved
SA	Level 1	$p = 0.002$	Improved
	Level 2	$p = 0.308$	No Different
	Combined	$p < 0.001$	Improved
Trust	Competence	$p = 0.033$	Improved
	Predictability	$p = 0.046$	Improved
	Reliability	$p = 0.025$	Improved
	Faith	$p = 0.728$	No Different
	Overall Trust	$p = 0.082$	No Different

metric included the IRT, ERT, and the total number of IOIs found.

CW was defined in six dimensions by NASA-TLX [97], SA was defined in three levels by SAGAT [76], and Trust was defined in five dimensions [220] . A paper-based rating form was used to capture these metrics and analysed individually.

Finally, the participant's objective performance (IRT, ERT, and IOIs found) [43] were recorded in an in-program logging feature as well as the manual reviews of the in-experiment video footage and analysed individually.

Table 7.28 presents a summarised list of the individual and the combined cognitive performance results, their significance value, as well as the outcome of the hypothesis testing of the means carried out comparing the evaluation scenario score to the baseline scenario score.

A Paired Sample T-Test was conducted to compare the CW and SA of a human operator managing multiple UAVs in opaque autonomy transparency and transparent autonomy information conditions. There was a significant reduction in the overall operator CW ($\bar{x} = 67.06\%$, $\sigma = 9.727$) and overall increased in SA ($\bar{x} = 0.610$, $\sigma = 0.012$) in the transparent navigation autonomy information configurations and opaque transparency ($\bar{x_{CW}} = 59.95\%$, $\sigma_{CW} = 16$, $\bar{x_{SA}} = 0.514$, $\sigma_{SA} = 0.012$) conditions: $N = 33$, $p_{CW} = 0.022$, $p_{SA} < 0.001$. No significant difference in the overall Trust between the conditions: $N = 33$, $p_{Trust} = 0.082$.

Table 7.29 presents a summarised list of the objective performance metric results and outcomes. There was no significant difference in the participant's IRT between the two conditions: $N = 33$, $p = 0.082$. However, a significant improvement was observed in the participant's ERT ($\bar{x} = 3.834$ secs, $\sigma = 0.424$) and the number of IOIs found ($\bar{x} = 45.577$, $\sigma = 9.668$). This outcome was desirable as an improvement in the ERT

Table 7.29: A summary of the participants' objective performance results and statistical significance outcome

Metrics	$p - value$	**Verdict**
Initial Response Time	$p = 0.082$	No Different
Event Response Time	$p = 0.001$	Improved
IOIs Found	$p < 0.001$	Improved

suggested that the time for an operator to react to situational change were reduced as a result of increase autonomy transparency, consequently, unexpected events that the machines were unable to handle could be addressed by the human teammates at the earliest possible time. Furthermore, increased transparency assisted in the human agent's mental process of developing a working model of the UAV systems, hence, the human agents were able to devote greater cognitive resource to perform mission tasks, resulting in an improvement in their objective achievement.

The combination of the results presented in Chapter 6 and this chapter strongly support the hypothesis that the increased autonomy transparency improves the human agent's (operator's) cognitive and objective performances. As a result, Contribution 3 is made and Research Question 3 is addressed.

CHAPTER 8

Conclusion

In realising the full military and commercial potentials of deploying multiple heterogeneous Unmanned Aerial Vehicles (UAVs) in close proximity, an effective way to enable one mission commander to simultaneously manage and supervise multiple agents (in a *one-to-many* operation ratio) is needed. This means that not only should the technology associated with the onboard capabilities of the UAVs be cutting-edge, but the interaction between the human operator and these capabilities must also be sophisticated.

The rest of this chapter summarises the findings of this research and its significance to the research problem by addressing each research question.

8.1 Addressing the Research Questions

This research had presented three bodies of work, each aimed at addressing one research question, and the combination of the outcomes had contributed positive evidence to reach the overarching research objective.

8.1.1 Question 1: Functional Transparency

Revisiting the first research question:

> *What are the effects of UAV functional transparency on human operator's Cognitive Workload (CW) and Situation Awareness (SA) in heterogeneous multiple UAV management?*

This question was answered through a three-step process:

1. The proposal of the novel Functional Capability Framework (FCF),
2. Its application and the software visual implementation, and
3. The evaluation of its effects on human CW and SA.

The first step was the proposal of the novel FCF. This step initially incorporated a thorough understanding of the subsystems and functional components of a UAV. The focus was then shifted to establishing the fundamental concepts of human information processing and Ecological Interface Design (EID) principles. Finally, efforts were combined to propose

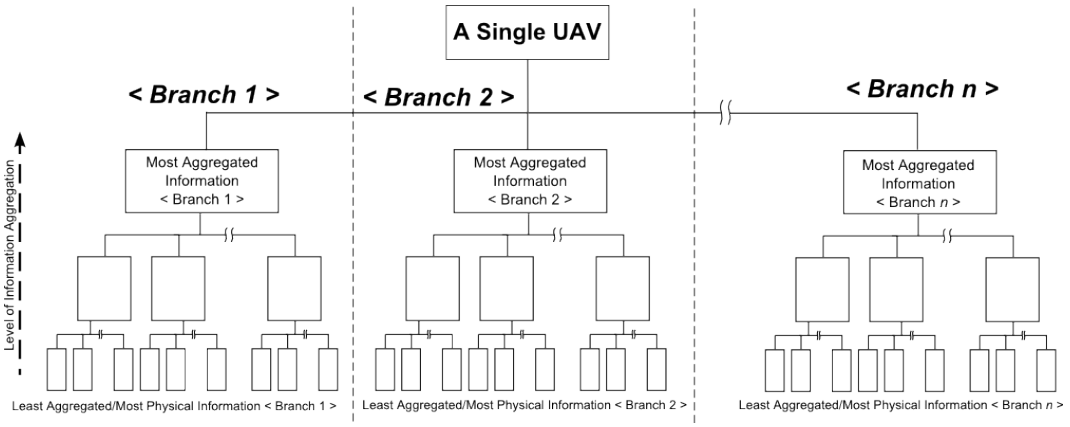


Figure 8.1: A simplified layout of the Functional Capability Framework

a derivation of the FCF that was aimed to be used in a hybrid system, and align with the human operator's ergonomic preferences. The final FCF had incorporated a three-layered (high, medium, and low) abstraction and a four-branched classification of a UAV's functional subsystems. As depicted in Figure 8.1, the highest Level Of Detail (LOD) represented the highest level of abstraction, and only the most aggregated information about each subsystem was included (i.e. a green indicator represented the subsystem health was "good"). The medium LOD represented a moderate level of abstraction. This included a hybrid level of information aggregation (i.e. numerical or graphical representation of sensory data). The low LOD represented the lowest level of abstraction, and only very low to no intelligence was required to collate or synthesise the raw data (i.e. data collected from onboard sensors). The four branches of functional subsystems included the *UAV Health, Navigation, Autopilot & States,* and *Payload.* Each branch was then abstracted into the three LODs, forming the completed FCF.

The second step was the application of the FCF and its software implementation for experimental purposes. This step required the analysis of a typical tactical surveillance UAV and a breakdown of its onboard subsystems and functionality. It was then applied to the FCF to form a practical FCF. This FCF was then required to be materialised by software implementation, where a set of specifications for the software visualisation and behaviour was proposed. Chapter 5 presented the detailed specification of this experimental software prototype.

Finally, the significance and the effects of the FCF on human CW and SA was verified through an experimental trial. The experiment involved three scenarios; a high LOD scenario, a hybrid LOD scenario (where the LODs of each subsystem was adjustable) and a low LOD scenario. The analysis of the results confirmed that the participants had exhibited a higher cognitive performance when using FCF with a hybrid LOD as compared to unregulated LOD (a low or a high LOD).

8.1.2 Question 2: Partial Autonomy Transparency

Revisiting the second research question:

> *What are the effects of UAV autonomy transparency on human operator's CW and SA when partial capability autonomy is represented visually in heterogeneous multiple UAV management?*

This question was answered by extending the contribution posed by the first research question. The partial capability autonomy referred to was the *navigation* branch autonomy capability, that is, the UAV's ability to re-plan alternative flight paths when they were confronted with potentially catastrophic events (i.e. dangerous weather or no-fly zones).

Firstly, a detailed review of the existing autonomy classification taxonomies was conducted with the aim of applying a recognised scale of autonomy to the functional subsystems of the UAVs in the hybrid system. For this reason, a simplified three-level autonomy scale, inspired by the Sheridan & Verplanck (SV) scale [211] and the Human-Automation Collaboration Taxonomy (HACT) [28], was adopted.

Secondly, this autonomy scope was applied to the *navigation* autonomy in three levels, high, medium and low Functional Level Of Autonomy (F-LOA). High F-LOA was where the autonomy had all the responsibilities and capabilities to acquire information (information autonomy), generate possible alternate flight paths, (solution generation autonomy), decide on the most appropriate flight path (decision autonomy), and carry out the decision (action autonomy). Medium F-LOA was where the autonomy was responsible for information acquisition, solution generation, and action implementation. And low F-LOA was that all the steps were the responsibility of the human operator.

Thirdly, given the aim of this research question was to investigate the effects of partial autonomy transparency on human cognitive performance, a visualisation of this autonomy was established. Graphical representation of the flight path was the chosen visualisation method to investigate the impact. The reason for this was because graphical representation was both visually comprehensible and non-intrusive. Hence, it was easy for the participants to comprehend the situation. The experimental software prototype was then modified to incorporate the autonomy capabilities and the visual implementations of the flight path.

Finally, the effects of the navigation autonomy transparency was verified through an experimental trial involving two scenarios; one had an interface that did not support autonomy transparency, while other other supported autonomy transparency. Taking the work established from research question 1, participants were invited to perform the experiment and reported their CW and SA. The analysis of the results had confirmed that with a 95% confidence, the participants had exhibited positive cognitive performance improvements with the interface that supported partial autonomy transparency.

8.1.3 Question 3: Complete Autonomy Transparency

Revisiting the third research question:

> *What are the effects of UAV autonomy transparency on human operator's CW, SA and Trust when the UAV's autonomy is represented visually in heterogeneous multiple UAV management?*

This question was answered by combining the first two research questions and extending the second research question. The aim of this research question was to investigate the effects of the (*complete*) autonomy capability's transparency (as opposed to *partial*) on the human operator's CW and SA, with the addition of system/automation Trust. A three-step process was invoked.

Firstly, a revision of the existing taxonomies in classifying LOAs was conducted with the aim of selecting a methodology that not only encapsulated the different modes of operation throughout a similar search mission to the experiments conducted for Questions 1 and 2, but also demonstrated use on other multiple UAV reconnaissance missions. For this reason, the *autonomy spectrum* was selected because it satisfied both the requirements from a human-centred perspective and had the capacity to incorporate all four subsystem capabilities (health, navigation, states, and payload) proposed in research question 1.

Secondly, a review of the modalities of human-machine interaction was conducted. In selecting a suitable mode to visually represent the full spectrum of the UAVs' autonomy capabilities, a message box that included messages that were constructed from plain English was proposed. Natural language communication was determined most direct and non-intrusive in communicating information to a human. The messages that appeared in the message box communicated the autonomy capabilities of the agents in three parts: the identity of the message sender (UAV), its autonomy capability to perform work (desire), and whether the desired work can be achieved (intention). This approach provided two advantages, 1) The messages were constructed from plain, first person, simple English- this meant that the message could be simply understood in a short amount of time during the mission, and 2) The operator had the capacity to revisit any message at any time, and he/she was not required to rely on his/her memory ability.

Finally, the effects of all-encompassing autonomy transparency was verified through an experimental trial involving two scenarios: one had an interface configured to show the message box, which supported autonomy transparency, while the other had not. Participants were invited to perform an experiment with these two configurations tested individually, and reported their CW, SA, and Trust in the system. The analysis of the results had confirmed that with a 95% confidence, the interface that was configured to support autonomy transparency had shown significant statistical improvements in both CW and SA, while no reduction of Trust was evident. Furthermore, evidence suggested that the inclusion of capability transparency had improved the human operator's objective mission performance.

This confirmed that autonomy transparency is a significant factor in reducing cognitive overhead in human-machine interaction.

8.2 Contribution

The focus of this research was to investigate the impacts of a UAV's capability transparency on a single human operator's performance in the context of managing multiple UAVs that had heterogeneous levels of functional autonomy. Through the course, five contributions have been made (Section 1.2.3).

Contribution 1 *A novel FCF to encapsulate agent subsystem functional capability and its derivation procedure.* The framework was proposed as a part of answering Research Question 1, and its derivation procedure was demonstrated through the design and development of the experiment software prototype used throughout this research.

Contribution 2 *The concept of Functional-LOA to enable a system's autonomy to be classified in its separate capabilities.* F-LOA was proposed as a part of answering Research Question 2 and further used to answer Research Question 3. It aimed at describing a UAV's autonomy levels based on it's functional capabilities, which had not been seen in previous research.

Contribution 3 *A novel approach to support natural language communication of autonomy capability in a hybrid system.* The message box approach proposed in Research Question 3 demonstrated a validated approach to support natural language communication of the UAVs' autonomy capabilities.

Contribution 4 *An experimental software prototype that is customisable, and incorporates all elements proposed in contributions 1 and 2.* The experiment software that is designed and developed has been implemented and used through the course this research to verify novel approaches and solutions to the research questions.

Contribution 5 *A complete set of experimental scenario designs, procedures, and results.* The scenario designs and experiment procedures were documented and presented in this thesis to support data collection. The analysed results that were collected during the experimentation also formed a part of this contribution.

This research had arrived at a terminus with a verified, novel approach (Figure 8.2) to positively incorporate UAV capability transparency in future designs of multiple heterogeneous UAV management interfaces. The findings presented throughout this thesis had successfully reached the overarching research objective.

The presence of functional and autonomy capability transparency had demonstrated a significant and positive affect on the human operator's CW, SA, Trust and objective performance. The transparency can be achieved through incorporating capability information in the design of future interfaces for the purpose of managing multiple heterogeneous UAVs in a common mission.

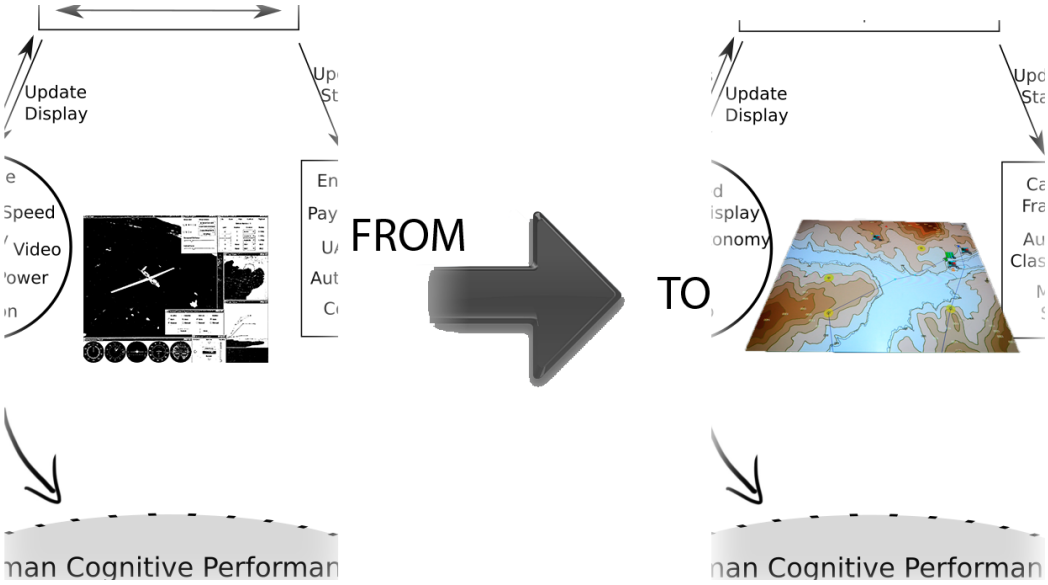


Figure 8.2: The images in the centre of the illustration on the left depicts an example of the existing piloting interface, the illustration on the right depicts the proposed command and control interface.

8.3 Research Limitation

Conducting rigorous, scientific research in the space of the research topic of this doctoral thesis is extremely challenging. As such, defining the research scope is important to ensure the scientific merit achieved in this thesis is still relevant in the larger problem space. In identifying the scope, three main constraints were in place and discussed in this section.

8.3.1 Transparency Visualisation and Interface Designs

This research placed much effort on the structure of information presentation, while lower emphasis was placed on interface design principles. The interface design principles are another distinct topic, even though it is well studied. However, for the particular interface designed as a part of this research, there are limitations.

The author understood that a suboptimal interface design can reduce the end user's cognitive performance. However, it should not impact the results of the studies in a comparative nature. Based on the findings of this research, further studies are required to achieve an optimal information transparency visualisation technique, and better interface design guidelines for future designs of user interfaces involving command and control of multiple unmanned systems.

8.3.2 Experiment Scenario Realism

The multidisciplinary nature of this research requires that a fine balance between ecological validity and experimental control be established. It is important that an abstract and simplification of a real world scenario must be conducted and implemented as experiment tasks to ensure study participants can obtain sufficient task competency to perform the experiment. Equally as important is to ensure that the tasks are not overly simplified to remove any possibility of generalising the findings to a real-world situation.

The participants available for this research were limited to lay people with very little to no experience in the UAV operations area. Therefore, the findings within this study are limited to human operators of multiple UAVs who have not received any specialised training. Further research is required to better understand the effects of capability-information transparency on professional UAV operators and mission commanders.

8.3.3 Experiment Task Familiarisation

Experiment design was focused on exposing the participant to an acceptable level of software prototype usage competency to produce experimental data. The familiarisation period built into the experiment will not be sufficient to allow participants a complete and highly proficient usage of the prototype.

This limitation was realised and consequences were considered to be acceptable to produce results for comparative studies. However, future studies can consider improving the familiarisation process, or if resources allow, using participants who had prior training in using the software system for the experiment.

8.4 Recommendations & Future Work

The concepts proposed in this research provide a verified approach for future research in three potential directions: theoretical development, practical enhancement, and physical implementation.

Theoretical development can be pursued further in the investigation of the quality and the quantity of UAV capability information, specifically the level of information transparency and its effects on human performance (as very recently commenced by *Mercado et al.* [145]). The scalability of the transparency approach (message box), and design implications that can further improve the ergonomics and usability of the interfaces.

Practical enhancements can be further pursued in the expandability and transferability of the FCF to other, more complex unmanned platforms. The potential applications of the FCF can potentially benefit a wide range of autonomous systems. Hence, further evaluative assessments should be performed to assess its feasibility in other heterogeneous, multi-agent domains.

Finally, the physical implementation development can be focused on the experimental platform by adapting the existing software prototype and expanding its capabilities., These

can include different modalities of interactions, i.e. haptic control, gesture recognition, and audio enhancements. The successful and significant implementation of these features can drive the future application of this software towards the real world. Ultimately, with the successful demonstration of its viability, the software could be commissioned for real world deployment to manage multiple heterogeneous UAVs.

With the rapid growth of the platform-level capabilities, the deployment for multiple heterogeneous UAVs in both the military and the civilian domain is near. Hence, the recommendations presented in this thesis, although framing a wide spectrum for future work shall be considered a priority.

APPENDIX A

Experiment 1 Result Analysis

A.1 Cognitive Workload Assumptions Testing

This section detailed the process of assumptions testing of the Cognitive Workload (CW) results collected in experiment 1 to support the statistical analyses presented in Chapter 5.

A.1.1 Mental Demand

Table A.1 presented the descriptive statistics of MD data for segments A, B and C. The outcomes of the Shapiro-Wilk's numerical tests of normality revealed that the data was not normally distributed. The hypothesis of the test of normality states:

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

The significance value (p) for the three segments were 0.019, 0.032, and 0.042 respectively, hence, with a Confidence Interval (CI) of 95%, one concluded that the the null hypothesis (H_0) could not be accepted. Furthermore, the numerical results of the skewness

Table A.1: Summary of the descriptives for the mental demand in the experiment

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**
Normal Distribution	Numerical: Reject $H_0(p = 0.019)$	Numerical: Reject $H_0(p = 0.032)$	Numerical: Reject $H_0(p = 0.042)$
Standard Deviation	$\sigma = 3.63$	$\sigma = 4.30$	$\sigma = 4.39$
Skewness	-0.816 $(SE = 0.472)$: Negative skew	-0.650 $(SE = 0.472)$: Negative skew	-0.761 $(SE = 0.472)$: Negative skew
Kurtosis	-0.426 $(SE = 0.918)$: Acceptable kurtosis	-0.563 $(SE = 0.918)$: Acceptable kurtosis	-0.316 $(SE = 0.918)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	13.29 (66.45%)	13.25 (66.25%)	14.17 (70.85%)

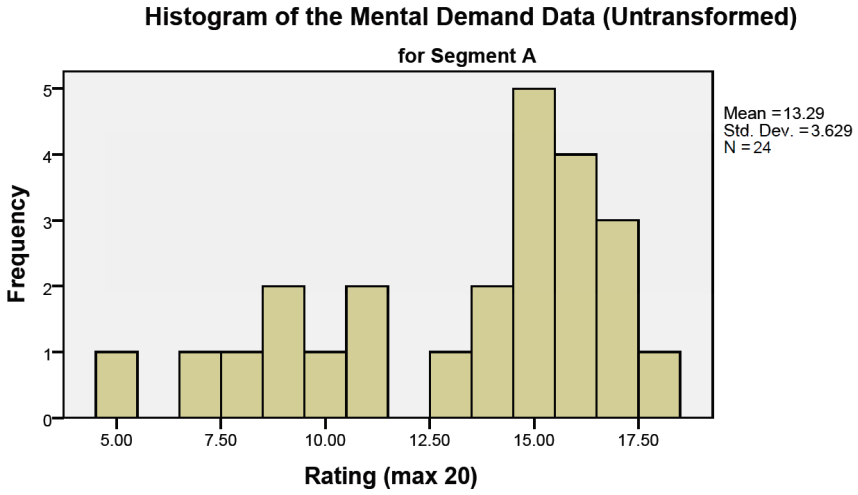


Figure A.1: Histogram of the untransformed mental demand data of segment A

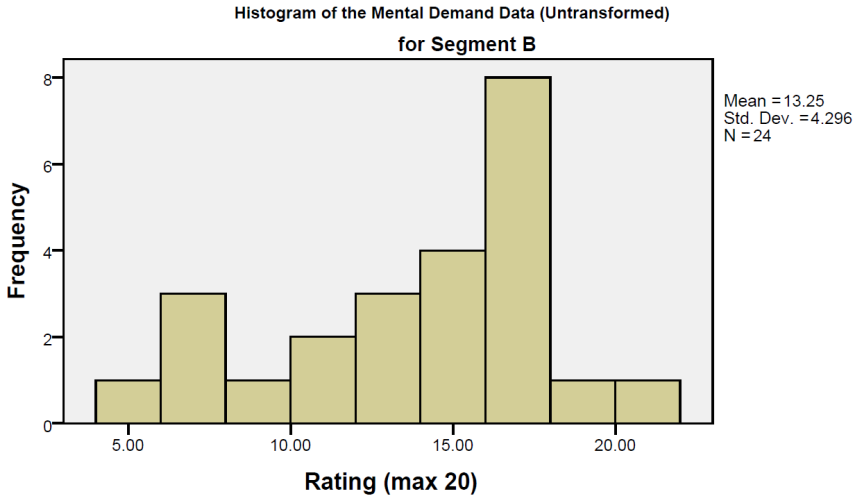


Figure A.2: Histogram of the untransformed mental demand data of segment B

analysis data were -0.816, -0.650, and -0.761 with a Standard Error (SE) of 0.472 for the three segments respectively, and figureA.1, A.2 and A.3 presented the histograms of the raw data, which also supported the negative skew, hence, contributing to the non-normal distribution. Therefore, a square root transformation $(X_{i_{tr}} = X_i^2)$was applied to adjust the distribution, enabling a parametric testing to be carried out.

Table A.2 illustrated the descriptive statistics for the transformed data.

The descriptives of the transformed MD data illustrated in table A.2 supports the three assumptions for a parametric T-Test; 1) Data was assumed to be normally distributed, 2) All the sample must be independently collected, and 3) There must be similar variances in the means of the datasets.

The first assumption was verified numerically with $p_A = 0.073$, $p_B = 0.182$, and $p_C = 0.235$ (where the subscripts denoted the segment identifier) greater than $\alpha =$

Table A.2: Summary of the descriptives and parametric test assumptions testing for the transformed mental demand data

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segments A & B**	**Segments B & C**
Normal Distribution	Graphical: Accept H_0 Numerical: Reject $H_0(p = 0.073)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.182)$	Graphical: Accept H_0 Numerical: Reject $H_0(p = 0.235)$	-	-
Standard Deviation	$\sigma = 87.42$	$\sigma = 104.32$	$\sigma = 112.54$	-	-
Skewness	-0.442 ($SE = 0.472$): Acceptable skewed	-0.1 ($SE = 0.472$): Acceptable skew	-0.232 ($SE = 0.472$): Acceptable skew	-	-
Kurtosis	-1.059 ($SE = 0.918$): Flatter curve	-0.674 ($SE = 0.918$): Acceptable kurtosis	-0.928($SE = 0.918$): Flatter curve	-	-
Sample Mean ($\bar{x}$)	189.29	193.25	219.17	-	-
Variances in the means	-	-	-	Accept $H_0(p = 0.393)$	Accept $H_0(p = 0.705)$

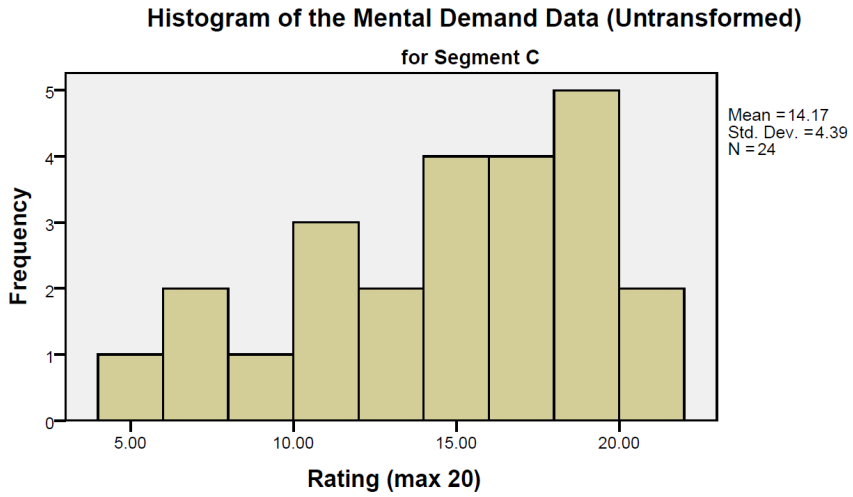


Figure A.3: Histogram of the untransformed mental demand data of segment C

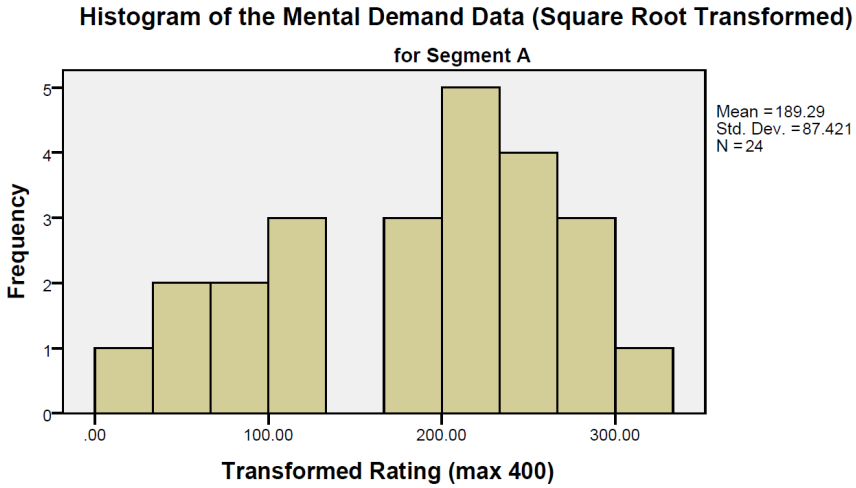


Figure A.4: Histogram of the transformed mental demand data of segment A

0.05, hence, at CI=95%, $H_0$was accepted, which suggested that the transformed data was normally distributed. The histogram illustrated in figues A.4, A.5 and A.6 presented an improved distribution of the data; correcting the skewness that was presented in the raw data. Furthermore, the Q-Q plots (figure A.7, A.8 and A.9) illustrated the data points to be following the quantile line, hence confirming that the distribution of the transformed data was normal.

Secondly, the author/experimenter noted that all samples were collected independently, hence the second assumption was assumed.

Finally, similar variances were achieved with $p_{A|B} = 0.393$ and $p_{B|C} = 0.705$ (where the subscripts $A|B$ and $B|C$ denoted the variances between Segments A and B, and B and C) being greater than the type-one error $\alpha = 0.05$, hence, at CI=95%, H_o was accepted given the hypothesis of the variance testing was:

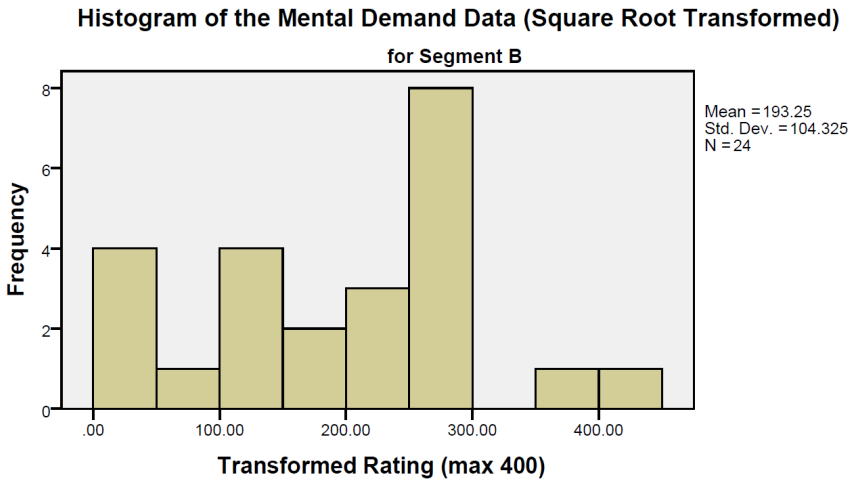


Figure A.5: Histogram of the transformed mental demand data of segment B

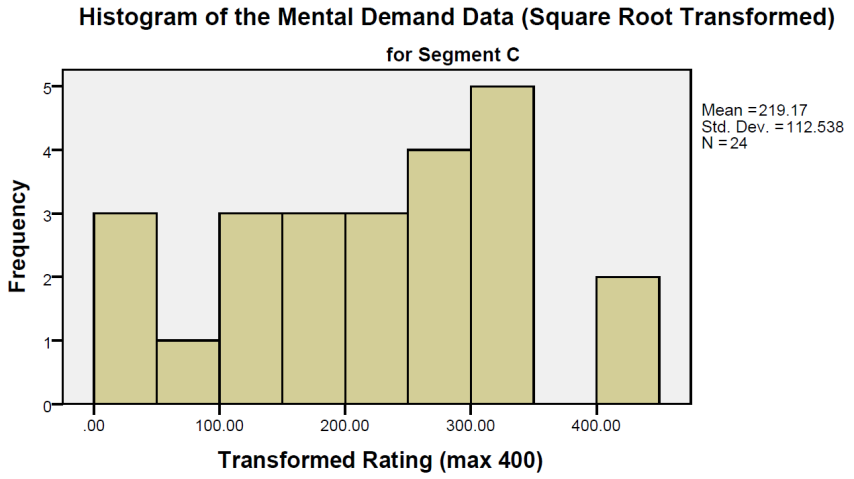


Figure A.6: Histogram of the transformed mental demand data of segment C

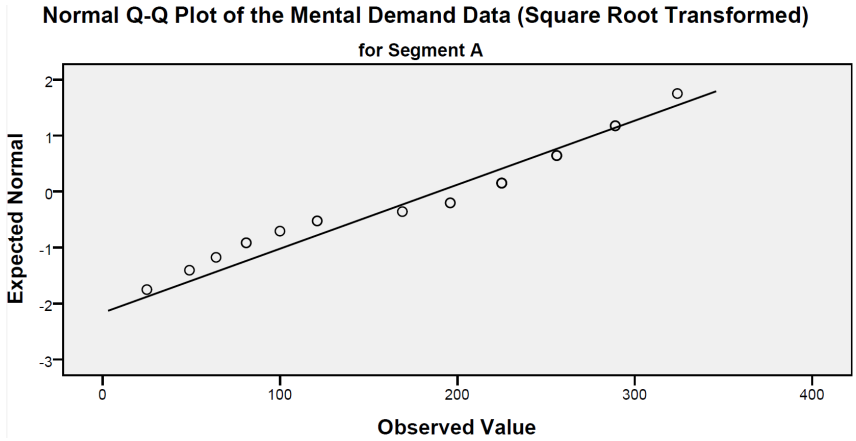


Figure A.7: Q-Q plot of the transformed mental demand data of segment A

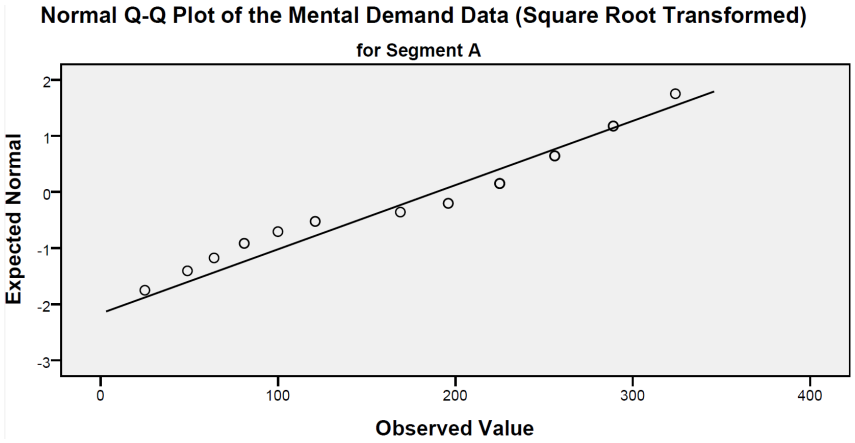


Figure A.8: Q-Q plot of the transformed mental demand data of segment B

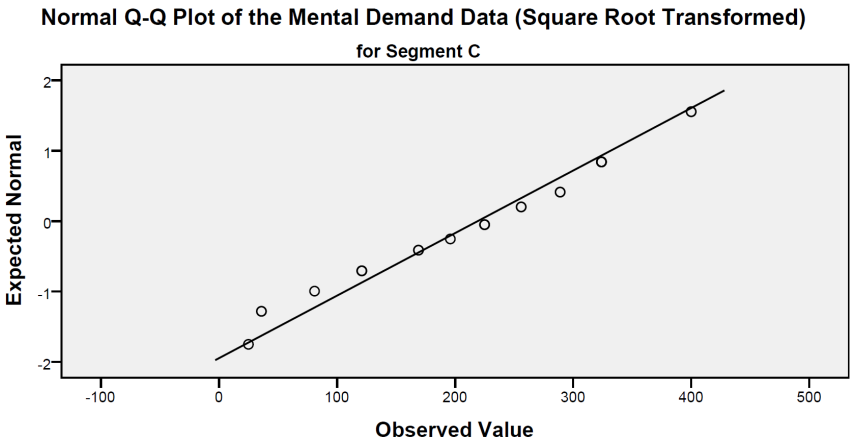


Figure A.9: Q-Q plot of the transformed mental demand data of segment C

- H_0: The pair of sample data set had similar variances
- H_a: The pair of sample data set did not have similar variances

Hypothesis testing of the means were conducted using the parametric paired-sample T-Test on the transformed data, and the non-parametric Wilcoxon Signed-Rank test on the original/raw data.

A.1.2 Physical Demand Assumptions Testing

Table A.3 presented the descriptive statistics of PD data for segments A, B and C. The outcomes of the Shapiro-Wilk's numerical tests of normality revealed that the data were not normally distributed. The hypothesis of the test of normality states:

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

Table A.3 illustrated the numerical test result for testing the distribution of normality of segments A, B and C. The significance value (p) of segment A suggested that at CI =

Table A.3: Summary of the descriptives for the physical demand data

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segment B & C**
Normal Distribution	Numerical: Reject $H_0(p = 0.001)$	Numerical: Accept $H_0(p = 0.063)$	Numerical: Accept $H_0(p = 0.050)$	-
Standard Deviation	$\sigma = 3.87$	$\sigma = 4.44$	$\sigma = 5.19$	-
Skewness	1.265 $(SE = 0.472)$: Positive skew	0.458 $(SE = 0.472)$: Acceptable skew	-0.099 $(SE = 0.472)$: Acceptable skew	-
Kurtosis	-0.724 $(SE = 0.918)$: Acceptable kurtosis	-0.929 $(SE = 0.918)$: Flatter curve	-1.089 $(SE = 0.918)$: Flatter Curve	-
Sample Mean ($\bar{x}$)	5.08 (25.4%)	7 (35%)	8.5 (42.5%)	-
Variances in the means	-	-	-	Accept $H_0(p = 0.395)$

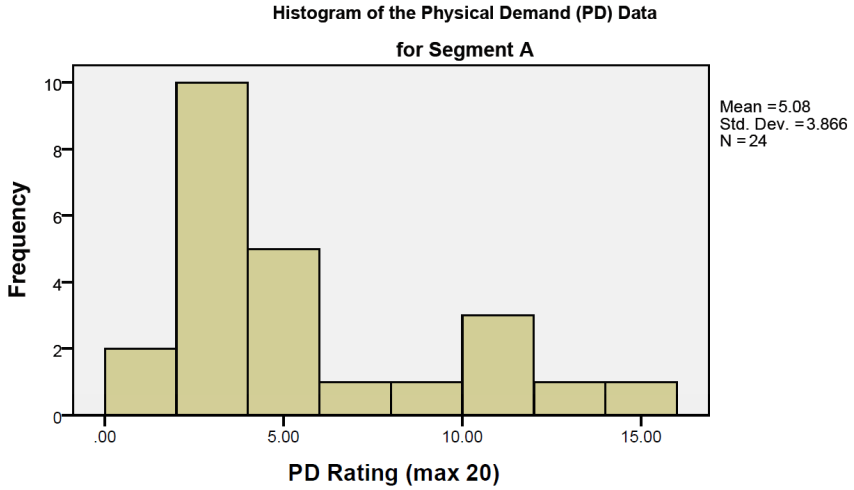


Figure A.10: Histogram of the physical demand data for segment A

95%, H_0 was to be rejected, while the H_0 of segment B and C was accepted. Furthermore, the skewness result suggested a positive skew in the data, and this was illustrated in the histogram of the PD data for segment A (figure A.10), hence a logarithmic transform was applied to segment A and B.

Table A.4 illustrated the descriptive statistics for the segment A and B's transformed data set. The numerical evaluation of the distribution of this new set of data exhibited

Table A.4: Summary of the descriptives for the physical demand data from segment A and B using the logarithmic transform

$(CI = 95\%, \alpha = 0.05)$	**Segment A**	**Segment B**	**Segment A & B**
Normal Distribution	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.332)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.086)$	-
Standard Deviation	$\sigma = 0.32$	$\sigma = 0.34$	-
Skewness	0.073 $(SE = 0.472)$: Acceptable skew	0.522 $(SE = 0.472)$: Positive skew	-
Kurtosis	-0.500 $(SE = 0.918)$: Acceptable kurtosis	-0.720 $(SE = 0.918)$: Acceptable kurtosis	-
Sample Mean $(\bar{x})$	0.59 (45.35%)	0.74 (56.88%)	-
Variances in the means	-	-	Accept $H_0(p = 0.633)$

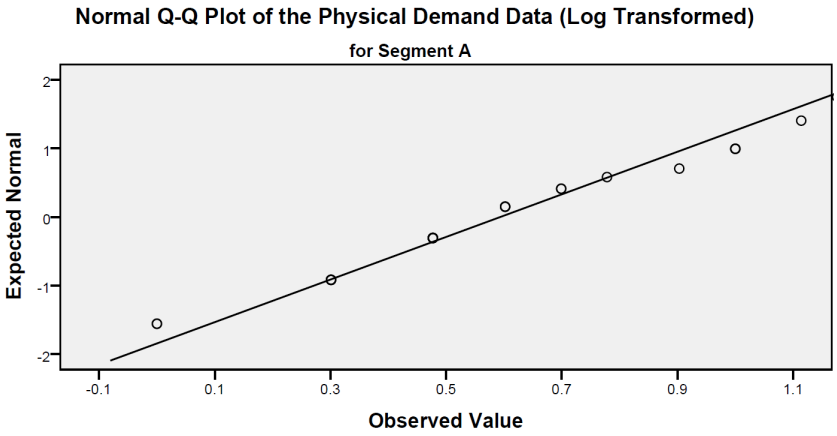


Figure A.11: Q-Q plot of the transformed physical demand data for segment A

normality with a CI = 95%, this means, H_0 of the test of normality hypothesis was accepted numerically. Furthermore, the Q-Q plots illustrated in figures A.11 and A.12 also supported the normal distribution of the new data set.

The second assumption being the independence of data, where the author/experimenter noted that all samples were collected independently, hence the second assumption could be assumed.

Finally, equal variances were achieved with $p_{A|B} = 0.395$ and $p_{B|C} = 0.633$ (where the subscripts $A|B$ and $B|C$ denoted the variances between the transformed data from segments A and B, and the original data from B and C) being greater than the type-one error, hence, at CI = 95%, H_o was accepted given the following variance testing hypothesis:

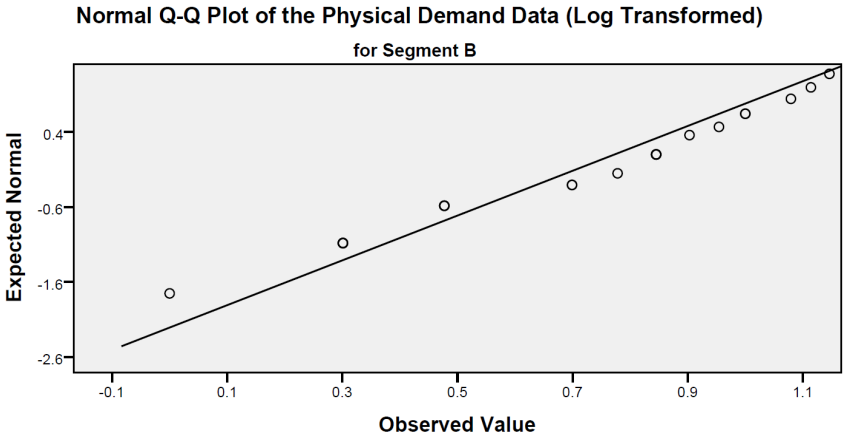


Figure A.12: Q-Q plot of the transformed physical demand data for segment B

- H_0: The pair of sample data set had similar variances
- H_a: The pair of sample data set did not have similar variances

Hypothesis testing of the means were conducted using the parametric paired-sample T-Test on the transformed data from segments A and B, and the raw data from segments B and C.

A.1.3 Temporal Demand

Table A.5 illustrates the outcomes of the assumptions testing for the parametric T-Test's for all three segments.

The test of normality revealed two outcomes, a numerically analysed outcome using the Shapiro-Wilk's method for a sample size $N = 24$, and a graphically analysed outcome using Q-Q plots. The assumptions for the test of normality states:

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

The significance value of $p = 0.048$ and $p = 0.005$ for segments A and C respectively, one can conclude that with a CI of 95%, H_0 could not be considered to be accepted, while the significant value of $p = 0.048$, one can conclude that with the same CI, H_0 of segment B was considered to be accepted.

On the other hand, a graphical analysis illustrated in figure A.13, A.14 and A.15 for segments A, B and C respectively showed that most of the sample points (circles) closely followed the trend line $y = x$, which also suggests the conclusion that H_0 was accepted, however, a few outliers were presented. The box plot illustrated in figure A.16, in combination with figures A.13 and A.15, detected that both segment A and C had a presence of outliers. As a result, the parametric tests should not be carried out.

Table A.5: Summary of the parametric T-Test's assumptions testing for the temporal demand in the experiment

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segments A & B**	**Segments B & C**
Normal Distribution	Graphical: Accept H_0 Numerical: Reject $H_0(p = 0.048)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.896)$	Graphical: Accept H_0 Numerical: Reject $H_0(p = 0.005)$	-	-
Standard Deviation	$\sigma = 3.41$	$\sigma = 4.11$	$\sigma = 4.08$	-	-
Sample Size	$N = 24$	$N = 24$	$N = 24$		
Skewness	-0.730 ($SE = 0.472$): Left skewed	-0.377 ($SE = 0.472$): Acceptable skew	-1.486 ($SE = 0.472$): Left skew	-	-
Kurtosis	-0.4 ($SE = 0.918$): Acceptable kurtosis	-0.003 ($SE = 0.918$): Acceptable kurtosis	2.959 ($SE = 0.918$): Acceptable kurtosis	-	-
Sample Mean ($\bar{x}$)	14.04 (70.20%)	12.67 (63.35%)	14.17 (70.85%)	-	-
Variances in the means	-	-	-	Accept $H_0(p = 0.463)$	Accept $H_0(p = 0.706)$

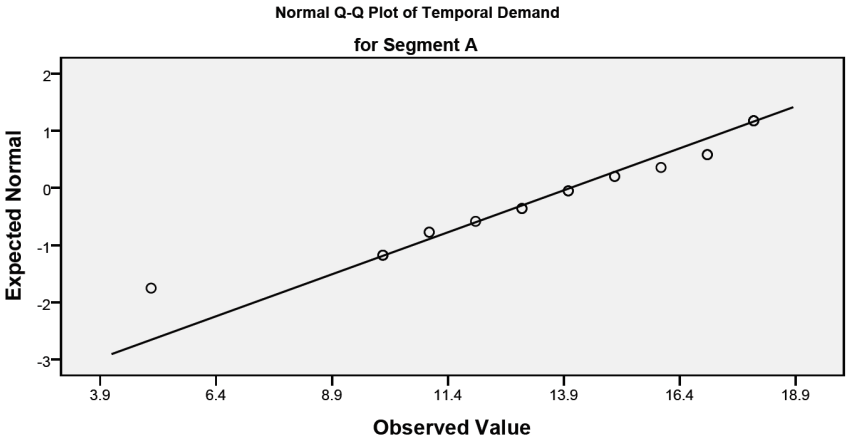


Figure A.13: Q-Q plot for the participants' temporal demand in Segment A

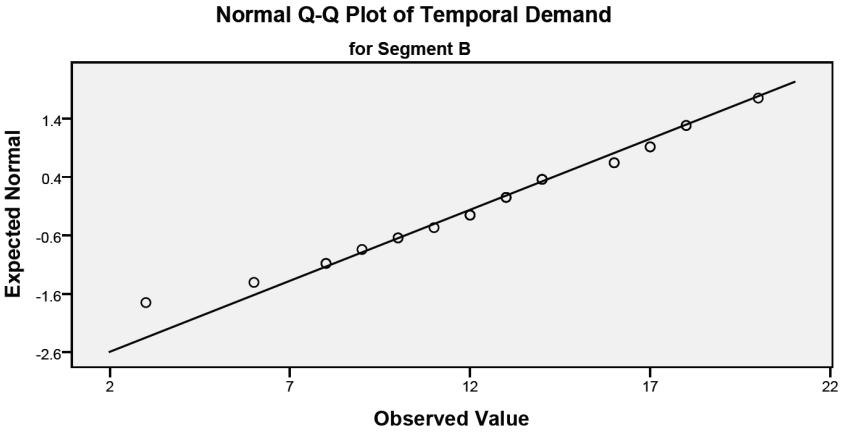


Figure A.14: Q-Q plot for the participants' temporal demand in Segment B

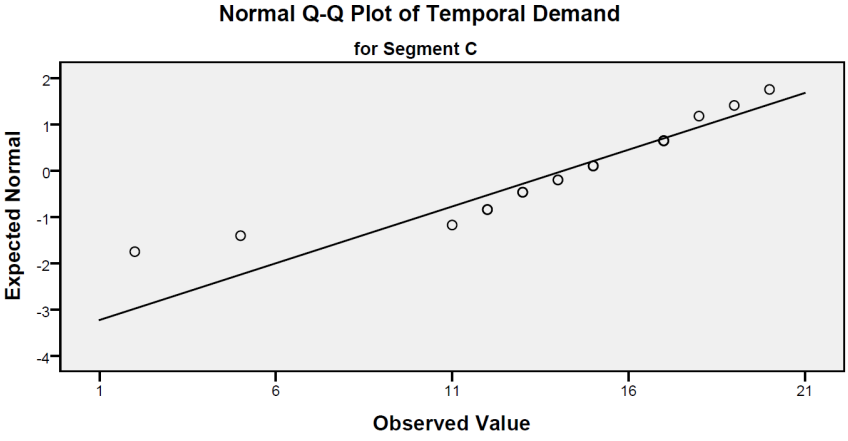


Figure A.15: Q-Q plot for the participants' temporal demand in Segment C

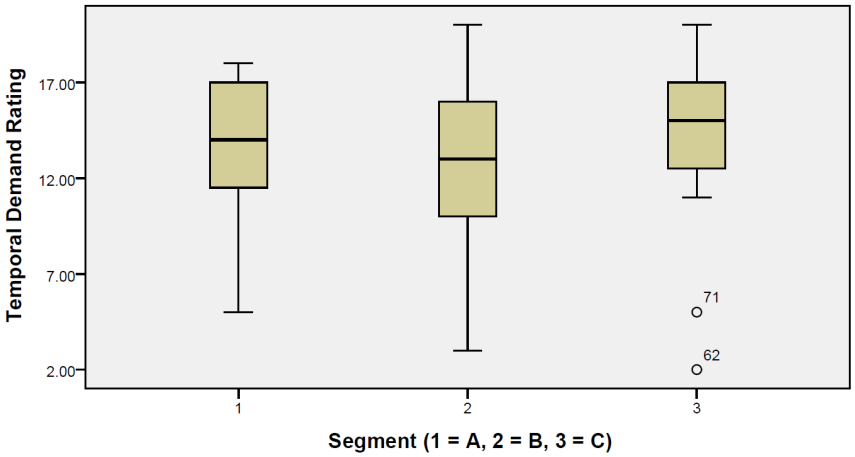


Figure A.16: Box plot for the participants' temporal demand for the three segments

Table A.6: Summary of the parametric T-Test's assumptions testing for the temporal demand with the outlier sample 14 removed (N = 23)

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segments A & B**
Normal Distribution	Graphical: Accept H_0 Numerical: Reject $H_0(p = 0.029)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.911)$	-
Standard Deviation	$\sigma = 2.87$	$\sigma = 3.64$	-
Sample Size	$N = 23$	$N = 23$	
Sample Independence	All samples collected independently as noted by the author/experimenter		-
Skewness	-0.178 $(SE = 0.481)$: Acceptable skew	-0.021 $(SE = 0.481)$: Acceptable skew	-
Kurtosis	-1.327 $(SE = 0.935)$: Flatter curve	-0.602 $(SE = 0.935)$: Acceptable kurtosis	-
Sample Mean $(\bar{x})$	14.43 (72.15%)	13.09 (65.45%)	-
Variances in the means	-	-	Accept $H_0(p = 0.428)$

The outliers with a magnitude of temporal demand load of $X_{14} = 5$ in segment A, and $X_{14} = 2$ and $X_{23} = 5$ in segment C were removed for the analysis to encourage an improved distribution of the data.

To enable equality of comparison, X_{14} for both segments A and B were removed as one pair of comparison, while X_{14} and X_{23} for both segments B and C were removed as a separate pair of comparison.

Table A.7: Summary of the parametric T-Test's assumptions testing for the temporal demand with the outlier samples 14 and 23 removed (N = 22)

($CI = 95\%$, $\alpha = 0.05$)	**Segment B**	**Segment C**	**Segments B & C**
Normal Distribution	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.894)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.311)$	-
Standard Deviation	$\sigma = 3.73$	$\sigma = 2.49$	-
Sample Size	$N = 22$	$N = 22$	
Skewness	-0.024 $(SE = 0.491)$: Acceptable skew	0.133 $(SE = 0.491)$: Acceptable skew	-
Kurtosis	-0.720 $(SE = 0.953)$: Acceptable kurtosis	-0.906 $(SE = 0.953)$: Acceptable kurtosis	-
Sample Mean $(\bar{x})$	13.09 (65.45%)	13.14 (65.70%)	-
Variances in the means	-		Accept $H_0(p = 0.083)$

Tables A.6 and A.7 presented the descriptive statistics of the new data sets with outliers removed, where the first pair (segment A and B) had a new sample size of 23, while the second pair (segment B and C) had a new sample size of 22.

Segment A's new data set presented a non-normal distribution with a significance value of $p = 0.894$ as presented in table A.6, while segment B's new data set presented a normal distribution. Further assessment of the data sets' skewness revealed that no abnormal skewness were detected, hence, a transformation could not be applied. Therefore, the new data set of segments A and B with the removal of the outlier could not pass the assumptions testing for a parametric T-Test, hence, a non-parametric test was used to compair the means.

Table A.7 illustrated that normal distributions of the data for segment B and C were presented both graphically (figures A.17 and A.7) and numerically, hence, no modification of the data was required.

The second assumption was to assume the dataset was independently sampled, which was satisfied given the procedure of the experiment and data collection.

The third assumption was to assume the data had similar variances between the means. This was analysed numerically using Levene's test of variances. TableA.7 illustrated the significance value of the variances in the means between segment B and C. The hypothesis for the test of variance in the means:

- H_0: The pair of data had similar variances in the means

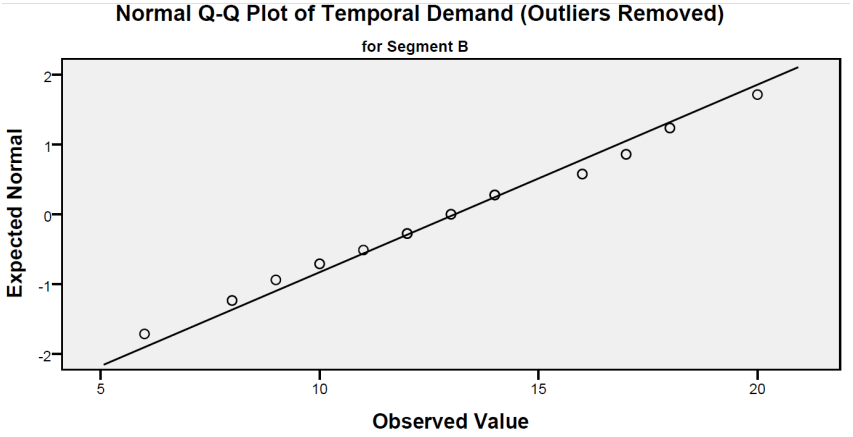


Figure A.17: Q-Q plot for the participants' temporal demand in Segment B

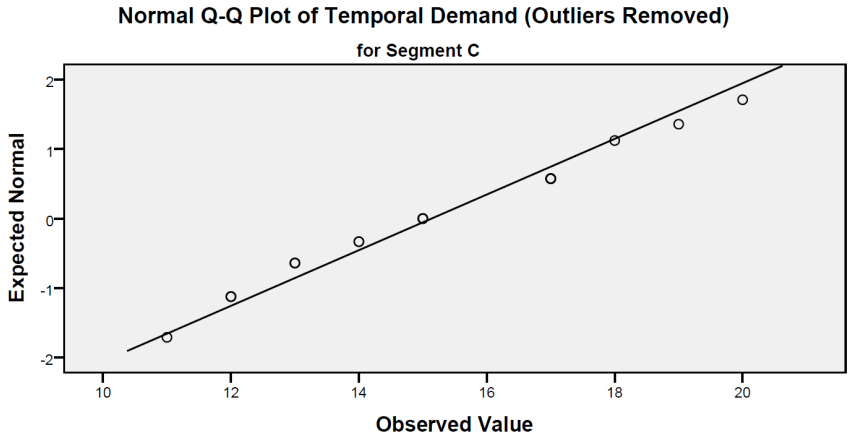


Figure A.18: Q-Q plot for the participants' temporal demand in Segment C

- H_a: The pair of data did not have similar variances in the means

Given the assumptions were satisfied through assumptions testing, with the skewness and kurtosis being within an acceptable range, a parametric test was carried out to compair the mean TD of segment B and C.

A.1.4 Performance

Table A.8 illustrates the outcomes of the assumptions testing for the parametric T-Test's for all three segments.

The test of normality revealed two outcomes, a numerically analysed outcome using the Shapiro-Wilk's method for a sample size $N = 24$, and a graphically analysed outcome using Q-Q plots. The assumptions for the test of normality states:

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

The significance value of $p = 0.388$, $p = 0.070$, and $p = 0.177$ for segments A, B and C respectively, one can conclude that with a CI of 95%, H_0 was considered to be accepted.

Table A.8: Summary of the parametric T-Test's assumptions testing for the participants' performance in the experiment

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segments A & B**	**Segments B & C**
Normal Distribution	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.388)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.070)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.177)$	-	-
Standard Deviation	$\sigma = 3.90$	$\sigma = 3.57$	$\sigma = 2.70$	-	-
Sample Independence	All samples collected independently as noted by the author/experimenter			-	-
Skewness	0.043 ($SE = 0.472$): Acceptable skew	-0.624 ($SE = 0.472$): Acceptable skew	-0.187 ($SE = 0.472$): Left skew	-	-
Kurtosis	-0.298 ($SE = 0.918$): Acceptable kurtosis	-0.380 ($SE = 0.918$): Acceptable kurtosis	-0.680 ($SE = 0.918$): Acceptable kurtosis	-	-
Sample Mean ($\bar{x}$)	9.33 (46.65%)	10.58 (52.90%)	10.54 (52.70%)	-	-
Variances in the means	-	-	-	Accept $H_0(p = 0.298)$	Accept $H_0(p = 0.249)$

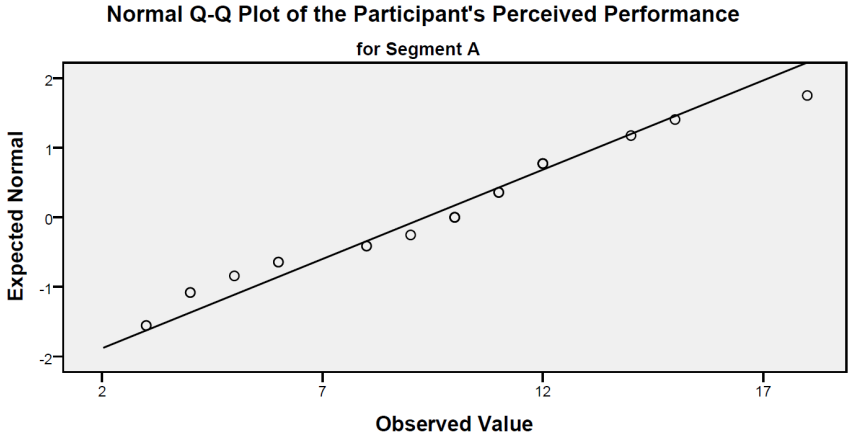


Figure A.19: Q-Q plot for the participants' performance in Segment A

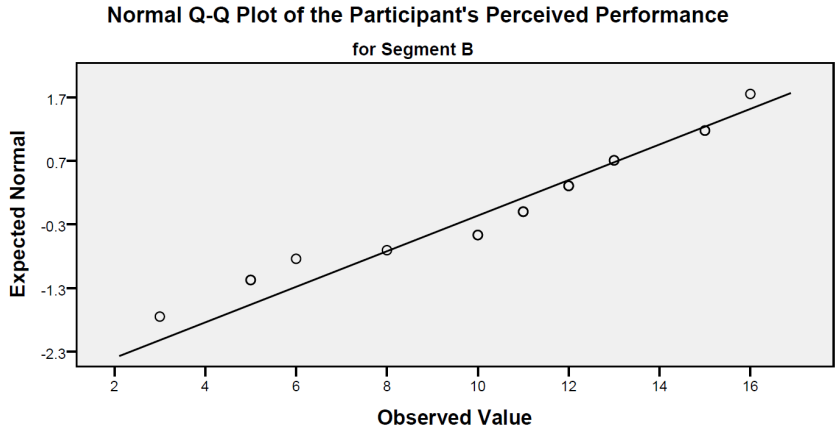


Figure A.20: Q-Q plot for the participants' performance in Segment B

Furthermore, figure A.19, A.20 and A.21 illustrated the Q-Q plots for segments A, B and C respectively, and as the sample points (circles) closely followed the trend line $y = x$, which also supports the conclusion that H_0 was accepted.

The second assumption was to assume the dataset was independently sampled, which

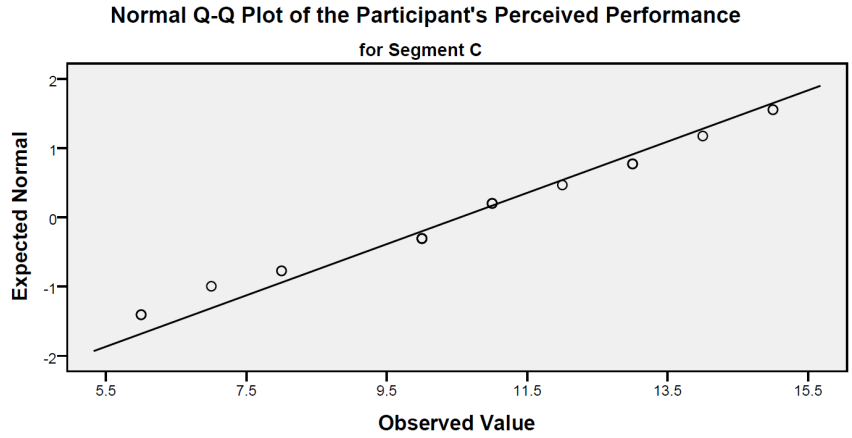


Figure A.21: Q-Q plot for the participants' performance in Segment C

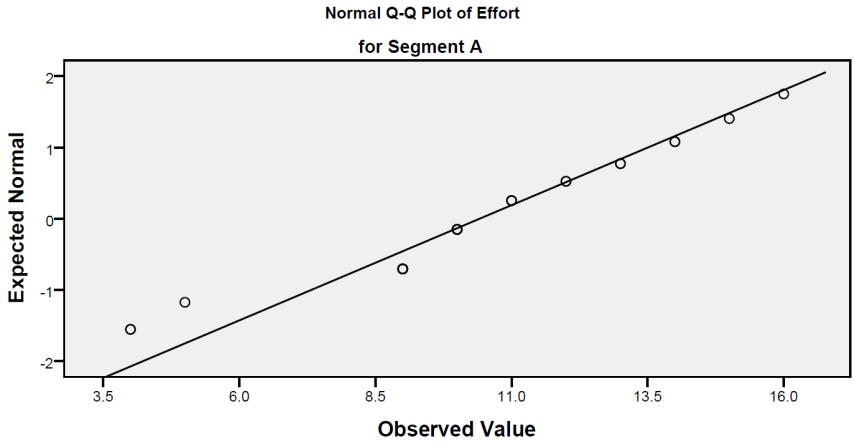


Figure A.22: Q-Q plot for the *effort* asserted in Segment A

was satisfied given the procedure of the experiment and data collection.

The third assumption was to assume the data had similar variances between the means. This was analysed numerically using Levene's test of variances. TableA.8 illustrated the significance value of the variances in the means between segment pairs A and B, and B and C. The hypothesis for the test of variance in the means:

- H_0: The pair of data had similar variances in the means
- H_a: The pair of data did not have similar variances in the means

Given the assumptions were satisfied through assumptions testing, with the skewness and kurtosis being within an acceptable range, a parametric test is be carried out.

A.1.5 Effort

Table A.9 illustrates the outcomes of the assumptions testing for the parametric T-Test's for all three segments.

The test of normality revealed two outcomes, a numerically analysed outcome using the Shapiro-Wilk's method for a sample size $N = 24$, and a graphically analysed outcome using Q-Q plots. The assumptions for the test of normality states:

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

The significance value of $p = 0.131$, $p = 0.958$, and $p = 0.788$ for segments A, B and C respectively, one can conclude that with a CI of 95%, H_0 was considered to be accepted. Furthermore, figure A.22, A.23 and A.24 illustrated the Q-Q plots for segments A, B and C respectively, and as the sample points (circles) closely followed the trend line $y = x$, which also supports the conclusion that H_0 was accepted.

The second assumption was to assume the dataset was independently sampled, which was satisfied given the procedure of the experiment and data collection.

Table A.9: Summary of the parametric T-Test's assumptions testing for the *effort* asserted by the participant throughout the experiment

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segments A & B**	**Segments B & C**
Normal Distribution	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.131)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.958)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.788)$	-	-
Standard Deviation	$\sigma = 3.09$	$\sigma = 3.39$	$\sigma = 3.47$	-	-
Skewness	-0.476 ($SE = 0.472$): Left skewed	0.109 ($SE = 0.472$): Acceptable skew	-0.114 ($SE = 0.472$): Left skew	-	-
Kurtosis	0.345 ($SE = 0.918$): Acceptable kurtosis	-0.593 ($SE = 0.918$): Acceptable kurtosis	-0.520 ($SE = 0.918$): Acceptable kurtosis	-	-
Sample Mean ($\bar{x}$)	10.42 (52.10%)	11.13 (55.65%)	12.25 (61.25%)	-	-
Variances in the means	-	-	-	Accept $H_0(p = 0.453)$	Accept $H_0(p = 0.838)$

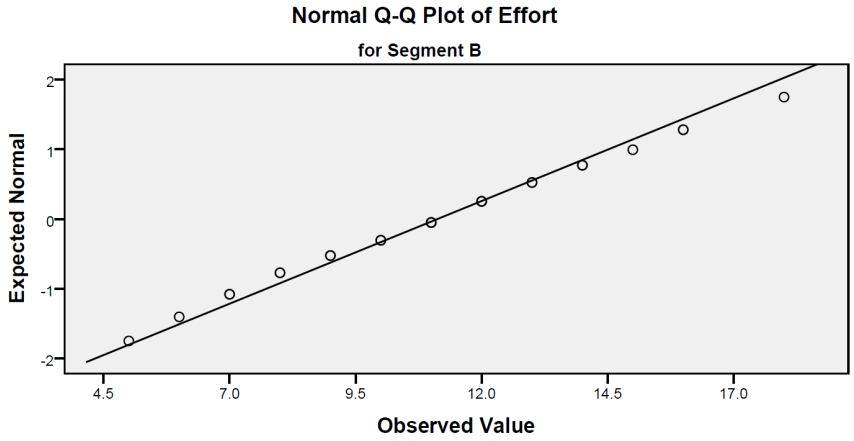


Figure A.23: Q-Q plot for the *effort* asserted in Segment B

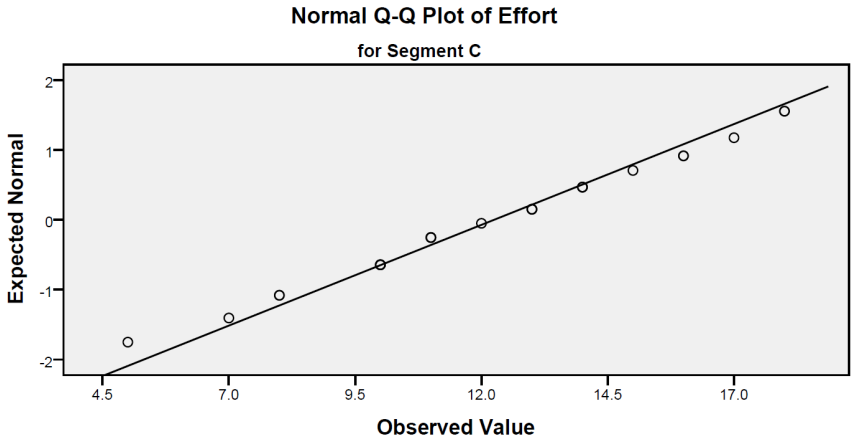


Figure A.24: Q-Q plot for the *effort* asserted in Segment C

The third assumption was to assume the data had similar variances between the means. This was analysed numerically using Levene's test of variances. TableA.9 illustrated the significance value of the variances in the means between segment pairs A and B, and B and C. The hypothesis for the test of variance in the means:

- H_0: The pair of data had similar variances in the means
- H_a: The pair of data did not have similar variances in the means

Given the assumptions were satisfied through assumptions testing, with the skewness and kurtosis being within an acceptable range, a parametric test could be carried out.

A.1.6 Frustration

Table A.10 illustrates the outcomes of the assumptions testing for the parametric T-Test's for all three segments.

The test of normality revealed two outcomes, a numerically analysed outcome using the Shapiro-Wilk's method for a sample size $N = 24$, and a graphically analysed outcome using Q-Q plots. The assumptions for the test of normality states:

Table A.10: Summary of the parametric T-Test's assumptions testing for the level of frustration experienced by the participants during the experiment

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segments A & B**	**Segments B & C**
Normal Distribution	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.316)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.095)$	Graphical: Accept H_0 Numerical: Accept $H_0(p = 0.126)$	-	-
Standard Deviation	$\sigma = 4.69$	$\sigma = 4.71$	$\sigma = 5.33$	-	-
Skewness	-0.022 ($SE = 0.472$): Acceptable skew	-0.375 ($SE = 0.472$): Acceptable skew	-0.363 ($SE = 0.472$): Acceptable skew	-	-
Kurtosis	-1.032 ($SE = 0.918$): Flatter curve	-1.026 ($SE = 0.918$): Flatter curve	-1.018 ($SE = 0.918$): Flatter curve	-	-
Sample Mean ($\bar{x}$)	8.5 (42.5%)	8.67 (43.4%)	9.92 (49.6%)	-	-
Variances in the means	-	-	-	Accept $H_0(p = 1.000)$	Accept $H_0(p = 0.351)$

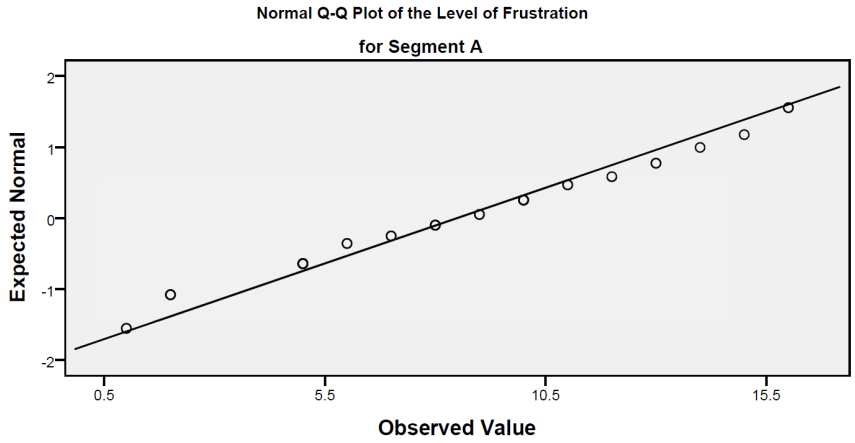


Figure A.25: Q-Q plot for the level of frustration experienced by the participants in Segment A

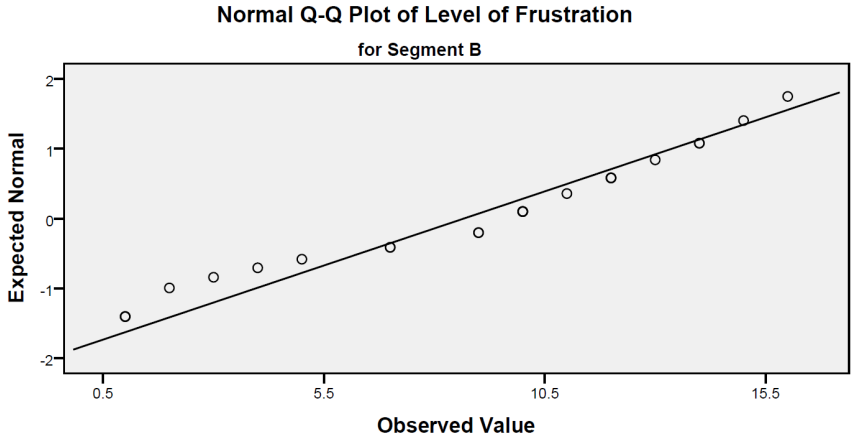


Figure A.26: Q-Q plot for the level of frustration experienced by the participants in Segment B

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

The significance value of $p = 0.316$, $p = 0.095$, and $p = 0.126$ for segments A, B and C respectively, one can conclude that with a CI of 95%, H_0 was considered to be accepted. Furthermore, figure A.25, A.26 and A.27 illustrated the Q-Q plots for segments A, B and C respectively, and as the sample points (circles) closely followed the trend line $y = x$, which also supported the conclusion that H_0 was accepted.

The second assumption was to assume the dataset was independently sampled, which was satisfied given the procedure of the experiment and data collection.

The third assumption was to assume the data had similar variances between the means. This was analysed numerically using Levene's test of variances. TableA.10 illustrated the significance value of the variances in the means between segment pairs A and B, and B and C. The hypothesis for the test of variance in the means:

- H_0: The pair of data had similar variances in the means

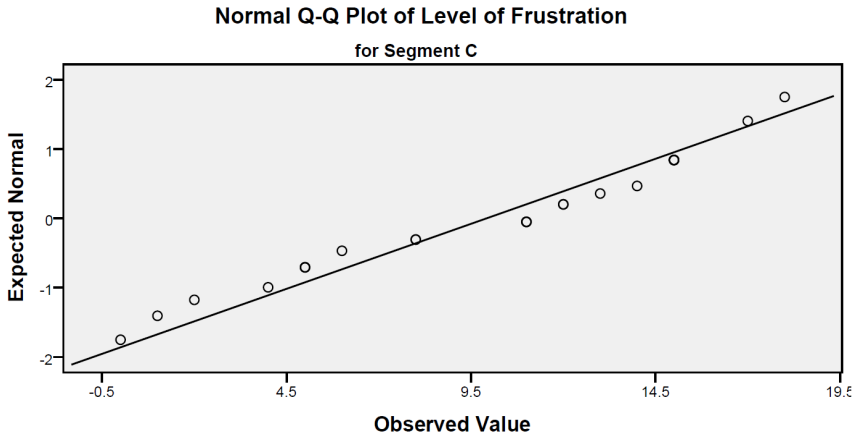


Figure A.27: Q-Q plot for the level of frustration experienced by the participants in Segment C

- H_a: The pair of data did not have similar variances in the means

Given the assumptions were satisfied through assumptions testing, with the skewness and kurtosis being within an acceptable range, a parametric test could be carried out.

A.1.7 Combined Cognitive Workload

Table A.11 illustrates the outcomes of the assumptions testing for the parametric T-Test's for all three segments.

The test of normality revealed two outcomes, a numerically analysed outcome using the Shapiro-Wilk's method for a sample size $N = 24$, and a graphically analysed outcome using Q-Q plots. The assumptions for the test of normality states:

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

The significance value of $p = 0.021$, $p < 0.001$, and $p = 0.768$ for Segments A, B and C respectively, one can conclude that with a CI of 95%, H_0 is considered to be rejected for Segments A and B, while it is considered to be accepted for Segment C. That is, the combined CW data collected for Segment A and B are normally distributed, while Segment C's data is not.

The second assumption is to assume the dataset was independently sampled, which was satisfied given the procedure of the experiment and data collection.

The third assumption is to assume the data had similar variances between the means. This was analysed numerically using Levene's test of variances. TableA.11 illustrated the significance value of the variances in the means between segment pairs A and B, and B and C. The hypothesis for the test of variance in the means:

- H_0: The pair of data had similar variances in the means
- H_a: The pair of data did not have similar variances in the means

Table A.11: Summary of the parametric T-Test's assumptions testing for the combined CW

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segments A & B**	**Segments B & C**
Normal Distribution	Numerical: Reject $H_0(p = 0.021)$	Numerical: Reject $H_0(p < 0.001)$	Numerical: Accept $H_0(p = 0.768)$	-	-
Standard Deviation	$\sigma = 10.25$	$\sigma = 14.71$	$\sigma = 8.46$	-	-
Skewness	-1.33 ($SE = 0.472$): Left skewed	-2.289 ($SE = 0.472$): acceptable skew	-0.424 ($SE = 0.472$): Left skew	-	-
Kurtosis	2.229 ($SE = 0.918$): Acceptable kurtosis	7.653($SE = 0.918$): Acceptable kurtosis	-0.225 ($SE = 0.918$): Acceptable kurtosis	-	-
Sample Mean ($\bar{x}$)	58.50%	59.42%	66.63%	-	-
Variances in the means	-	-	-	Accept $H_0(p = 0.578)$	Accept $H_0(p = 0.732)$

The testing of the combined CW results are presented in Chapter 5 Section 5.2.7.

A.2 Situation Awareness Assumptions Testing

This section detailed the process of assumptions testing of the Situation Awareness (SA) results collected in experiment 1 to support the statistical analyses presented in Chapter 5.

The assumption required to perform a parametric T-Test is threefolds: 1) The results must be normally distributed, 2) The results must be independently sampled, and 3) The results must be have similar variances between the means.

Table A.12 illustrates the outcomes of the assumptions testing for the parametric T-Test's for all three segments.

The test of normality revealed two outcomes, a numerically analysed outcome using the Shapiro-Wilk's method for a sample size (N) of 24, and a graphically analysed outcome using Q-Q plots. The assumptions for the test of normality states:

- H_0: The sample dataset is normally distributed
- H_a: The sample dataset is not normally distributed

The significance value of the (normal) distributions for segments A, B and C are $p = 0.445$, $p = 0.029$, and $p = 0.016$ respectively, one can conclude that with a CI of 95%, H_0 was considered to be accepted for segment A, and rejected for segments B and C.

The second assumption was to assume the dataset was independently sampled, which was satisfied given the procedure of the experiment and data collection.

The third assumption was to assume the data had similar variances between the means. This was analysed numerically using Levene's test of variances. TableA.12 illustrated the significance (p) value of the variances in the means between segment pairs A and B, and B and C. The hypothesis for the test of variance in the means:

- H_0: The pair of data had similar variances in the means
- H_a: The pair of data did not have similar variances in the means

The assumptions testing outcome revealed the data collected from segment A was not considered to be in normal distribution, and transformation is not necessary (due to the insignificant skewness of the data). However, segments B and C was normally distributed. Since the assumptions for the non-parametric test is the same as the parametric test without the normality requirement, segment A (B and C) satisfies the assumptions to carry out the Wilcoxon Signed Rank Test. Therefore, the parametric paired sample T-Test is used to compare segment B with C, while the non-parametric Wilcoxon Signed-Rank Test is used to compare segment A and B's SA.

The testing of the SA results are presented in Chapter 5 Section 5.3.

Table A.12: Summary of the parametric T-Test's assumptions testing for the combined SA, verdict compared at 95% CI

($CI = 95\%$, $\alpha = 0.05$)	**Segment A**	**Segment B**	**Segment C**	**Segments A & B**	**Segments B & C**
Normal Distribution	Accept $H_0(p = 0.445)$	Reject $H_0(p = 0.029)$	Reject $H_0(p = 0.016)$	-	-
Standard Deviation	$\sigma = 0.21$	$\sigma = 0.24$	$\sigma = 0.18$	-	-
Skewness	-0.444 ($SE = 0.472$): Acceptable skew	0.359 ($SE = 0.472$): Acceptable skew	-0.426 ($SE = 0.472$): Acceptable skew	-	-
Kurtosis	-0.159 ($SE = 0.918$): Acceptable kurtosis	-0.839 ($SE = 0.918$): Acceptable kurtosis	-1.157 ($SE = 0.918$): Acceptable kurtosis	-	-
Sample Mean ($\bar{x}$)	0.58 (57.99%)	0.57 (57.12%)	0.53 (52.92%)	-	-
Variances in the means	-	-	-	Accept $H_0(p = 0.895)$	Accept $H_0(p = 0.503)$

APPENDIX B

Experiment 2 Result Analysis

B.1 Cognitive Workload Assumptions Testing

This section detailed the process of assumptions testing of the Cognitive Workload (CW) results collected in experiment 2 to support the statistical analyses presented in Chapter 6.

B.1.1 Mental Demand

Table B.1 presented the descriptives of the raw MD data. The baseline scenario's data were considered not to be normally distributed according to Shapiro-Wilk's test of normality with a significance value of $p = 0.036$ and a negative skew was present. The graphical verification (figure B.1) also confirmed the numeric testing where the data points were non-constantly deviating from line of normality. Therefore with respect to the hypothesis:

- H_0: The data set were in normal distribution
- H_a: The data set were not in normal distribution

The graphical and numerical tests revealed that there were significant statistical evidence to reject H_0, and given there was a negative skew in the data, a square transform was both the baseline and evaluation data sets.

Table B.2 presented the descriptives of the transformed MD data. The transformed data revealed an acceptable skewness and kurtosis of both scenarios with the numerical

Table B.1: Descriptive statistics of the original mental demand data

($CI = 95\%, \alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p = 0.036)$	Numerical: Accept $H_0(p = 0.102)$
Standard Deviation	$\sigma = 3.496$	$\sigma = 3.96$
Skewness	-0.585 ($SE = 0.365$): Negative skew	-0.558($SE = 0.365$): Negative skew
Kurtosis	0.852 ($SE = 0.717$): Acceptable kurtosis	-0.099 ($SE = 0.717$): Acceptable kurtosis
Sample Mean ($\bar{x}$)	$\bar{x}_{baseline} = 13.31$	$\bar{x}_{evaluation} = 12.69$
Sample Size (N)	43	43

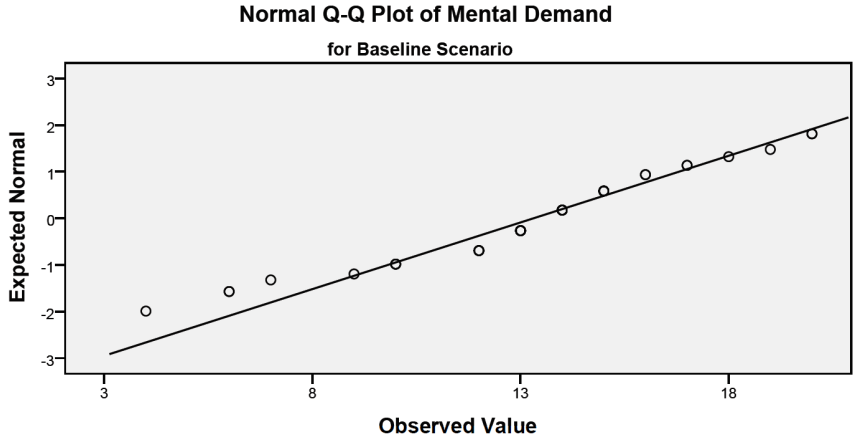


Figure B.1: Q-Q plot of the original mental demand data for the baseline scenario

Table B.2: Descriptive statistics of the (square) transformed mental demand data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.088)$	Numerical: Accept $H_0(p = 0.454)$
Standard Deviation	$\sigma = 88.16$	$\sigma = 94.09$
Skewness	0.385 $(SE = 0.365)$: Positive skew	0.186 $(SE = 0.365)$: Positive skew
Kurtosis	0.571 $(SE = 0.717)$: Acceptable kurtosis	-0.433 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 189.07$	$\bar{x}_{evaluation} = 176.36$
Sample Size (N)	43	43
Sample Variance	Accept H_0: $p = 0.379$	

test of normality confirming the normally distributed nature of the data set. This was also supported by the placement of the sample points in the Q-Q plot illustrated in figure B.2 and B.3, as they were closely following the line of normality.

The final two rows in table B.2 presented the outcomes of the assumptions requirements for a parametric paired-sample test.

The four assumptions that were required to be satisfied are: 1) The data set must be normally distributed, 2) The depedent variables of the data set must be continuous, 3) The independent variables of the data set must be categorically collected, and 4) The dataset must have similar variances. These assumptions were verified following the normality hypothesis:

The hypothesis for sample independence assumption stated (tested using *Pearson's Chi-Square* test):

- H_0: The data set were considered statistically independent
- H_a: The data set were considered not to be statistically independent

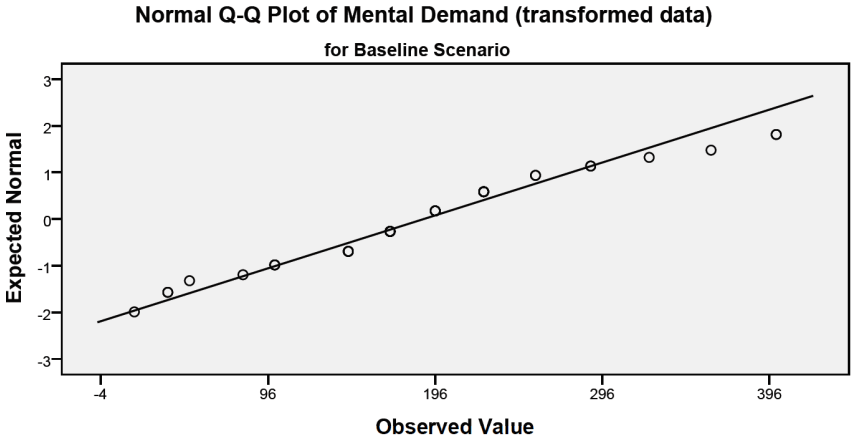


Figure B.2: Q-Q plot of the transformed mental demand rating of the baseline scenario

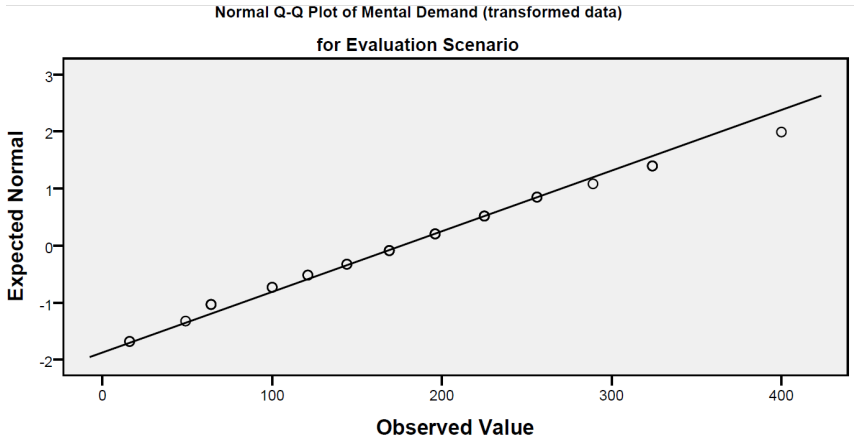


Figure B.3: Q-Q plot of the transformed mental demand rating of the evaluation scenario

And the hypothesis for similar variances ($\alpha_{variance} = 0.05$) stated:

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

From the stated hypothesis and the information presented in table B.2, the transformed data set with the outliers removed had satisfied the assumptions requirement to carry out a parametric paired-sample T-Test.

B.1.2 Physical Demand

An analysis of the raw dataset from both scenarios revealed a positive skew (table B.3), which was corrected using a square-root transformation. The transformed data presented a set of analysis which satisfied the requirements of the assumptions testing, hence, a T-Test was used.

Table B.3 presented the descriptives of the raw PD data. The dataset were not considered to be normally distributed according to Shapiro-Wilk's test of normality with a

Table B.3: Descriptive statistics of the raw combined cognitive workload data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p < 0.001)$	Numerical: Reject $H_0(p = 0.001)$
Standard Deviation	$\sigma = 3.995$	$\sigma = 4.132$
Skewness	1.016 $(SE = 0.365)$: Positive skew	0.716 $(SE = 0.365)$: Positive skew
Kurtosis	-0.055 $(SE = 0.717)$: Acceptable kurtosis	-0.669 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 4.548$	$\bar{x}_{evaluation} = 4.738$
Sample Size (N)	43	43

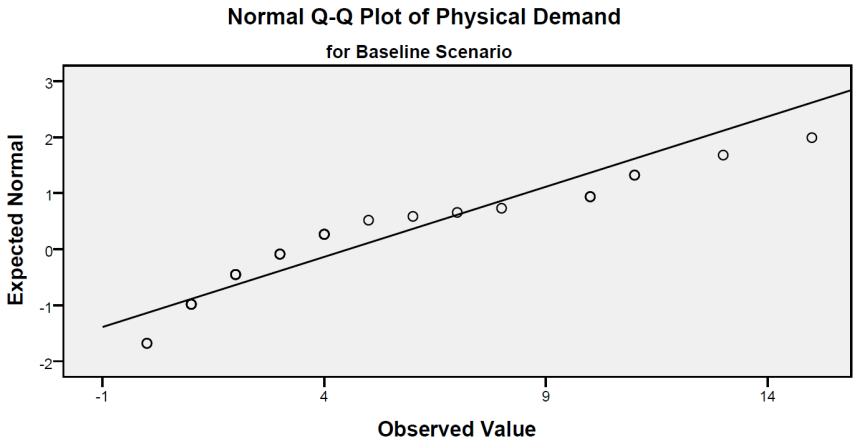


Figure B.4: Q-Q plot of the original physical demand data for the baseline scenario

positive skew, furthermore, the graphical verification also supported the skewness where the Q-Q plot in figure B.4 and B.5 of both scenarios' data points strayed around the line of normality. A square-root transformation was applied to the data to rectify the distribution and the skewness.

The transformed was analysed and the descriptive statistics were illustrated in table B.4, along with the values of testing of the assumptions for a parametric test.

The final two rows in table B.4 presented the outcomes of the assumptions requirements for a parametric paired-sample test.

The four assumptions that were required to be satisfied are: 1) The data set must be normally distributed, 2) The depedent variables of the data set must be continuous, 3) The independent variables of the data set must be categorically collected, and 4) The dataset must have similar variances. These assumptions were verified following the normality hypothesis:

And the hypothesis for similar variances stated:

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

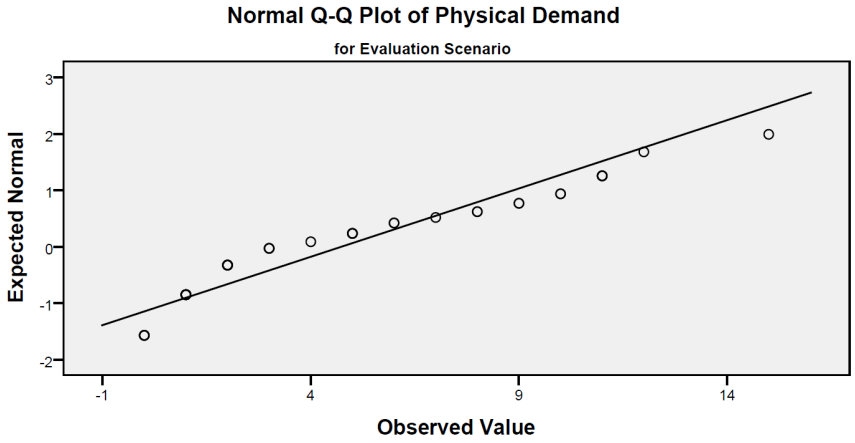


Figure B.5: Q-Q plot of the original physical demand data for the evaluation scenario

Table B.4: Descriptive statistics of the (square-root) transformed physical demand data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.071)$	Numerical: Accept $H_0(p = 0.054)$
Standard Deviation	$\sigma = 0.98$	$\sigma = 1.06$
Skewness	0.113 $(SE = 0.365)$: Acceptable skew	-0.070 $(SE = 0.365)$: Acceptable skew
Kurtosis	-0.446 $(SE = 0.709)$: Acceptable kurtosis	-0.871 $(SE = 0.709)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 1.9$	$\bar{x}_{evaluation} = 1.91$
Sample Size (N)	43	43
Sample Variance	Accept $H_0(p = 0.308)$	

From the stated hypothesis and the information presented in table B.4, the transformed data set with the outliers removed had satisfied the assumptions requirement to carry out a parametric paired-sample T-Test.

B.1.3 Temporal Demand

The raw data set was tested to determine whether it was suitable to analyse the differences in the mean magnitude of load using the parametric test. Table B.5 illustrated that the raw data set did not have a normal distribution when analysed numerically, and it was left skewed. Furthermore, figures B.6 and B.7 illustrated the graphical evidence also supported the non-normal distribution of the data as the sample points did not closely follow the line of normality.

A negative skew was evident in the data set, hence a square transform was applied to the data, and the descriptive statistics of the transformed data was presented in table B.6.

Based on the following hypothesis, one did not have significant statistical evidence to accept H_0 of the data collected from the evaluation scenario:

Table B.5: Descriptive statistics of the raw temporal demand data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.053)$ Graphical: Reject H_0	Numerical: Reject $H_0(p = 0.019)$ Graphical: Reject H_0
Standard Deviation	$\sigma = 4.56$	$\sigma = 5.11$
Skewness	-0.589 $(SE = 0.365)$: Negative skew	-0.487 $(SE = 0.365)$: Negative skew
Kurtosis	0.158 $(SE = 0.717)$: Acceptable kurtosis	0.855 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 11.33$	$\bar{x}_{evaluation} = 10.4$
Sample Size (N)	42	42

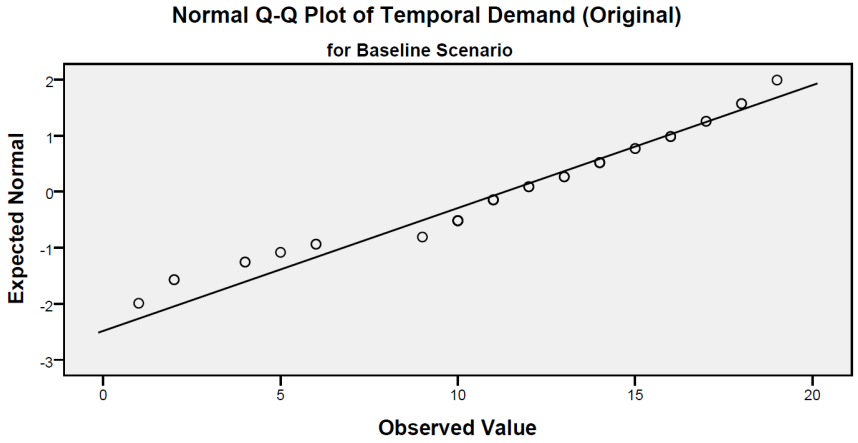


Figure B.6: Q-Q plot of the original temporal demand data for the baseline scenario

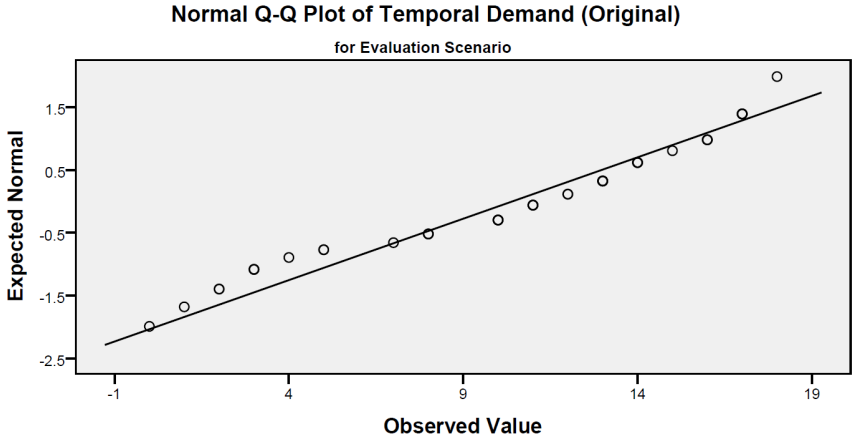


Figure B.7: Q-Q plot of the original temporal demand data for the evaluation scenario

Table B.6: Descriptive statistics and the assumptions testing summary of the square transformed temporal demand data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.156)$ Graphical: Reject H_0	Numerical: Reject $H_0(p = 0.023)$ Graphical: Reject H_0
Standard Deviation	$\sigma = 94.895$	$\sigma = 97.288$
Skewness	0.330 $(SE = 0.365)$: Acceptable skew	0.233 $(SE = 0.365)$: Acceptable skew
Kurtosis	-0.525 $(SE = 0.717)$: Acceptable kurtosis	-1.037 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 148.76$	$\bar{x}_{evaluation} = 133.74$
Sample Size (N)	42	42
Sample Variance	Accept $H_0(p = 0.678)$	

- H_0: The data set were in normal distribution
- H_a: The data set were not in normal distribution

The assumptions for a parametric test included the distribution of the data must be normal, which both the original and the transformed data sets were not able to satisfy, hence a non-parametric testing of the means was used.

The non-parametric, Wilcoxon Signed-Rank Test had three assumptions

1. The dependent variable should be ordinal or continuous, which was met due to the nature of the magnitude of load of the data set
2. The independent variables should consist of two categorical data sets, which was met as the data reflect the magntitude of load of the participants during two separate experimental trials
3. The distribution of the differences between the groups must be symmetrical in shape, which was met as the differences between the raw and transformed data collected in the baseline and evaluation scenarios were symmetrical.

With the assumptions of the Wilcoxon Signed-Rank test met for both the original and the transformed data, two sets of analysis were carried out for the original and the transformed data set.

B.1.4 Performance

The data set was tested to determine whether it was suitable to analyse the differences in the mean magnitude of load using the parametric test. Table B.7 presented the descriptive statistics and the summary of the assumptions tests.

To perform the parametric T-Test, the three assumptions were met as presented in table B.7:

Table B.7: Descriptive statistics of the perceived performance data

($CI = 95\%, \alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.222)$ Graphical: Accept H_0	Numerical: Accept $H_0(p = 0.306)$ Graphical: Accept H_0
Standard Deviation	$\sigma = 3.75$	$\sigma = 3.76$
Skewness	-0.143 ($SE = 0.365$): Negative skew	-0.149 ($SE = 0.365$): Negative skew
Kurtosis	-0.264 ($SE = 0.717$): Acceptable kurtosis	-0.772 ($SE = 0.717$): Acceptable kurtosis
Sample Mean ($\bar{x}$)	$\bar{x}_{baseline} = 9.0$	$\bar{x}_{evaluation} = 8.33$
Sample Size (N)	42	42
Sample Variance	Accept H_0 ($p = 0.674$)	

1. The data set must be normally distributed: Assumption met (H_0 accepted) with CI = 95% - analysed using Shapiro-Wilk's test of normality, and the Q-Q plots (figures B.8 and B.9) had also supported the verdict with the sample points closely tracking the line of normality

2. The data set must have similar variances: Assumption met (H_0 accepted) with CI = 95% - analysed using Levene's Test for Equality for Variances

3. The independent variable must consists of at least two categorical sets: Assumption met as the independent variable (performance rating) was collected in two separate experiment trials

The hypothesis for the tests were defined as follows:

- H_0: The data set were in normal distribution, considered statistically independent, and had similar variances

- H_a: The data set were not in normal distribution, considered statistically dependent, and did not have similar variances

With the assumptions satisfactorially met, the parametric paired-sample T-Test was conducted to evaluate the differences in the means.

B.1.5 Effort

Table B.8 presented the descriptives of the raw effort data. The baseline scenario's data were considered not to be normally distributed according to Shapiro-Wilk's test of normality with a significance value of $p = 0.044$ and a negative skew was present. This was verified graphically in figure B.10 where the data points were fluctuating from line of normality. Therefore with respect to the hypothesis:

- H_0: The data set were in normal distribution

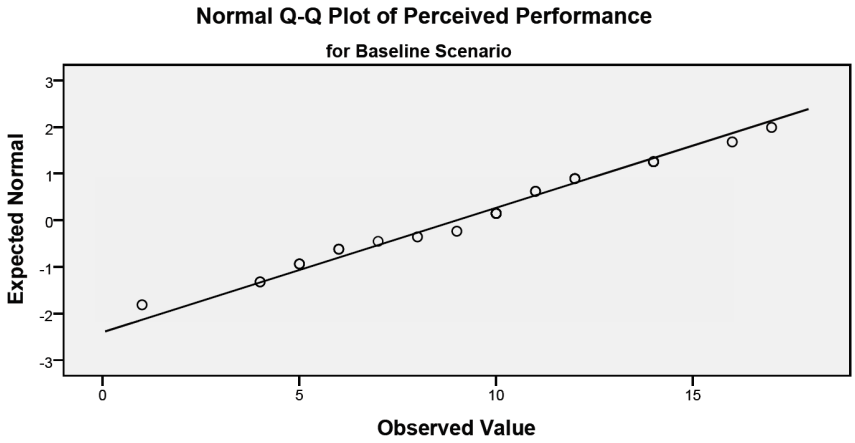


Figure B.8: Q-Q plot of the perceived performance data for the baseline scenario

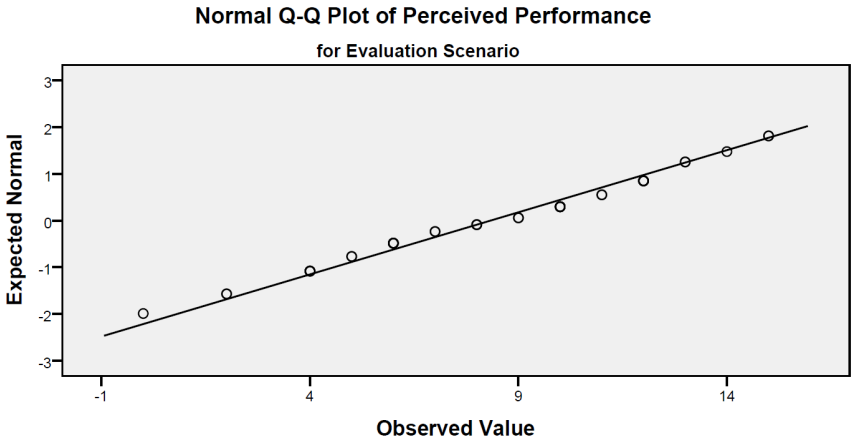


Figure B.9: Q-Q plot of the perceived performance data for the evaluation scenario

Table B.8: Descriptive statistics of the original effort data

($CI = 95\%$, $\alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p = 0.044)$	Numerical: Accept $H_0(p = 0.066)$
Standard Deviation	$\sigma = 4.042$	$\sigma = 4.431$
Skewness	-0.625 $(SE = 0.365)$: Negative skew	-0.469$(SE = 0.365)$: Negative skew
Kurtosis	-0.098 $(SE = 0.717)$: Acceptable kurtosis	-0.583 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 13.31$	$\bar{x}_{evaluation} = 12.69$
Sample Size (N)	42	42

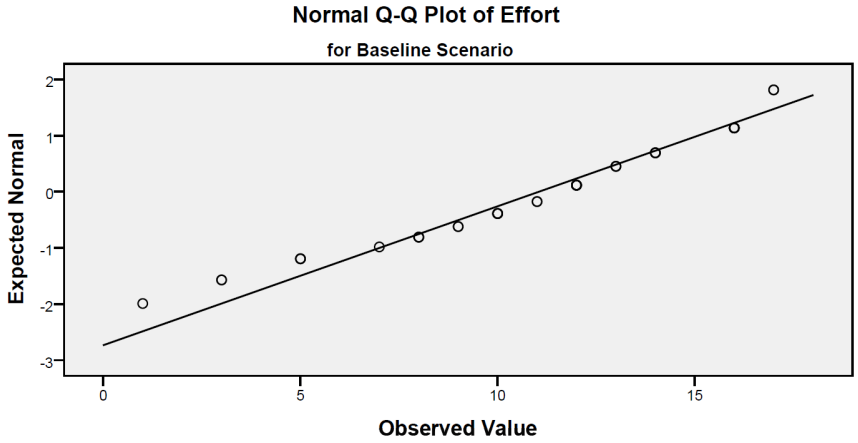


Figure B.10: Q-Q plot of the original effort data for the baseline scenario

Table B.9: Descriptive statistics of the (square) transformed effort data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.084)$	Numerical: Accept $H_0(p = 0.052)$
Standard Deviation	$\sigma = 81.86$	$\sigma = 76.58$
Skewness	0.180 $(SE = 0.365)$: Acceptable skew	0.447 $(SE = 0.365)$: Positive skew
Kurtosis	-0.824 $(SE = 0.717)$: Flatter curve	-0.378 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 138$	$\bar{x}_{evaluation} = 105.83$
Sample Size (N)	42	42
Sample Variance	Accept H_0: $p = 0.612$	

- H_a: The data set were not in normal distribution

The graphical and numerical tests revealed that there were significant statistical evidence to reject H_0, and given there was a negative skew in the data, a square transform was both the baseline and evaluation data sets.

Table B.9 presented the descriptives of the transformed effort data. The transformed data revealed an acceptable skewness and kurtosis of both scenarios with the numerical test of normality confirming the normally distributed nature of the data set.

The final two rows in table B.9 presented the outcomes of the assumptions for a parametric paired-sample test.

The three assumptions that were required to be satisfied are: 1) The data set must be normally distributed, 2) The depedent variables of the data set must be continuous, 3) The independent variables of the data set must be categorically collected, and 4) The dataset must have similar variances. These assumptions were verified following the normality hypothesis:

And the hypothesis for similar variances stated:

- H_0: The data set had similar variances in the means

Table B.10: Descriptive statistics of the frustration level data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p = 0.043)$ Graphical: Reject H_0	Numerical: Reject $H_0(p = 0.002)$ Graphical: Reject H_0
Standard Deviation	$\sigma = 4.75$	$\sigma = 5.01$
Skewness	-0.020 $(SE = 0.365)$: Acceptable skew	0.412 $(SE = 0.365)$: Positive skew
Kurtosis	-1.177 $(SE = 0.717)$: Flatter curve	-1.221 $(SE = 0.717)$: Flatter curve
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 7.83$	$\bar{x}_{evaluation} = 7.1$
Sample Size (N)	42	42
Equality of Variances	Accept H_0 $(p = 0.734)$	

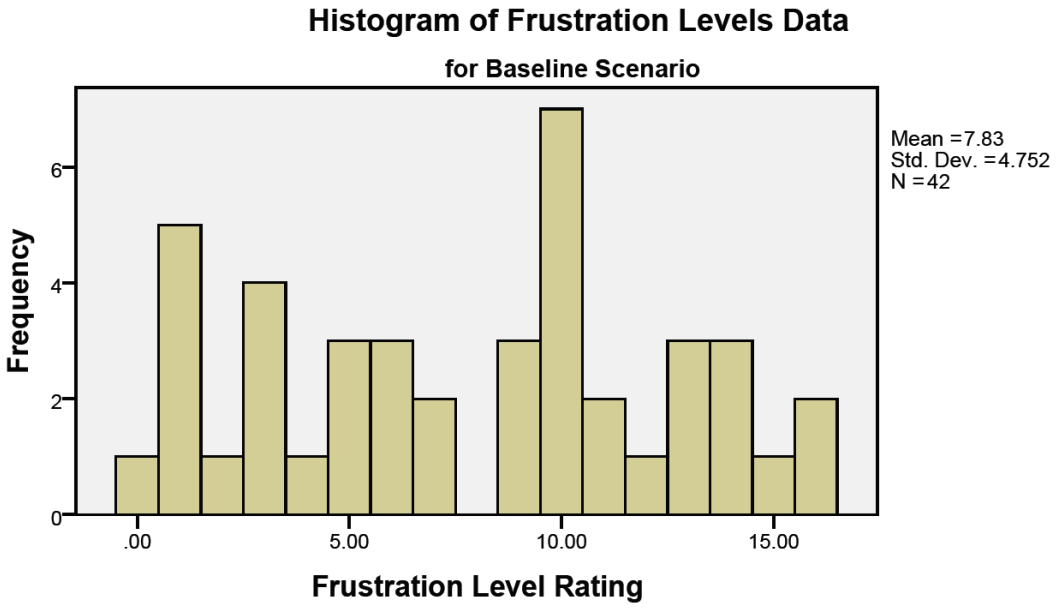


Figure B.11: Histogram of the frustration level data for the baseline scenario

- H_a: The data set did not have similar variances in the means

From the stated hypothesis and the information presented in the table, all the assumptions were satisfactorily met, hence, the parametric test could be carried out.

B.1.6 Frustration

The data set was tested to determine whether it was suitable to analyse the differences in the mean magnitude of load using the parametric test. Table B.10 illustrated the descriptive statistics related to the data set.

Based on the following hypothesis, one did not have significant statistical evidence to accept H_0 both numerically and graphically. That was to say, the data set did not have a normal distribution (rejecting H_0) when analysed numerically, and inspected graphically through the histograms in figures B.11 and B.12, however, the skewness was within

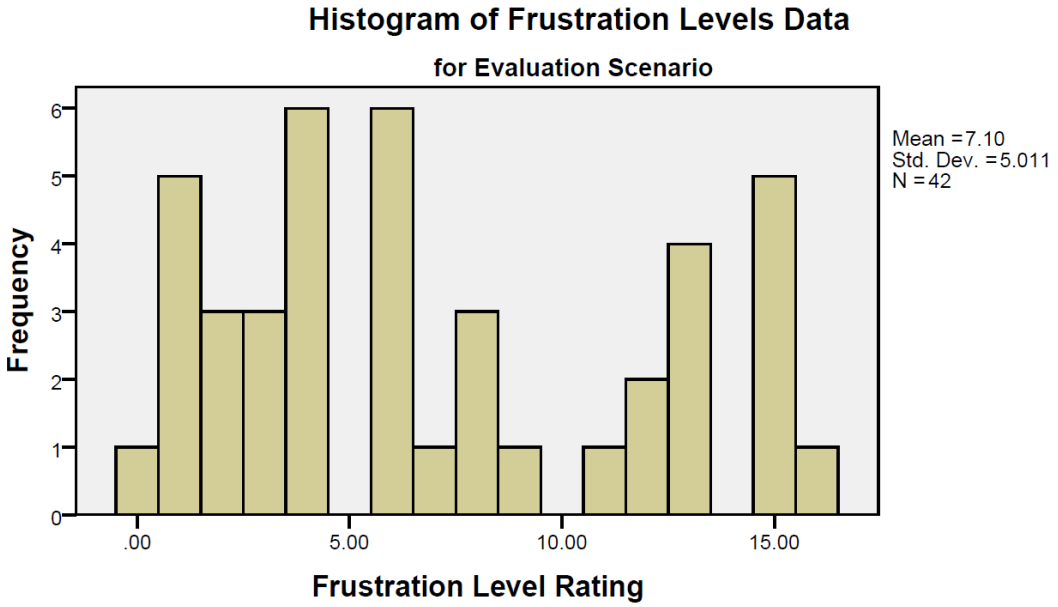


Figure B.12: Histogram of the frustration level data for the evaluation scenario

acceptable range.

- H_0: The data set were in normal distribution
- H_a: The data set were not in normal distribution

The hypothesis for sample independence assumption stated (tested using *Pearson's Chi-Square* test):

- H_0: The data set were considered statistically independent
- H_a: The data set were considered not to be statistically independent

Given the inability for the assumptions of a parametric test to be met, testing of the assumptions of the non-parametric testing of the means was carried out.

The non-parametric, Wilcoxon Signed-Rank Test had three assumptions

1. The dependent variable should be ordinal or continuous, which was met due to the continuous scaling of the magnitude of load of the data set
2. The independent variables should consist of two categorical data sets, which was met as the data reflect the magntitude of load of the participants during two separate experimental trials
3. The distribution of the differences between the groups must be symmetrical in shape, which was also met as the differences between the baseline and evaluation scenarios were symmetrical, illustrated by the boxplot in figure B.13

With the assumptions of the Wilcoxon Signed-Rank test met for the data set, analysis was carried out.

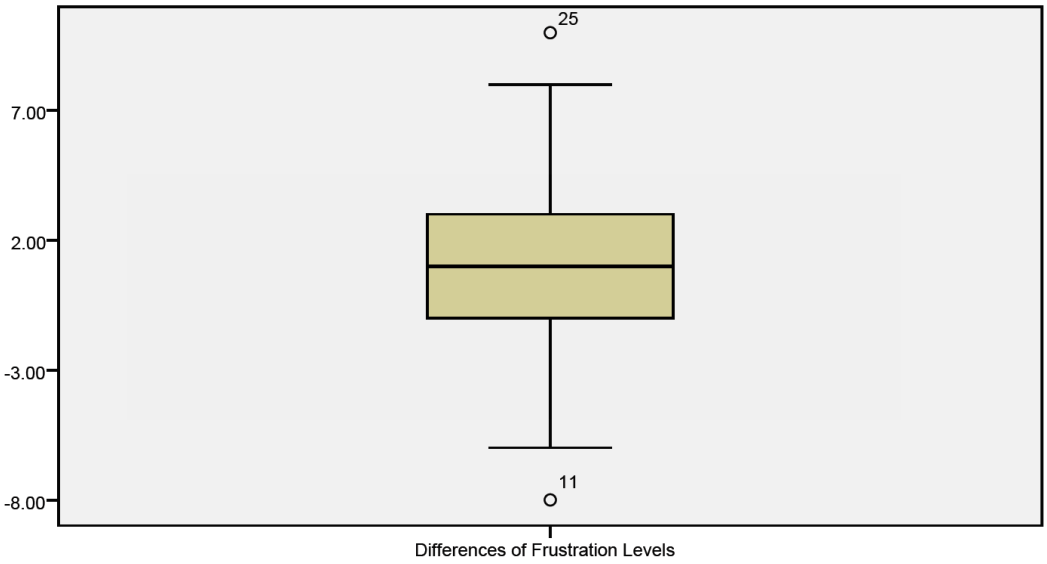


Figure B.13: Boxplot of the differences in the distribution between the frustration level felt in the baseline and the evaluation scenario

Table B.11: Descriptive statistics of the raw combined cognitive workload data

($CI = 95\%, \alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.076)$	Numerical: Accept $H_0(p = 0.260)$
Standard Deviation	$\sigma = 13.91$	$\sigma = 15.02$
Skewness	-0.529 ($SE = 0.361$): Negative skew	-0.557 ($SE = 0.361$): Negative skew
Kurtosis	0.481 ($SE = 0.709$): Acceptable kurtosis	0.893 ($SE = 0.709$): Acceptable kurtosis
Sample Mean ($\bar{x}$)	$\bar{x}_{baseline} = 55.86$	$\bar{x}_{evaluation} = 51.95$
Sample Size (N)	43	43

B.1.7 Combined Cognitive Workload

Table B.11 presented the descriptives of the raw combined CW data. The dataset were considered normally distributed according to Shapiro-Wilk's test of normality, however, a negative skew was present. The graphical verification also supported the skewness where the Q-Q plot in figure B.11 of the baseline data's data points (circles) strayed around the $y = x$ line of normality. And the Q-Q plot of the evaluation data (figure B.12) suggested that at the end of the distribution, several data points no longer closely followed the line of normality, hence, a transformation was performed.

Table B.12 presented the descriptives of the transformed combined CW data. In this set of data, a set of outliers were detected (as illustrated in figure B.16) to have a significantly higher CW rating for both scenarios. As a result, sample 29 was removed from both sets of data, and table B.13 presented the descriptive statistics of the set of transformed data with the outliers removed.

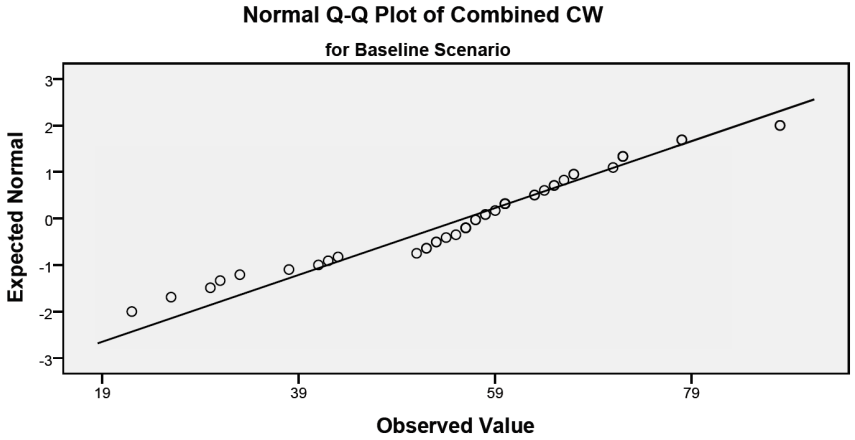


Figure B.14: Q-Q plot of the original combined CW data for the baseline scenario

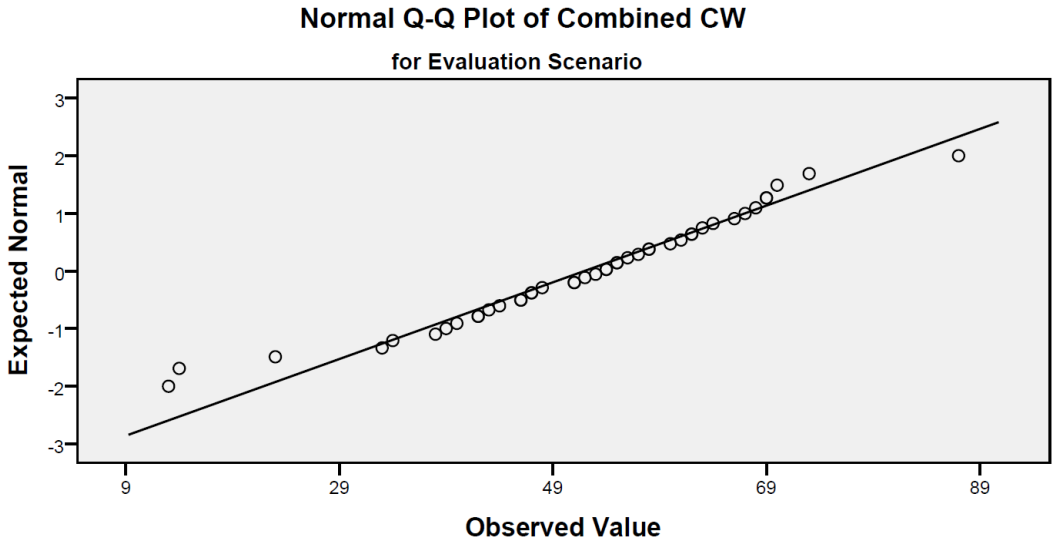


Figure B.15: Q-Q plot of the original combined CW data for the evaluation scenario

Table B.12: Descriptive statistics of the (square) transformed combined cognitive workload data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.224)$	Numerical: Accept $H_0(p = 0.439)$
Standard Deviation	$\sigma = 1481.97$	$\sigma = 1481.56$
Skewness	0.331 $(SE = 0.361)$: Acceptable skew	0.538 $(SE = 0.361)$: Positive skew
Kurtosis	0.928 $(SE = 0.709)$: Acceptable kurtosis	1.039 $(SE = 0.709)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 3309.44$	$\bar{x}_{evaluation} = 2919.63$
Sample Size (N)	43	43

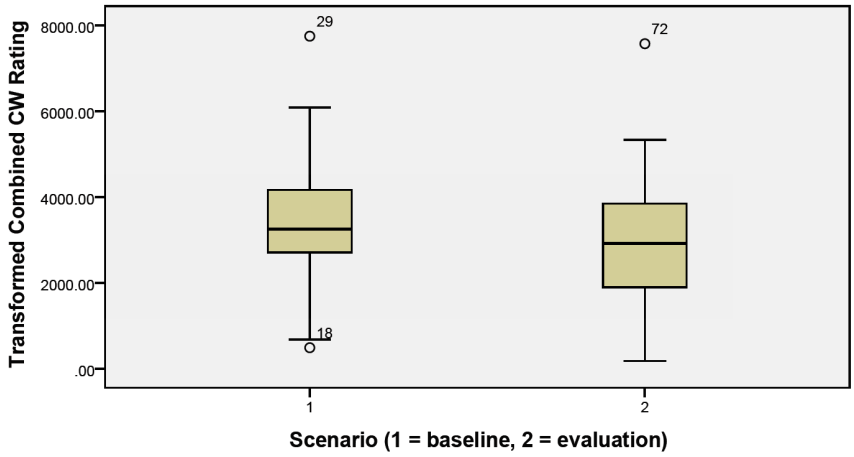


Figure B.16: Boxplot of the transformed combined cognitive workload rating, where Scenario 1 and 2 denoted the baseline and evluation scenarios

Table B.13: Descriptive statistics and the assumptions testing outcome of the transformed combined cognitive workload data with outliers removed

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.325)$	Numerical: Accept $H_0(p = 0.776)$
Standard Deviation	$\sigma = 1326.17$	$\sigma = 1307.19$
Skewness	-0.232 $(SE = 0.365)$: Acceptable skew	-0.082 $(SE = 0.365)$: Acceptable skew
Kurtosis	-0.228 $(SE = 0.717)$: Acceptable kurtosis	-0.597 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 3203.86$	$\bar{x}_{evaluation} = 2808.93$
Sample Size (N)	42	42
Similar Variance	Accept $H_0(p = 0.747)$	

Table B.13 presented the descriptives of the final set of combined CW which were used to compare the means.

The final two rows in table B.13 presented the outcomes of the assumptions requirements for a parametric paired-sample test.

The four assumptions that were required to be satisfied are: 1) The dataset must be normally distributed, 2) The dependent variables of the data set must be continuous, 3) The independent variables of the data set must be categorically collected, and 4) The dataset must have similar variances. These assumptions were verified following the normality hypothesis:

- H_0: The data set were in normal distribution
- H_a: The data set were not in normal distribution

And the hypothesis for similar variances stated:

Table B.14: Descriptive statistics and the assumptions testing outcome of the level 1 SA data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p < 0.001)$	Numerical: Reject $H_0(p = 0.003)$
Standard Deviation	$\sigma = 0.19$	$\sigma = 0.207$
Skewness	-0.515 $(SE = 0.365)$: Negative skew	-0.012 $(SE = 0.365)$: Acceptable skew
Kurtosis	0.005 $(SE = 0.717)$: Acceptable kurtosis	0.6 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 0.524$	$\bar{x}_{evaluation} = 0.472$
Sample Size (N)	42	42
Similar Variance	Accept $H_0(p = 0.398)$	

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

From the stated hypothesis and the information presented in table B.13, the transformed data set with the outliers removed had satisfied the assumptions requirement to carry out a parametric paired-sample T-Test.

B.2 Situation Awareness Assumptions Testing

This section detailed the process of assumptions testing of the Situation Awareness (SA) results collected in experiment 2 to support the statistical analyses presented in Chapter 6 Section 6.3.

B.2.1 Level 1 SA

Table B.14 presented the descriptives of the level 1 SA data. Assumptions for the test of normality was specified as:

- H_0: The data set were normally distributed
- H_a: The data set were not normally distributed

The data set recorded from both the baseline and the evaluation scenario revealed that the null hypothesis (H_0) was rejected at CI = 95% according to Shapiro-Wilk's test of normality, hence, the data sets were not normally distributed.

One of the four assumptions that must be met to perform a parametric T-Test was the distribution of normality, the remaining three included: The dependent variables of the data set must be continuous; the independent variables of the data set must be categorically collected; and the data set must have similar variances. These assumptions were verified following the normality hypothesis stated above, and the variance hypothesis:

Table B.15: Descriptive statistics and the assumptions testing outcome of the Level 2 SA data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p < 0.001)$	Numerical: Reject $H_0(p = 0.001)$
Standard Deviation	$\sigma = 0.157$	$\sigma = 0.299$
Skewness	0.261 $(SE = 0.365)$: Acceptable skew	0.304 $(SE = 0.365)$: Acceptable skew
Kurtosis	-1.303 $(SE = 0.717)$: Flatter curve	-0.714 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 0.171$	$\bar{x}_{evaluation} = 0.518$
Sample Size (N)	42	42
Similar Variance	Reject $H_0(p = 0.003)$	

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

Since the latter assumptions were met and the normality assumption could not be met, an equivalent non-parametric test was used.

B.2.2 Level 2 SA

The four assumptions that must be met are: 1) The dataset must be normally distributed, 2) The dependent variables of the data set must be continuous, 3) The independent variables of the data set must be categorically collected, and 4) The data set must have similar variances. These assumptions were verified following the normality hypothesis stated above, and the variance hypothesis:

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

And the hypothesis for the homogenity of variance were:

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

Table B.15 that presented the descriptives of the level 2 SA data revealed that at CI = 95%, both the first and the last assumption could not accept H_0, that was to say, the data set was not normally distributed and their variances were not similar.

Given the non-parametric testing of the means required an assumption of similar variance, a variance-stabilising transform to the data was required. The two transforms that could be applied to perform variance-stabilisation were the natural-log transform, and the square-root transform. Since the data included sample with zero percent, the natural-log

Table B.16: Descriptive statistics and the assumptions testing outcome of the (square-root) transformed level 2 SA data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p < 0.001)$	Numerical: Reject $H_0(p < 0.001)$
Standard Deviation	$\sigma = 0.265$	$\sigma = 0.257$
Skewness	-0.268 $(SE = 0.365)$: Acceptable skew	-1.001 $(SE = 0.365)$: Negative skew
Kurtosis	-1.692 $(SE = 0.717)$: Flatter curve	1.437 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 0.321$	$\bar{x}_{evaluation} = 0.673$
Sample Size (N)	42	42
Similar Variance	Accept $H_0(p = 0.063)$	

Table B.17: Descriptive statistics of the combined SA data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p = 0.039)$	Numerical: Accept $H_0(p = 0.090)$
Standard Deviation	$\sigma = 0.155$	$\sigma = 0.246$
Skewness	-0.556 $(SE = 0.361)$: Negative skew	-0.084 $(SE = 0.361)$: Acceptable skew
Kurtosis	1.234 $(SE = 0.709)$: Acceptable kurtosis	-0.291 $(SE = 0.709)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 0.28$	$\bar{x}_{evaluation} = 0.51$
Sample Size (N)	43	43

transform could not be applied, hence, the square-root transformation was performed to achieve the outcome presented in table B.16.

Table B.16 presented the descriptive statistics for the transformed SA data for level 2 SA where at CI = 95%, the variances were similar.

B.2.3 Combined SA

Table B.17 presented the descriptives of the raw combined SA data. Assumptions for the test of normality was specified as:

- H_0: The data set were normally distributed
- H_a: The data set were not normally distributed

The data set recorded from the baseline scenario revealed that the null hypothesis (H_0) was rejected at CI = 95% according to Shapiro-Wilk's test of normality, hence, the data set was not normally distributed . The data set recorded from the evaluation scenario revealed that H_0 was accepted at CI = 95%, hence, the data set was normally distributed. Upon further examining the boxplot of the data in figure B.17, it was apparent that sample 27 was an outlier of the baseline data, hence, table B.18 presented the descriptive statistics

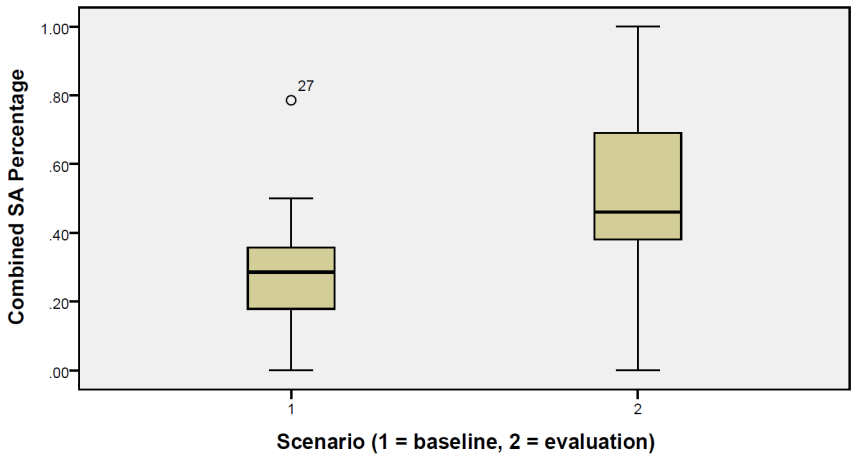


Figure B.17: Boxplot of the combined SA data where Scenario 1 and 2 denoted the baseline and evluation scenarios

Table B.18: Descriptive statistics and the assumptions testing outcome of the combined SA data with outliers removed

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.052)$	Numerical: Accept $H_0(p = 0.119)$
Standard Deviation	$\sigma = 0.135$	$\sigma = 0.241$
Skewness	-0.174 $(SE = 0.365)$: Acceptable skew	0.092 $(SE = 0.365)$: Acceptable skew
Kurtosis	-0.821 $(SE = 0.717)$: Flatter curve	-0.156 $(SE = 0.717)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 0.272$	$\bar{x}_{evaluation} = 0.498$
Sample Size (N)	42	42
Similar Variance	Reject $H_0(p = 0.007)$	

of the combined SA percentage with the outlier removed from both the baseline and the evaluation data sets.

Four assumptions must be met to reliability perform a parametric T-Test. However, the test could still be carried out even if the assumptions could not be fully met, resulting in a less reliable analysis. The four assumptions that were: 1) The dataset must be normally distributed, 2) The dependent variables of the data set must be continuous, 3) The independent variables of the data set must be categorically collected, and 4) The data set must have similar variances. These assumptions were verified following the normality hypothesis stated above, and the variance hypothesis:

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

Assumptions 1, 2 and 3 were satisfactorily met, however, at CI = 95%, H_0 was rejected, which meant that assumption 4 was not met, that is, the data set did not have similar variances in the means. A variance-stabilising transform by applying a square-root to

the data set could be applied to rectify the non-homogeneous variance, adjusting the significance value to $p = 0.459$, however, consequently, the significance value of test of normality became $p_{baseline} = 0.020$ and $p_{evaluation} < 0.001$, hence, both the baseline and the evaluation scenarios rejected H_0, hence, denoting that at CI = 95%, both data sets were not normally distributed after a square-root transform was applied.

As such, a parametric paired sample T-Test was applied to test for the differences in the means using the untransformed data, while a non-parametric Wilcoxon Signed-Rank test was applied to test for the differences in the means using the square-root transformed data.

APPENDIX C

Experiment 3 Result Analysis

C.1 Cognitive Workload Assumptions Testing

This section detailed the process of assumptions testing of the Cognitive Workload (CW) results collected in experiment 3 to support the statistical analyses presented in Chapter 7.

For each of the CW attributes to perform a parametric T-Test, four assumptions were required to be satisfied to apply the parametric test: 1) The data set must be normally distributed, 2) The dependent variables of the data set must be continuous, 3) The independent variables of the data set must be categorically collected, and 4) The dataset must have similar variances. Due to the nature of the data collection process, both assumptions 2 and 3 were satisfied. The hypothesis for normality, determined using a single or a combination of the numerical Shapiro-Wilk's and the graphical method, stated:

- H_0: The data set were in normal distribution
- H_a: The data set were not in normal distribution

And the hypothesis for similar variances, determined using the Levene's statistics, stated:

- H_0: The data set had similar variances in the means
- H_a: The data set did not have similar variances in the means

The results were first analysed to determine whether the assumptions could be satisfactorily met, if not, attempts were made to ensure that the assumptions could be met. However, if the data sets were not suitable for parametric tests, the non-parametric tests were used.

The first test (two-tailed test) was to determine the significance of the difference between the scenarios, where:

- H_0: The two sets of data were not statistically different
- H_a: The two sets of data were statistically different

If H_0 was rejected, a one-tailed test was conducted to determine the direction of the differences. The hypothesis stated:

- H_0: The baseline data were significantly greater than the evaluation data
- H_a: The baseline data were significantly less than the evaluation data

Table C.1: Descriptive statistics of the mental demand data at CI=95%

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Reject $H_0(p = 0.022)$	Numerical: Accept $H_0(p = 0.308)$
Standard Deviation	$\sigma = 9.042$	$\sigma = 7.787$
Skewness	-0.079 $(SE = 0.414)$: Acceptable skew	0.346$(SE = 0.414)$: Acceptable skew
Kurtosis	-1.424 $(SE = 0.809)$: Flatter curve	-0.681 $(SE = 0.809)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_A} = 16.65$	$\bar{x}_B = 13.64$
Sample Size (N)	32	32
Similar Variances	Accept H_0 $(p = 0.116)$	

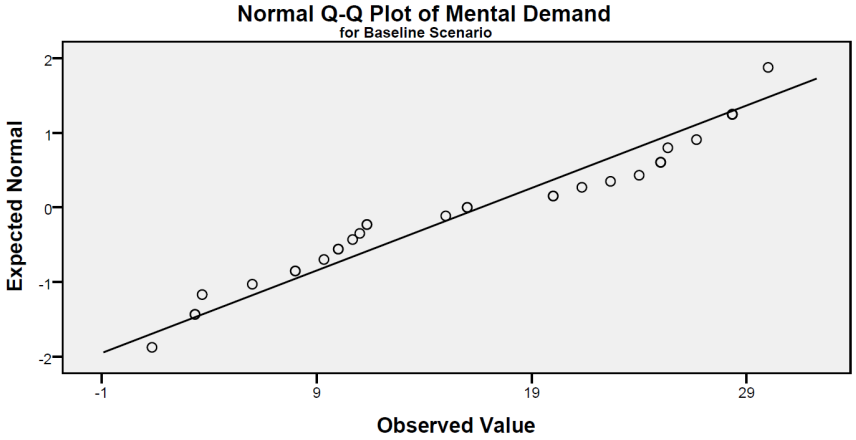


Figure C.1: Q-Q plot of the original mental demand data for the baseline scenario

C.1.1 Mental Demand

Table C.1 presented the descriptives of the raw MD data. The baseline scenario's data were considered not to be normally distributed at CI = 95% according to Shapiro-Wilk's test of normality with a significance value of $p = 0.022$. The graphical test illustrated in figure C.1 also demonstrated the non-normal distribution in the baseline data.

However, the descriptive statistics of the raw MD data of the evaluation scenario was in fact normally distributed with a significance value of $p = 0.308$.

Therefore, the parametric test assumptions could not be met for both sets of data, and transformation of any kind was not suitable to be applied, hence, a non-parametric test was used to analyse the differences in the operators' MD between the baseline and the evaluation scenarios.

C.1.2 Physical Demand

Table C.2 presented the descriptives of the raw PD data. Both scenarios' data were considered non-normally distributed at CI = 95% according to Shapiro-Wilk's test of normality with a significance value of $p_{baseline} < 0.001$ and $p_{evaluation} = 0.005$, and both sets of

Table C.2: Descriptive statistics of the physical demand data at CI=95%

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0(p < 0.001)$	Reject $H_0(p = 0.005)$
Standard Deviation	$\sigma = 5.2$	$\sigma = 5.243$
Skewness	0.813 $(SE = 0.393)$: Positive skew	0.736$(SE = 0.393)$: Positive skew
Kurtosis	-0.904 $(SE = 0.768)$: Flatter curve	-0.489 $(SE = 0.768)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_A} = 5.361$	$\bar{x}_B = 6.222$
Sample Size (N)	36	36
Similar Variances	Accept H_0 $(p = 0.994)$	

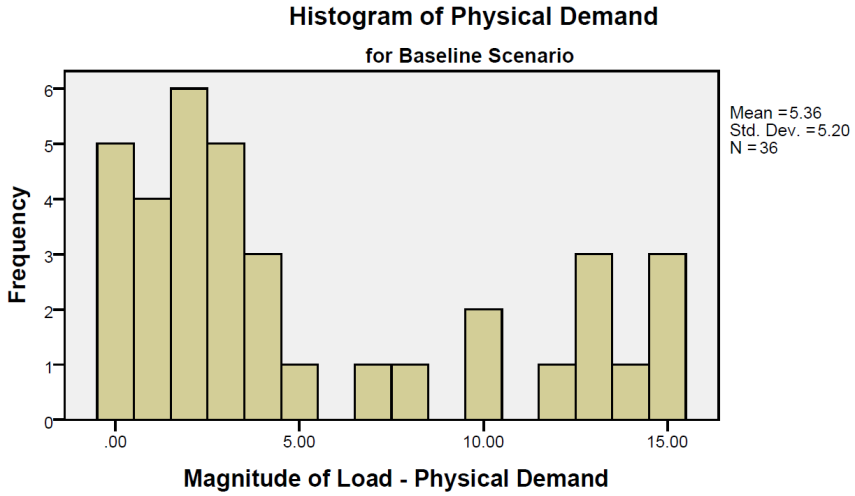


Figure C.2: Histogram of the physical demand data for the baseline scenario

data presented a positive skew, hence, the normality assumption could not be met for the application of the parametric test of the means.

In order to rectify the skewness, consequently rearranging the data into a normal distribution, a logrithmic or root transformation was necessary. Given the distribution and information presented in figure C.2 and C.3 for the baseline and the evaluation scenario (respectively), it was clear that the transformation could not be applied as a large number of sample points were zero (0), hence, a root/logrithmic calculation could not be successfully performed.

The assumptions of the non-parametric test were similar to the parametric test, however, it did not require the data sets to be normally distributed. The PD results were able to satisfy the remaining assumptions, hence, the non-parametric Wilcoxon Signed Rank Test was carried out to evaluate the differences between the PD experienced by the participants from the scenarios.

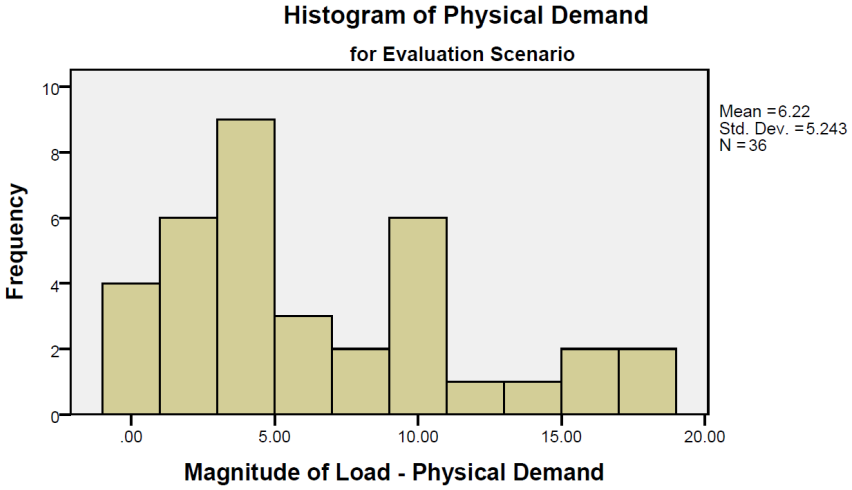


Figure C.3: Histogram of the physical demand data for the evaluation scenario

Table C.3: Descriptive statistics of the temporal demand data at CI=95%

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0(p = 0.031)$	Reject $H_0(p = 0.018)$
Standard Deviation	$\sigma = 4.239$	$\sigma = 4.399$
Skewness	-0.911 $(SE = 0.393)$: Negative skew	-0.803$(SE = 0.393)$: Negative skew
Kurtosis	1.222 $(SE = 0.768)$: Acceptable kurtosis	0.094 $(SE = 0.768)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_A} = 13.17$	$\bar{x}_B = 12.72$
Sample Size (N)	36	36

C.1.3 Temporal Demand

Table C.3 presented the descriptives of the raw PD data. Both scenarios' data were considered non-normally distributed at CI = 95% according to Shapiro-Wilk's test of normality with a significance value of $p_{baseline} = 0.031$ and $p_{evaluation} = 0.018$, and both sets of data presented a negative skew, hence, the normality assumption could not be met for the application of the parametric test of the means.

In order to rectify the skewness, consequently rearranging the data into a normal distribution, a squared transformation was necessary.

Table C.4 presented the descriptive statistics of the TD data with a squared transform applied. The analysis of the transformed data revealed that both scenarios' data sets were able to satisfy the normality distribution and similar variance assumption, and that the skewness and kurtosis of both sets of data were in range. Hence, a parametric test was applied to analyse the differences between the pair of comparison.

Table C.4: Descriptive statistics of the temporal demand data at CI=95% with a squared transformation applied

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.397)$	Accept $H_0(p = 0.380)$
Standard Deviation	$\sigma = 99.592$	$\sigma = 99.691$
Skewness	0.159 $(SE = 0.393)$: Acceptable skew	-0.030$(SE = 0.393)$: Acceptable skew
Kurtosis	-0.384 $(SE = 0.768)$: Acceptable kurtosis	-0.636 $(SE = 0.768)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_A} = 190.83$	$\bar{x}_B = 180.67$
Sample Size (N)	36	36
Similar Variance	Accept H_0 $(p = 0.826)$	

Table C.5: Descriptive statistics of the perceived performance data at CI=95%

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.304)$	Accept $H_0(p = 0.087)$
Standard Deviation	$\sigma = 3.843$	$\sigma = 3.977$
Skewness	0.208 $(SE = 0.393)$: Acceptable skew	-0.014$(SE = 0.393)$: Acceptable skew
Kurtosis	-0.622 $(SE = 0.768)$: Flatter curve	-0.892 $(SE = 0.768)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_A} = 10.97$	$\bar{x}_B = 9.89$
Sample Size (N)	36	36
Similar Variances	Accept H_0 $(p = 1.0)$	

C.1.4 Performance

Table C.5 presented the descriptives of the performance data. As the descriptives illustrated that the data set for both scenarios met the assumptions for the parametric hypothesis testing. The distribution for both scenarios were normal with a significance value of $p_{baseline} = 0.304$ and $p_{evaluation} = 0.087$ at CI = 95%, while the variances tested using Levene's test revealed that the significance value $p_{Levene} = 1.0$, hence accepting all H_0.

C.1.5 Effort

Table C.6 presented the descriptives of the original effort data. Along with the Q-Q plot and the histogram presented in figure C.4 and C.5, an outlier was identified in the baseline scenario data, which caused a negative skew to the data. The scores from this outlier was removed from both the baseline and the evaluation data sets for fairness.

The analysis was rerun and the results were presented in table C.7. Evidently, the removal of the outlier adjusted the data sets and at CI=95%, both data sets were in normal distribution ($p_{baseline} = 0.267$ and $p_{evaluation} = 0.607$) and had similar variance ($p = 0.13$).

Since the parametric test assumptions were all met, the pair sample T-Test was used to

Table C.6: Descriptive statistics of the perceived effort data at CI=95%

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.064)$	Accept $H_0(p = 0.616)$
Standard Deviation	$\sigma = 3.989$	$\sigma = 4.381$
Skewness	-0.379 $(SE = 0.393)$: Negative skew	-0.154$(SE = 0.393)$: Acceptable skew
Kurtosis	1.185 $(SE = 0.768)$: Acceptable kurtosis	-0.459 $(SE = 0.768)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_A} = 13.17$	$\bar{x}_B = 12.19$
Sample Size (N)	36	36
Similar Variances	Accept H_0 $(p = 0.414)$	

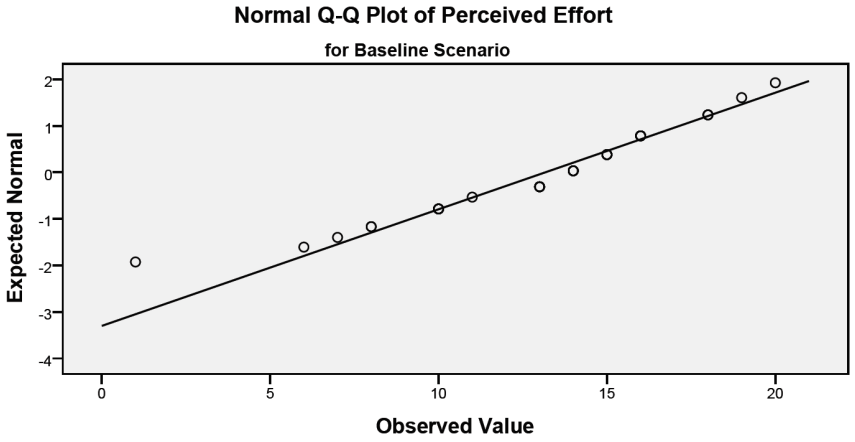


Figure C.4: Q-Q plot of the original effort data for the baseline scenario

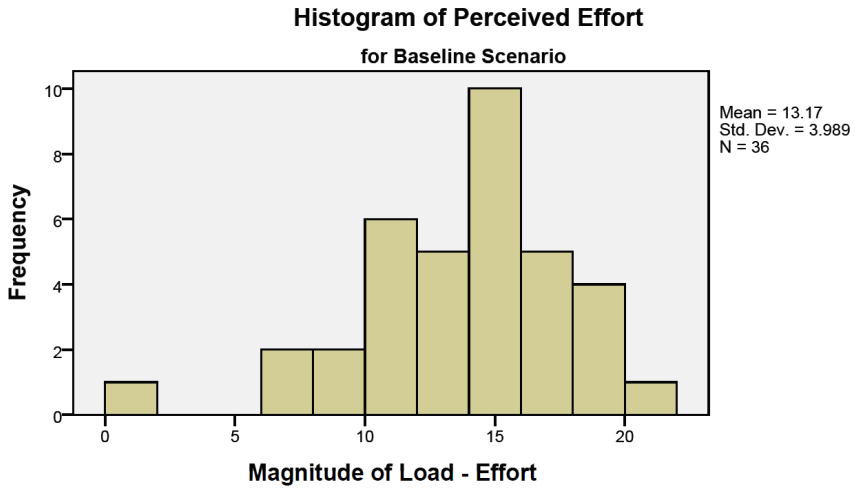


Figure C.5: Histogram of the original effort data for the baseline scenario

Table C.7: Descriptive statistics of the perceived effort data with the removal of the outlier at CI=95%

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.267)$	Accept $H_0(p = 0.607)$
Standard Deviation	$\sigma = 3.45$	$\sigma = 4.433$
Skewness	-0.355 $(SE = 0.398)$: Acceptable skew	-0.118$(SE = 0.398)$: Acceptable skew
Kurtosis	-0.375 $(SE = 0.778)$: Acceptable kurtosis	-0.515 $(SE = 0.778)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_A} = 13.51$	$\bar{x}_B = 12.14$
Sample Size (N)	35	35
Similar Variances	Accept H_0 $(p = 0.130)$	

Table C.8: Descriptive statistics of the participants' frustration level data at CI=95%

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.078)$	Reject $H_0(p = 0.011)$
Standard Deviation	$\sigma = 5.511$	$\sigma = 6.17$
Skewness	-0.355 $(SE = 0.393)$: Acceptable skew	-0.102$(SE = 0.393)$: Acceptable skew
Kurtosis	-0.854 $(SE = 0.768)$: Flatter curve	-1.235 $(SE = 0.768)$: Flatter curve
Sample Mean $(\bar{x})$	$\bar{x_A} = 10.53$	$\bar{x}_B = 9.86$
Sample Size (N)	36	36
Similar Variances	Accept H_0 $(p = 0.591)$	

analyse the differences in the means of the data sets.

C.1.6 Frustration

Table C.8 presented the descriptives of the raw frustration level data. The evaluation scenario's data were considered not normally distributed ($p = 0.011$) at CI = 95%. Transformation also could not be applied due to the fact that the distribution of the baseline data was normal, and skewness of both data sets were acceptable.

Given that both data sets were not all normally distributed, but had similar variances, the non-parametric Wilcoxon Signed-Rank test was used.

C.1.7 Combined Cognitive Workload

Table C.9 presented the descriptives of the raw combined CW data. The data set were not in normal distribution according to Shapiro-Wilk's test of normality, a negative skew was present.

Figure C.6 presented the boxplot for both the scenarios' data set. Evidently, samples 14, 23 and 27 were considered to be outliers. Hence, these sample points were removed from both the baseline and the evaluation data sets.

Table C.9: Descriptive statistics of the raw combined cognitive workload data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0(p = 0.001)$	Reject $H_0(p = 0.009)$
Standard Deviation	$\sigma = 16.102$	$\sigma = 16.514$
Skewness	-1.377 $(SE = 0.393)$: Negative skew	-0.891 $(SE = 0.393)$: Negative skew
Kurtosis	2.714 $(SE = 0.728)$: Acceptable kurtosis	0.103 $(SE = 0.709)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	63.17	58.72
Sample Size (N)	36	36

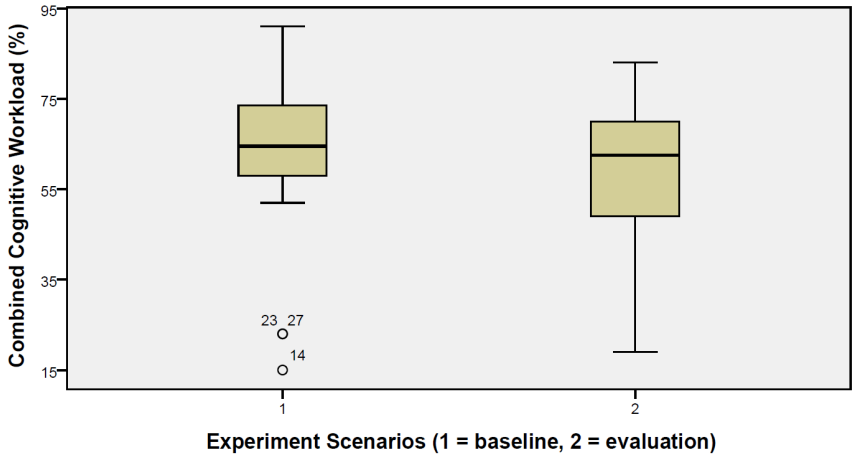


Figure C.6: Boxplot of the original combined CW data set

Table C.10 presented the descriptive statistics of the data sets from both scenarios with the removal of the three outliers. As a result, the normality and similar variance assumptions for the parametric test were satisfied at CI=95%. The baseline scenario had a normal distribution with a significance value of $p_{baseline} = 0.357$ and the evaluation scenario had a significance value of $p_{evaluation} = 0.087$. The data sets had a shared Levene's significance value of $p_{Levene} = 0.099$, hence, the data sets were similar in their variances.

Given that the assumptions for a pair-sample T-Test were met, the parametric test of the means was carried out.

C.2 Situation Awareness Assumptions Testing

This section detailed the process of assumptions testing of the Situation Awareness (SA) results collected in experiment 3 to support the statistical analyses presented in Chapter 7.

Table C.10: Descriptive statistics of the combined CW data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.357)$	Accept $H_0(p = 0.087)$
Standard Deviation	$\sigma = 9.727$	$\sigma = 12.997$
Skewness	0.504 $(SE = 0.409)$: Positive skew	0.737 $(SE = 0.409)$: Positive skew
Kurtosis	-0.011 $(SE = 0.798)$: Acceptable kurtosis	0.244 $(SE = 0.798)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 67.06$	$\bar{x}_{evaluation} = 61.94$
Sample Size (N)	33	33
Similar Variances	Accept H_0 $(p = 0.099)$	

Table C.11: Descriptive statistics and the assumptions testing outcome of the level 1 SA data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Numerical: Accept $H_0(p = 0.488)$	Numerical: Reject $H_0(p = 0.016)$
Standard Deviation	$\sigma = 16.369$	$\sigma = 15.781$
Skewness	-0.478 $(SE = 0.403)$: Negative skew	-0.970 $(SE = 0.403)$: Negative skew
Kurtosis	-0.278 $(SE = 0.788)$: Acceptable kurtosis	0.611 $(SE = 0.788)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_{baseline}} = 43.26$	$\bar{x_{evaluation}} = 52.892$
Sample Size (N)	34	34

C.2.1 Level 1 SA

Table C.11 presented the descriptives of the level 1 SA data. Evidently, the data collected from the baseline scenario was normally distributed. However, the evaluation scenario data was not, and also as identified from the boxplot in figure C.7, there was an outlier detected in the evaluation scenario's data set. Both data set had also presented negative skews in the shape of whose distributions.

The rectification process involved the removal of the outlier from both scenarios' data sets initially, and then followed by the application of a a squared transformation to both data sets.

After the removal of the outlier and the application of the squared transform, table xyz presented the descriptive statistics.

The outcomes presented in table C.12 demonstrated that the data sets with the outlier removed and square transformed were able to meet the assumptions for a parametric test, hence, the paired sample T-Test was used to compare the means.

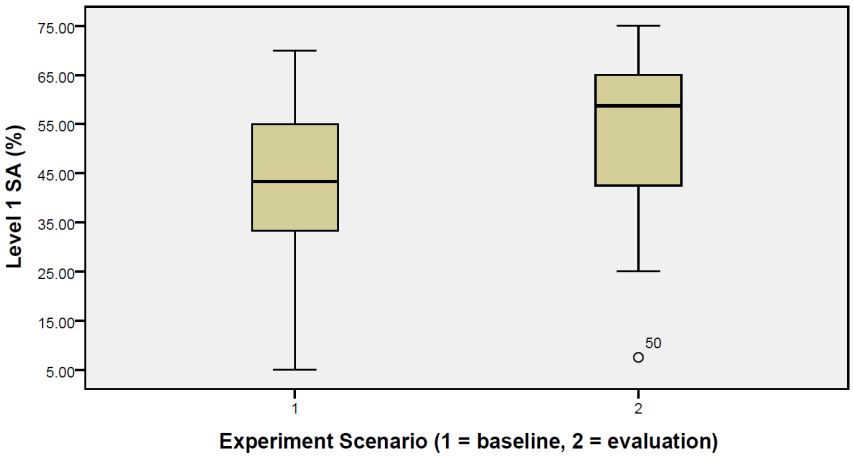


Figure C.7: Boxplot of the original level 1 SA data set

Table C.12: Descriptive statistics and the assumptions testing outcome of the square transformed level 1 SA data with the removal of an outlier

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.380)$	Accept $H_0(p = 0.171)$
Standard Deviation	$\sigma = 1299.112$	$\sigma = 1396.718$
Skewness	0.327 $(SE = 0.409)$: Acceptable skew	-0.236 $(SE = 0.409)$: Acceptable skew
Kurtosis	-0.802 $(SE = 0.798)$: Flatter curve	-0.991 $(SE = 0.798)$: Flatter curve
Sample Mean $(\bar{x})$	$\bar{x_{baseline}} = 2195.30$	$\bar{x_{evaluation}} = 3129.69$
Sample Size (N)	33	33
Similar Variances	Accept H_0 $(p = 0.507)$	

Table C.13: Descriptive statistics and the assumptions testing outcome of the level 2 SA data

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0(p = 0.016)$	Reject $H_0(p = 0.380)$
Standard Deviation	$\sigma = 15.781$	$\sigma = 14.943$
Skewness	-0.970 $(SE = 0.403)$: Negative skew	0.132 $(SE = 0.403)$: Acceptable skew
Kurtosis	0.611 $(SE = 0.788)$: Acceptable kurtosis	-0.374 $(SE = 0.788)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x_{baseline}} = 52.891$	$\bar{x_{evaluation}} = 56.303$
Sample Size (N)	34	34

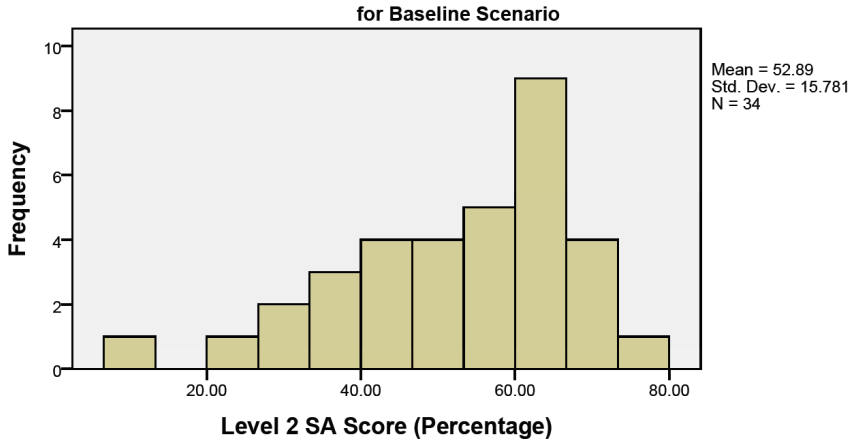


Figure C.8: Histogram of the participants' level 2 SA of the baseline scenario

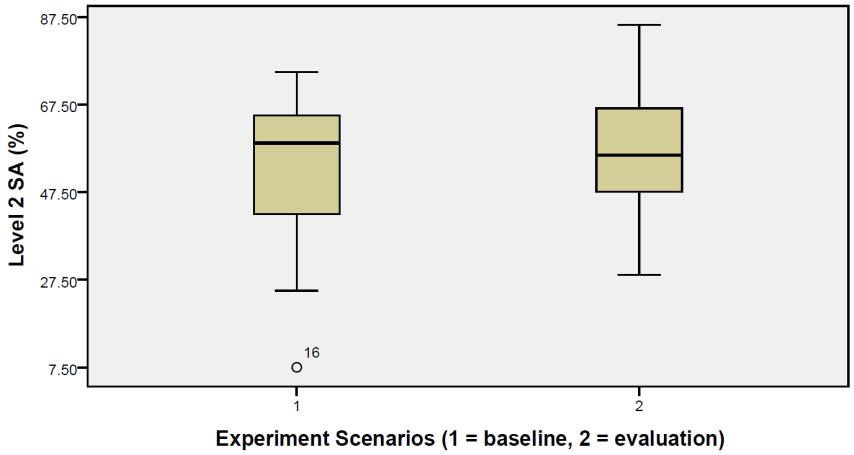


Figure C.9: Boxplot of the participants' level 2 SA in both scenarios, where 1 denoted the baseline scenario data, and 2 denoted the evaluation scenario data

C.2.2 Level 2 SA

Table C.13 presented the descriptives of the level 2 SA data. Evidently, the data collected from the baseline scenario was not normally distributed at CI=95% with a significance value of $p = 0.016$ (figure C.8), while the evaluation scenario was normally distributed. To contribute to this further, an outlier was identified in the baseline scenario as illustrated in figure C.9.

This outlier sample's data was removed from both the baseline and the evaluation scenario data for fairness, and table C.14 presented the subsequent descriptive statistics data.

The outcomes presented in table C.14 demonstrated that the data sets with the outlier removed, however, the baseline data set was still not normal as it was still negative skewed.

At this point, no transformation was applicable to process the data to simultaenously enable both sets of data to be in normal distribution. Hence, with a non-normal set of results, the outcomes were able to satisfy the assumptions required for a non-parametric

Table C.14: Descriptive statistics of the level 2 SA results with the removal of the outlier

($CI = 95\%, \alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0 (p = 0.035)$	Accept $H_0 (p = 0.423)$
Standard Deviation	$\sigma = 13.802$	$\sigma = 14.335$
Skewness	-0.634 ($SE = 0.409$): Negative skew	0.208 ($SE = 0.409$): Acceptable skew
Kurtosis	-0.579 ($SE = 0.798$): Flatter curve	-0.336 ($SE = 0.798$): Flatter curve
Sample Mean ($\bar{x}$)	$\bar{x}_{baseline} = 54.2677$	$\bar{x}_{evaluation} = 56.1428$
Sample Size (N)	33	33
Similar Variances	Accept H_0 ($p = 0.753$)	

Table C.15: Descriptive statistics of the raw combined situation awareness data

($CI = 95\%, \alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0 (p = 0.025)$	Reject $H_0 (p = 0.092)$
Standard Deviation	$\sigma = 14.179$	$\sigma = 12.267$
Skewness	-0.987 ($SE = 0.403$): Negative skew	0.225 ($SE = 0.403$): Acceptable skew
Kurtosis	1.118 ($SE = 0.788$): Acceptable kurtosis	-0.901 ($SE = 0.788$): Flatter curve
Sample Mean ($\bar{x}$)	50.074	60.36
Sample Size (N)	34	34

test.

C.2.3 Combined SA

Table C.15 presented the descriptives of the raw combined SA data. The data set were not in normal distribution according to Shapiro-Wilk's test of normality, a negative skew was present.

Figure C.10 presented the boxplot for both the scenarios' data set. Evidently, sample 16 was considered to be an outlier. Hence, it was removed from both the baseline and the evaluation data sets.

Table C.16 presented the descriptive statistics of the data sets from both scenarios with the removal of sample 16. As a result, the normality and similar variance assumptions for the parametric test were satisfied at CI=95%. The baseline scenario had a normal distribution with a significance value of $p_{baseline} = 0.114$ and the evaluation scenario had a significance value of $p_{evaluation} = 0.100$. The data sets had a shared Levene's significance value of $p_{Levene} = 0.720$, hence, the data sets were similar in their variances.

Given that the assumptions for a pair-sample T test were met, the parametric test of the means was carried out.

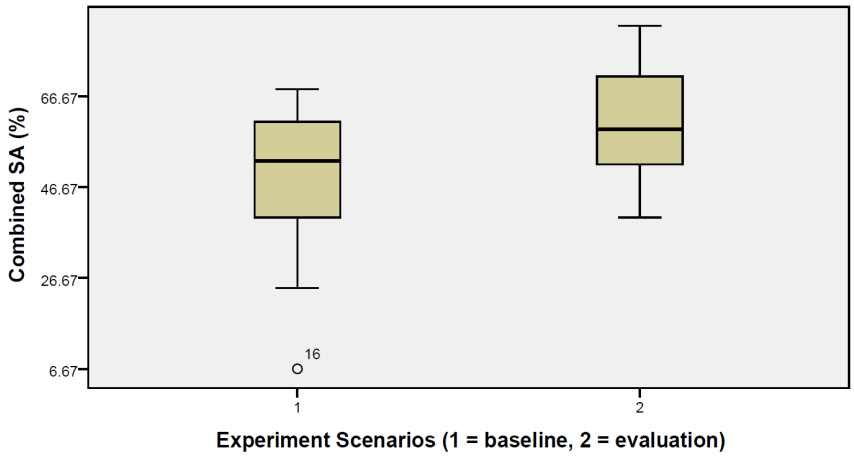


Figure C.10: Boxplot of the original combined SA data set

Table C.16: Descriptive statistics of the combined

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.114)$	Accept $H_0(p = 0.100)$
Standard Deviation	$\sigma = 12.111$	$\sigma = 11.94$
Skewness	0.454 $(SE = 0.409)$: Positive skew	0.243 $(SE = 0.409)$: Positive skew
Kurtosis	-0.790 $(SE = 0.798)$: Flatter curve	-0.898 $(SE = 0.798)$: Flatter curve
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 51.389$	$\bar{x}_{evaluation} = 60.96$
Sample Size (N)	33	33
Similar Variances	Accept H_0 $(p = 0.720)$	

C.3 Trust In Automation Assumptions Testing

This section detailed the process of assumptions testing of the Trust In Automation results collected in experiment 3 to support the statistical analyses presented in Chapter 7.

C.3.1 Competence

Table C.17 presented the descriptives of the competence level reported by the participants in both experiment scenarios. From the table, both data sets exhibited similar variences, though they were not normally distributed. Hence, a non-parametric Wilcoxon Signed-Rank test was used to analyse the means.

C.3.2 Predictability

Table C.18 presented the descriptives of the predictability level reported by the participants in both experiment scenarios. From the table, both data sets exhibited similar variances, though they were not normally distributed. Hence, a non-parametric Wilcoxon Signed-Rank test was used to analyse the means.

Table C.17: Descriptive statistics of the competence level reported by the participants during both experiment scenarios

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0 (p = 0.050)$	Reject $H_0 (p = 0.004)$
Standard Deviation	$\sigma = 1.762$	$\sigma = 1.507$
Skewness	-0.211 $(SE = 0.403)$: Acceptable skew	-0.793 $(SE = 0.403)$: Negative skew
Kurtosis	-0.757 $(SE = 0.788)$: Acceptable kurtosis	0.289 $(SE = 0.788)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 4.53$	$\bar{x}_{evaluation} = 4.97$
Sample Size (N)	34	34
Similar Variances	Accept H_0 $(p = 0.244)$	

Table C.18: Descriptive statistics of the predictability level reported by the participants during both experiment scenarios

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0 (p = 0.007)$	Reject $H_0 (p = 0.003)$
Standard Deviation	$\sigma = 1.498$	$\sigma = 1.234$
Skewness	-0.385 $(SE = 0.403)$: Acceptable skew	-0.090 $(SE = 0.403)$: Acceptable skew
Kurtosis	-0.954 $(SE = 0.788)$: Flatter curve	-0.292 $(SE = 0.788)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 4.62$	$\bar{x}_{evaluation} = 5.15$
Sample Size (N)	34	34
Similar Variances	Accept H_0 $(p = 0.133)$	

C.3.3 Reliability

Table C.19 presented the descriptives of the reliability level reported by the participants in both experiment scenarios. From the table, both data sets exhibited similar variances, though they were not normally distributed. Hence, a non-parametric Wilcoxon Signed-Rank test was used to analyse the means.

Table C.19: Descriptive statistics of the reliability level reported by the participants during both experiment scenarios

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0 (p = 0.109)$	Reject $H_0 (p = 0.017)$
Standard Deviation	$\sigma = 1.688$	$\sigma = 1.553$
Skewness	-0.246 $(SE = 0.403)$: Acceptable skew	-0.667 $(SE = 0.403)$: Negative skew
Kurtosis	-0.607 $(SE = 0.788)$: Acceptable kurtosis	-0.011 $(SE = 0.788)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 4.38$	$\bar{x}_{evaluation} = 4.79$
Sample Size (N)	34	34
Similar Variances	Accept H_0 $(p = 0.109)$	

Table C.20: Descriptive statistics of the amount of faith reported by the participants during both experiment scenarios

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.002)$	Reject $H_0(p = 0.005)$
Standard Deviation	$\sigma = 1.774$	$\sigma = 1.613$
Skewness	-0.943 $(SE = 0.403)$: Negative skew	-0.837 $(SE = 0.403)$: Negative skew
Kurtosis	0.373 $(SE = 0.788)$: Acceptable kurtosis	0.517 $(SE = 0.788)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 4.94$	$\bar{x}_{evaluation} = 5.06$
Sample Size (N)	34	34
Similar Variances	Accept H_0 $(p = 0.598)$	

Table C.21: Descriptive statistics of the amount of overall trust reported by the participants during both experiment scenarios

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.008)$	Reject $H_0(p = 0.007)$
Standard Deviation	$\sigma = 1.739$	$\sigma = 1.472$
Skewness	-0.777 $(SE = 0.403)$: Negative skew	-0.508 $(SE = 0.403)$: Negative skew
Kurtosis	0.166 $(SE = 0.788)$: Acceptable kurtosis	-0.531 $(SE = 0.788)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 4.65$	$\bar{x}_{evaluation} = 5.12$
Sample Size (N)	34	34
Similar Variances	Accept H_0 $(p = 0.402)$	

Non-Parametric Testing (Wilcoxon Signed-Rank Test)

C.3.4 Faith

Assumptions Testing

Table C.20 presented the descriptives of the faith level reported by the participants in both experiment scenarios. From the table, both data sets exhibited similar variances, though they were not normally distributed. Hence, a non-parametric Wilcoxon Signed-Rank test was used to analyse the means.

C.3.5 Overall Trust

Table C.21 presented the descriptives of the overall trust level reported by the participants in both experiment scenarios. From the table, both data sets exhibited similar variances, though they were not normally distributed, and transformations could not reorganise the data to be in normal distribution without severely obscuring the data. Hence, a non-parametric Wilcoxon Signed-Rank test was used to analyse the means.

Table C.22: Descriptive statistics of the IRT recorded from the participants during both experiment scenarios

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0(p = 0.001)$	Reject $H_0(p = 0.016)$
Standard Deviation	$\sigma = 34.619$	$\sigma = 22.503$
Skewness	1.787 $(SE = 0.421)$: Positive skew	1.242 $(SE = 0.421)$: Positive skew
Kurtosis	5.296 $(SE = 0.821)$: Acceptable kurtosis	1.986 $(SE = 0.821)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 63.624$	$\bar{x}_{evaluation} = 50.516$
Sample Size (N)	31	31

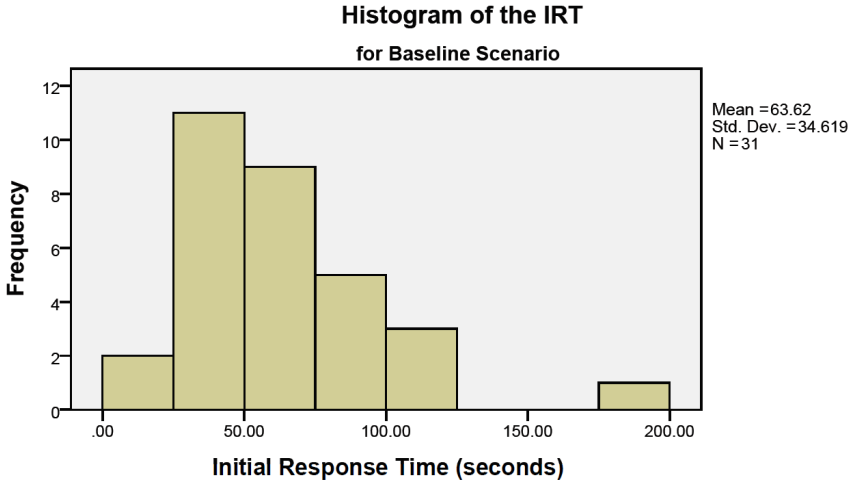


Figure C.11: Histogram of the baseline ERT data recorded from the participants

C.4 Operator Performance Assumptions Testing

This section detailed the process of assumptions testing of the operator objective performance results collected in experiment 3 to support the statistical analyses presented in Chapter 7.

C.4.1 Initial Response Time

Table C.22 presented the descriptives of the IRT data recorded from both experiment scenarios. From the table, both data sets were not normally distributed and severely positive skewed. These were further verified by visually inspecting the histograms of the baseline data set in figure C.11 and the evaluation data set in figureC.12.

A natural log transformation ($X_{ln} = log_e(X_{data})$) was applied to both data sets prior to further analysis. Table C.23 presented the descriptive statistics of the transformed data sets.

From the outcomes presented in table C.23, both transformed data sets had a normal distribution with a significance value of $p_{baseline} = 0.519$ and the evaluation scenario had

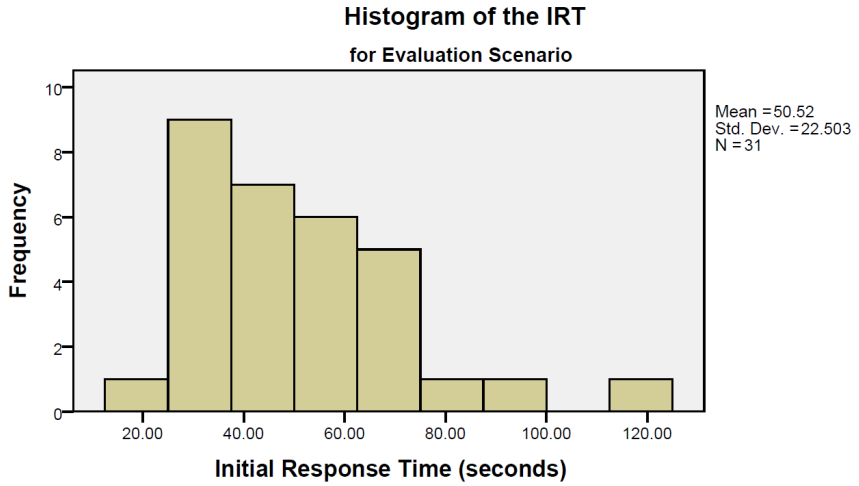


Figure C.12: Histogram of the evaluation IRT data recorded from the participants

Table C.23: Descriptive statistics of both sets of IRT data with the removal of the outlier and the application of a logrithmic transform

$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.939)$	Accept $H_0(p = 0.994)$
Standard Deviation	$\sigma = 0.519$	$\sigma = 0.424$
Skewness	-0.122 $(SE = 0.421)$: Acceptable skew	0.148 $(SE = 0.421)$: Acceptable skew
Kurtosis	0.269 $(SE = 0.821)$: Acceptable kurtosis	-0.271 $(SE = 0.821)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	$\bar{x}_{baseline} = 4.025$	$\bar{x}_{evaluation} = 3.834$
Sample Size (N)	31	31
Similar Variances	Accept H_0 $(p = 0.344)$	

a significance value of $p_{evaluation} = 0.424$. The data sets had a shared Levene's significance value of $p_{Levene} = 0.344$, hence, the data sets were similar in their variances.

Given that the assumptions for a pair-sample T-Test were met, this parametric test of the means was carried out.

C.4.2 Event Response Time

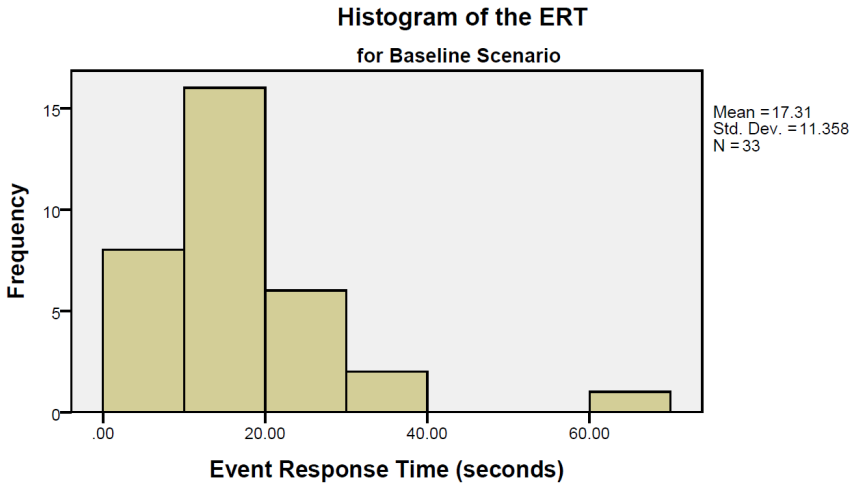
Table C.24 presented the descriptives of the ERT data recorded from both experiment scenarios. From the table, both data sets were not normally distributed and severely positive skewed. These were further verified by visually inspecting the histograms (figure C.13a and C.13a) and Q-Q plots of the baseline data and evaluation data (figure C.13a and C.13b).

Furthermore, sample 17 from the baseline data set were determined to be an outlier as illustrated in figure C.13, which was removed from both data sets for further analysis.

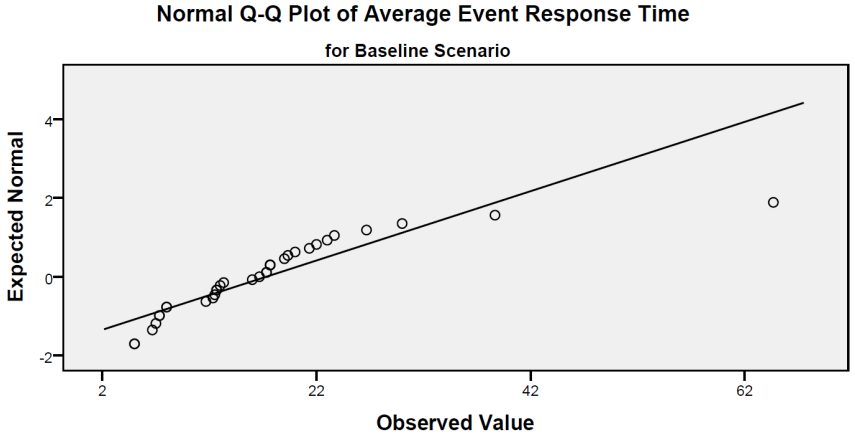
As seen from table C.24, there were significant positive skew associated with the data, hence, table C.25 presented the descriptive statistics of the data sets from both scenarios

Table C.24: Descriptive statistics of the ERT recorded from the participants during both experiment scenarios

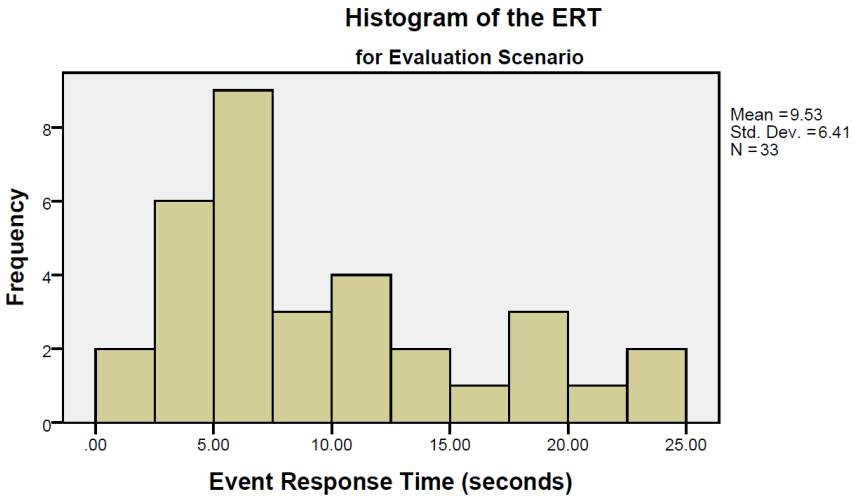
$(CI = 95\%, \alpha = 0.05)$	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Reject $H_0(p < 0.001)$	Reject $H_0(p = 0.01)$
Standard Deviation	$\sigma = 11.358$	$\sigma = 6.41$
Skewness	2.477 $(SE = 0.409)$: Positive skew	0.907 $(SE = 0.409)$: Positive skew
Kurtosis	8.899 $(SE = 0.798)$: Acceptable kurtosis	0.004 $(SE = 0.798)$: Acceptable kurtosis
Sample Mean $(\bar{x})$	17.309	9.531
Sample Size (N)	33	33



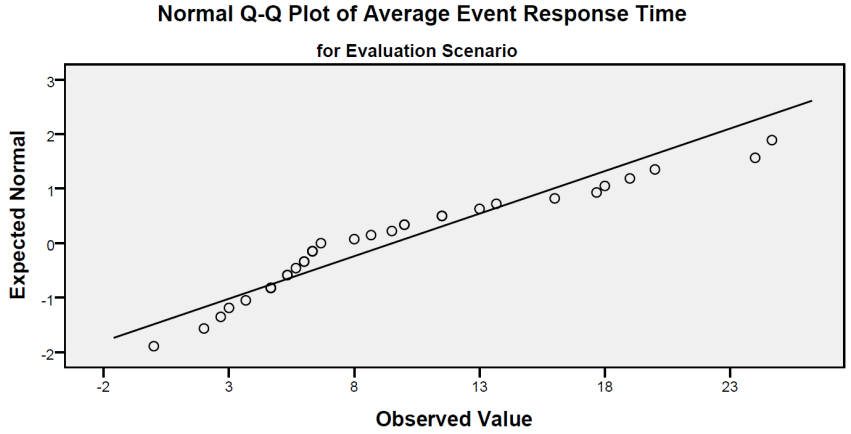
(a) Histogram of the baseline ERT data recorded from the participants



(b) Q-Q plot of the baseline ERT data recorded from the participants



(a) Histogram of the evaluation ERT data recorded from the participants



(b) Q-Q plot of the evaluation ERT data recorded from the participants

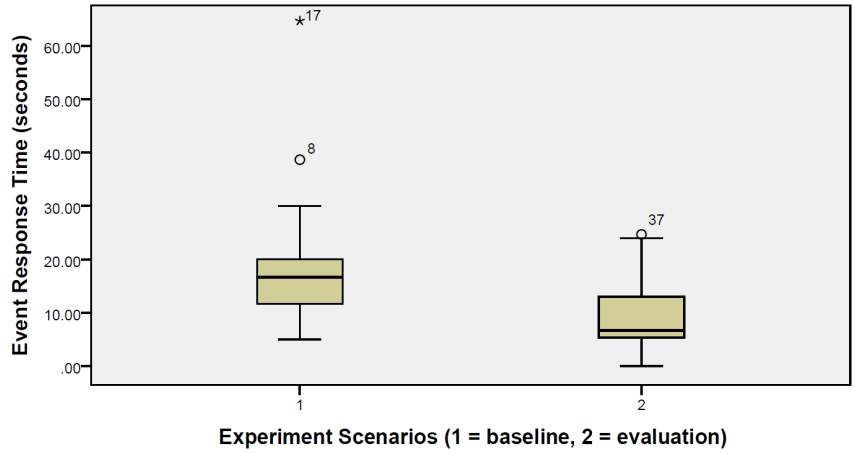


Figure C.13: Boxplot of both sets of ERT data

Table C.25: Descriptive statistics of both sets of ERT data with the removal of the outlier and the application of a logrithmic transform

($CI = 95\%, \alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.391)$	Accept $H_0(p = 0.618)$
Standard Deviation	$\sigma = 0.574$	$\sigma = 0.691$
Skewness	0.101 ($SE = 0.409$): Acceptable skew	-0.068 ($SE = 0.409$): Acceptable skew
Kurtosis	0.315 ($SE = 0.798$): Acceptable kurtosis	-0.614 ($SE = 0.798$): Acceptable kurtosis
Sample Mean ($\bar{x}$)	$\bar{x}_{baseline} = 2.588$	$\bar{x}_{evaluation} = 2.088$
Sample Size (N)	33	33
Similar Variances	Accept H_0 ($p = 0.299$)	

Table C.26: Descriptive statistics of the number of IOIs found (in percentage) data with the removal of three outliers

($CI = 95\%, \alpha = 0.05$)	**Baseline Scenario**	**Evaluation Scenario**
Normal Distribution	Accept $H_0(p = 0.262)$	Accept $H_0(p = 0.305)$
Standard Deviation	$\sigma = 12.807$	$\sigma = 9.668$
Skewness	-0.121 ($SE = 0.421$): Acceptable skew	0.260 ($SE = 0.421$): Acceptable skew
Kurtosis	-0.947 ($SE = 0.821$): Acceptable kurtosis	0.046 ($SE = 0.821$): Flatter curve
Sample Mean ($\bar{x}$)	$\bar{x}_{baseline} = 37.302$	$\bar{x}_{evaluation} = 45.577$
Sample Size (N)	31	31
Similar Variances	Accept H_0 ($p = 0.067$)	

with the removal of sample 17 and a logarithmic transformation applied. The table illustrated that the normality and similar variance assumptions for the parametric test were satisfied at CI=95%. The baseline scenario had a normal distribution with a significance value of $p_{baseline} = 0.391$ and the evaluation scenario had a significance value of $p_{evaluation} = 0.618$. The data sets had a shared Levene's significance value of $p_{Levene} = 0.299$, hence, the data sets were similar in their variances.

Given that the assumptions for a pair-sample T-Test were met, this parametric test of the means was carried out.

C.4.3 Items-Of-Interest Found

The initial data presented one outlier sample from the baseline results and two from the evaluation results, these outliers significantly influenced the distribution and variances of the data set, hence, they the three outliers were removed from both scenario data sets. Table C.26 presented the descriptive statistics of the data set with these outliers removed. The data sets were in normal distribution with CI=95% according to Shapiro-Wilk's test with significance values of $p_{baseline} = 0.262$ and $p_{evaluation} = 0.305$. Furthermore, also at CI=95%, the variances were similar with Levene's test $p_{Levene} = 0.067$. Hence, the

assumptions of the parametric T-Test had been met, and analysis was conducted with the pair-sample T-Test.

References

[1] (2003). New horizons for combat UAVs.

[2] (2010). Eyes of the army: U. s. army roadmap for unmanned aircraft systems 2010-2035. *Technical report*, U. S. Army UAS Center of Excellence.

[3] (2010). *National Search And Rescue Manual*. Australian Maritime Safety Authority, GPO Box 2181 Canberra ACT 2601 Australia.

[4] (2013). Unmanned systems integrated roadmap fy2013-2038. *Technical report*, Department of Defense.

[5] Allen, J. E. & Guinn, C. (1999). Mixed-initiative interaction. *Intelligent Systems and their Applications*, *14*(5), 14–23.

[6] Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine*, *22*(4), 27–38.

[7] Alvarado, M., Gonzalez, F., Fletcher, A., & Doshi, A. (2015). Towards the development of a low cost airborne sensing system to monitor dust particles after blasting at open-pit mine sites.

[8] Amelink, M. & Aguero, C. (2010). Structuring information during the design process of a complex human-machine interface. In *The 2nd IEEE International Conference on Information Management and Engineering (ICIME)*.

[9] Amelink, M. H. J. (2010). *Ecological Automation Design, Extending Work Domain Analysis* (Dissertation). Delft University of Technology.

[10] Aubert, T., Corjon, J., Gautreault, F., Laurent, M., Corjon, J., Causse, M., & Dehais, F. (2010). Improving situation awareness of a single human operator interacting with multiple unmanned vehilces: First results. In *HUMans Operating Unmanned Systems*.

[11] Barber, B. (1983). *The Logic and Limits of Trust*. Rutger University Press.

[12] Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation usage decision: Controlling intent and appraisal errors in a target detection task. *Human Factors*, *49*(3), 429–437.

[13] Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction*, *3*(2), 74–99.

[14] Bernstein, N. (1967). *Coordination and Regulation of Movement*. New York: Pergamon.

[15] Billings, C. (1997). *Aviation Automation: The Search for a Human-Centered Approach*. Mahwah: Erlbaum.

[16] Billings, C. E. & Woods, D. D. (1994). *Human Performance in Automated Systems: Current Research and Trends*, chapter Concerns about Adaptive Automation in Aviation Systems, pp. 24–29. Hillsdale: Lawrence Erlbaum.

[17] Bisantz, A. M., Llinas, J., Seong, Y., Finger, R., & Jian, J.-Y. (2000). Empirical investigations of trust-related systems vulnerabilities in aided, adversarial decision making. *Interim Report ADA389378*, State University of New York at Buffalo center of Multisource Information Fusion.

[18] Bitan, Y. & Meyer, J. (2007). Self-initiated and respondent action in a simulated control task. *Ergonomics*, *50*(5), 763–788.

[19] Boyd, J. (1987). A discourse on winning and losing.

[20] Bradshaw, J. M., Feltovich, P. J., Jung, H., Kulkarni, S., Taysom, W., & Uszok, A. (2004). Dimensions of adjustable autonomy and mixed-initiative interaction. In M. Nickles, M. Rovatsos, & G. Weiss (Eds.), *Agents and Computational Autonomy*, volume 2969 of *Lecture Notes in Computer Science*, pp. 17–39, Springer Berlin Heidelberg, ISBN 978-3-540-22477-8.

[21] Bratman, M. E. (1987). *Intent, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

[22] Bratman, M. E., Israel, D. J., & Pollack, M. E. (1988). Plans and resource-bounded practial reasoning. *Computational Intelligence*, *4*(4), 349–355.

[23] Brennan, R. W., Wang, J., & Norrie, D. H. (2004). A control-based approach to distributed control system design. In *Proceedings of the Canadian Engineering Education Association*.

[24] Breton, R. & Rousseau, R. (2005). The C-OODA: A cognitive version of the OODA loop to represent C^2 activities. In *International Command and Control Research and Technology Symposium*, Command and Control Process Modeling Group, Defence Research and Development Canada - Valcartier.

[25] Brick, T. & Scheutz, M. (2007). Incremental natural language processing for HRI. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 263–270.

[26] Broadbent, D. E. (Ed.) (2013). *Perception and Communication*. London U.K.: Pergamon Press.

[27] Brown, S., Toh, J., & Sukkarieh, S. (2005). A flexible human-robot team framework for information gathering mission. In *Proceedings of Australasian Conference of Robotics and Automation*.

[28] Bruni, S., Marquez, J. J., Brzezinski, A., Nehme, C., & Boussemart, Y. (2007). Introducing a human-automation collaboration taxonomy (HACT) in command and control decision-support systems. In *Proceedings of 12th International Command and Control Research and Technology Symposia*.

[29] Burns, C. M., Jr., G. S., Jamieson, G. A., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of ecological interface design for nuclearn process control: Situation awareness effects. *Human factors*, *50*(4), 663–679.

[30] Byers, J. C., Bittner, A. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32, pp. 1145–1149.

[31] Casbeer, D. W., Beard, R. W., McLain, T. W., Li, S.-M., & Mehra, R. K. (2005). Forest fire monitoring with multiple small UAVs. In *2005 American Control Conference*.

[32] Casbeer, D. W., Kingston, D. B., Beard, R. W., & McLain, T. W. (2006). Cooperative forest fire surveillance using a team of small unmanned air vehicles. *International Journal of Systems Science*, *37*(6), 351–360.

[33] Castillo-Effen, M., Alvis, W., & Castillo, C. (2005). Modeling and visualization of multiple autonomous hetergeneous vehilces. In *IEEE International Conference on Systems, Man and Cybernetics*.

[34] Chaput, A. J. Aircraft design and design laboratory (2004). Retrieved from `http://www.ae.utexas.edu/ASE261KChaput/`.

[35] Chen, J. Y. C. & Barnes, M. J. (2012). Control of multiple robots in dynamic tasking environments. *Ergonomics*, *55*, 1043–1058.

[36] Chen, J. Y. C. & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors*, *54*, 157–174.

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

[38] Chen, J. Y. C., Barnes, M. J., & Harper-Sciarini, M. (2011). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, *41*(4), 435–454.

[39] Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. *Technical report*, U.S. Army Research Laboratory.

[40] Chen, T. B., Campbell, D., Coppin, G., & Gonzalez, F. (2013). Management of heterogeneous UAVs through a capability framework of UAV's functional autonomy. In *15th Australian International Aerospace Congress*, Melbourne, Australia.

[41] Chen, T. B., Campbell, D., Coppin, G., Mooij, M., & Gonzalez, F. (2012). A capability framework visualisation for multiple heterogeneous uavs: From a mission commander's perspective. In *28th International Congress of the Aeronautical Sciences*, Brisbane Convention and Exhibition Centre, Brisbane.

[42] Chen, T. B., Campbell, D., Gonzalez, F., & Coppin, G. (2014). The effect of autonomy transparency in human-robot interactions: A preliminary study on operator cognitive workload and situation awareness in multiple heterogeneous UAV management. In *Proceedings of the Australasian Conference on Robotics and Automation*.

[43] Chen, T. B., Campbell, D., Gonzalez, F., & Coppin, G. (2015). Increasing autonomy transparency through capability communication in multiple heterogeneous UAV management. In *Inproceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[44] Chen, T. B., Gonzalez, F., Campbell, D., & Coppin, G. (2014). Management of multiple heterogeneous UAVs using capability and autonomy visualisation: Theory, experiment and result. In *International Conference on Unmanned Aircraft Systems*, Orlando, Florida.

[45] Clare, A. S., Maere, P. C. P., & Cummings, M. L. (2012). Assessing operator strategies for real-time replanning of multiple unmanned vehicles. *Intelligent Decision Technologies*, *6*, 221–231.

[46] Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

[47] Cleary, M., M., A., Adams, M. B., & Kolitz, S. (2000). Metrics for embedded collaborative intelligent systems. *Technical report*, Charles Stark Draper Laboratory, Inc.

[48] Clough, B. T. (2002). Metrics, schmetrics! how the heck do you determine a UAV's autonomy anyway? In *Performance Metrics for Intelligent Systems Workshop*, Gaithersburg, Maryland.

[49] Cohen, P. R. & Oviatt, S. L. (1995). The role of voice input for human-machine communication. In *Proceedings of the National Academy of Science*.

[50] Committee on Autonomous Vehicles in Support of Naval Operations, N. R. C. (2005). *Autonomous Vehicles in Support of Naval Operations*. The National Academies Press. Retrieved from `http://www.nap.edu/openbook.php?record_id=11379`.

[51] Coppin, G. & Legras, F. (2012). Autonomy spectrum and performance perception issues in swarm supervisory control. In *Proceedings of the IEEE*, volume 100.

[52] Coppin, G., Legras, F., & Bazalgette, D. (2008). FRA-1: UAV swarm control - SMAART project "interacting with multi-agent systems / UAV swarms.

[53] Coppin, G., Legras, F., & Saget, S. (2009). Supervision of autonomous vehicles: Mutual modeling and interaction management. In D. Harris (Ed.), *Engineering*

Psychology and Cognitive Ergonomics, volume 5639 of *Lecture Notes in Computer Science*, pp. 489–497, Springer Berlin Heidelberg, ISBN 978-3-642-02727-7.

[54] Coutaz, J. (1987). PAC: An object oriented model for implementing user interfaces. *ACM SIGCHI Bulletin*, *19*(2), 37–41.

[55] Cummings, M. (2007). Collaborative human-computer decision making for command and control resource allocation. *Technical report*, Massachusetts Institute of Technology, Cambridge.

[56] Cummings, M. & Bruni, S. (2009). *Collaborative Human-Automation Decision Making*. 437-447, Springer.

[57] Cummings, M. & Guerlain, S. (2007). Developing operator capacity estimates for supervisory control of autonomous vehicles. *Human Factors*, *49*, 1–15.

[58] Cummings, M. L., Bertucelli, L., Macbeth, J., & Surana, A. (2014). Task versus vehicle-based control paradigm in multiple unmanned vehicle supervision by a single operator. *IEEE Transactions on Human-Machine Systems*, *44*(3), 353–361.

[59] Cummings, M. L. & Brzezinski, A. (2010). Global vs. local decision support for multiple independent UAV schedule management. *International Journal of Applied Decision Sciences*, *3*(3), 188–205.

[60] de Dios, J. R. M., Merino, L., Ollero, A., Ribeiro, L. M., & Viegas, X. (2007). Multi-UAV experiments: Application to forest fires. In *Multiple Heterogeneous Unmanned Aerial Vehicles*, pp. 207–228, Springer.

[61] de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In *6th International Conference on Virtual, Augmented and Mixed Reality*.

[62] Dehais, F., Tessier, C., & Chaudron, L. (2003). GHOST: Experimenting conflicts countermeasures in the pilot's activity. In *the 18th Joint Conference on Artificial Intelligence*.

[63] Dehais, F., Tessier, C., Christophe, L., & Reuzeau, F. (2010). The perseveration syndrome in the pilot's activity: Guidelines and cognitive countermeasures. In P. Palanque, J. Vanderdonckt, & M. Winckler (Eds.), *Human Error, Safety and Systems Development*, volume 5962 of *Lecture Notes in Computer Science*, pp. 68–80, Springer Berlin Heidelberg, ISBN 978-3-642-11749-7.

[64] del Arco, J. C., Alejo, D., Arrue, B. C., Cobano, J. A., Heredia, G., & Ollero, A. (2015). Multi-UAV ground control station for gliding aircraft. In *23rd Mediterranean Conference on Control and Automation (MED)*.

[65] Delogne, R. (1999). The B-HUNTER UAV system. *Technical report*, EAGLE Temporary Association.

[66] Deutsch, M. (1958). Trust and suspicion. *The Journal of Conflict Resolution*, *2*(4), 265–279.

[67] Djuknic, G. M., Freidenfelds, J., & Okunev, Y. (1997). Establishing wireless communications services via high-altitude aerionautical platforms: A concept whose time has come? *IEEE Comunications Magazine*, *35*(9), 128–135.

[68] Donmez, B., carl Nehme, & Cummings, M. L. (2010). Modeling workload impact in multiple unmanned vehicle supervisory control. *IEEE Transactions on Systems Man and Cybernetics, Part A: Systems and Humans*, *40(6)*, 1180–1190.

[69] Drury, J. L., Riek, L., & Rackliffe, N. (2006). A decomposition of UAV-related situation awareness. In *1st Annual Conference on Human-Robot Interaction*.

[70] Drury, J. L. & Scott, S. D. (2008). Awareness in unmanned aerial vehicle operations. *The International C2 Journal*, *2*(1), 1.

[71] Duval, T. (2012). Models for design, implementation and deployment of 3d collaborative virtual environments. *Technical report*, Universitè Rennes.

[72] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

[73] Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*, volume 3, pp. 789–795. doi:10.1109/NAECON.1988.195097.

[74] Endsley, M. R. (1993). Situation awareness and workload: Flip sides of the same coin. In *Proceedings of the 7th International Symposium on Aviation Psychology*.

[75] Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 65 – 84. doi:10.1518/001872095779049499.

[76] Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems: Situation awareness. *Human Factors*, *37*(1), 32–64.

[77] Endsley, M. R. (2000). *Direct Measurement of Situation Awareness: Validity and Use of SAGAT*. Lawrence Erlbaum Associates.

[78] Endsley, M. R. & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492.

[79] Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). A comparative analysis of SAGAT and SART for evaluations of situation awareness. *Proc. Human Factors and Ergonomics Society Annual Meeting*, *42*(5), 82–86.

[80] Engel, F. L. (1980). *Ergonomic Aspects of Visual Display Terminals*, chapter Information Selection from Visual Displays. London: Taylor and Francis Ltd.

[81] Ezzedine, H., Kolski, C., & Peninou, A. (2005). Agent-oriented design of human-computer interface: Application to supervision of an urban transport network. *Engineering Applications of Artificial Intelligence*, *18*, 255–270.

[82] Feitshans, G. L., Rowe, A. J., & Davis, J. E. (2008). Vigilant spirit control station (VSCS) "the face of COUNTER". In *AIAA Guidance, Navigation and Control Conference and Exhibit*.

[83] Franke, J. L., Zaychik, V., Spura, T. M., & Alves, E. E. (2005). Inverting the operator/vehicle ratio: Approach to next generation UAV command and control. In *Association for Unmanned Vehicle Systems International*.

[84] Freeman, E., Freeman, E., Sierra, K., & Bates, B. (2004). *Head First Design Patterns*. O'Reilly Media Inc.

[85] Frische, F. & Ludtke, A. (2013). SA-tracer: A tool for assessment of UAV swarm operator SA during mission execution. In *IEEE International Multi-Disciplinary conference on Cognitive Methods in Situation Awareness and Decision Support*.

[86] Fuchs, C., Ferreira, S., Sousa, J., & Gonçalves, G. (2013). Adaptive consoles for supervisory control of multiple unmanned aerial vehicles. In *Human-Computer Interaction. Interaction Modalities and Techniques*, pp. 678–687.

[87] Galitz, W. O. (2007). *The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques*, chapter Choose the Proper Colors, pp. 691–726. John Wiley & Sons.

[88] Goldberg, A. (1990). Information models, views, and controllers. *Dr. Dobb's Journal*, *15*(7), 54–61.

[89] Gonçalves, R., Ferreira, S., Pinto, J., Sousa, J., & Gonçalves, G. (2011). Authority sharing in mixed initiative control of multiple uninhabited aerial vehicles. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics*, volume 6781 of *Lecture Notes in Computer Science*, pp. 530–539, Springer Berlin Heidelberg, ISBN 978-3-642-21740-1. Retrieved from `http://dx.doi.org/10.1007/978-3-642-21741-8_56`, `doi:10.1007/978-3-642-21741-8_56`.

[90] Gonzalez, F., Narayan, P., Castro, M. P. G., Walker, R., & Zeller, L. (2011). Development of an autonomous unmanned aerial system to collect time-stamped samples from the atmosphere and localize potential path sources. *Journal of Field Robotics, Special Issue: Safety, Security, and Rescue Robotics*, *28*(6), 961–976.

[91] Gonzalez, L. F., Lee, D.-S., & Walker, R. A. (2009). Optimal mission path planning (MPP) for an air sampling unmanned aerial system. In *Proceedings of the 2009 Australasian Conference on Robotics & Automation*, pp. 1–9, Australian Robotics & Automation Association.

[92] Gonzalez, L. F., Richardson, D., & Vercoe, M. (2014). Development of multi-rotor localised surveillance using multi-spectral sensors for plant biosecurity. *Technical report*, Australian Research Centre for Aerospace Automation.

[93] Goodrich, M. A., Morse, B. S., Gerhardt, D., Cooper, J. L., Quigley, M., Adams, J. A., & Humphrey, C. (2007). Supporting wilderness search and rescue using a camera-equipped mini uav. *Journal of Field Robotics*, *25*(1-2), 89–110.

[94] Gupta, S. G., Ghonge, M. M., & Jawandhiya, P. M. (2013). Review of unmanned aircraft systems (UAS). *2*(4), 1646–1658.

[95] Ham, D. H. & Yoon, W. C. (2001). Design of information content and layout for process control based on goal-means domain analysis. *Cognition, Technology and Work*, *3*, 205–223.

[96] Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

[97] Hart, S. G. & Staveland, L. E. (1988). *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*, chapter 7, pp. 139–178. Elsevier Science Publishers B.V. (North-Holland).

[98] Hart, S. G. & Wempe, T. E. (1979). Cockpit display of traffic information: Airline pilot opinions about content, symbology, and format. *Technical report*, NASA Ames Research Center, Moffett Field, CA, NASA.

[99] Hendy, K. C., Hamilton, K. M., & Landry, L. N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors*, *35*(4), 579–601.

[100] Henry, J. C. (2006). Electroencephalography: Basic principles, clinical applications, and related fields. *Neurology*, *67*(11), 2092.

[101] Hooey, B. L., Gore, B. F., Wickens, C. D., Scott-Nash, S., Socash, C., Salud, E., & Foyle, D. C. (2011). *Human Modeling in Assisted Transportation*, chapter Modeling Pilot Situation Awareness, pp. 207–213. Springer.

[102] How, J. P. (2004). Multi-vehicle experimental platform for distributed coordination and control. *Technical report*, Massachusetts Institute of Technology Cambridge, Department of Aeronautics and Astronautics.

[103] Huang, H. M. (2007). Autonomy levels for unmanned systems (ALFUS) framework: safety and application issues. In *Proceedings of the 2007 Workshop on Performance Metrics for Intelligent Systems*, PerMIS '07, pp. 48–53, New York, NY, USA: ACM, ISBN 978-1-59593-854-1.

[104] Huang, H. M., Messina, E., & Albus, J. (2007). *Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume II: Framework Models*. National Institute of Standards and Technology, 1 edition. NIST Special Publication 1011-II-1.0.

[105] Huang, H. M., Messina, E., Wade, R., English, W., Novak, B., & Albus, J. (2005). A framework for autonomy levels for unmanned systems (ALFUS). In *Proceedings of the AUVSI's Unmanned Systems North America*.

[106] Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, *57*(1), 94–103.

[107] Inagaki, T. (2003). Automation and the cost of authority. *International Journal of Industrial Ergonomics*, *31*(3), 169–174.

[108] Israel, M., Mende, M., & Keim, S. (2015). Uavrc, a generic MAV flight assistance software. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

[109] Jamieson, G. A. (2007). Ecological interface design for petrochemical process control: An empirical assessment. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *37*(6), 906–920.

[110] Jamieson, G. A. & Vicente, K. J. (2001). Ecological interface design for petrochemical applications: Supporting operator adaptation, continuous learning, and distributed, collaborable work. *Computers and Chemical Engineering*, *25*(7–8), 1055–1074.

[111] Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirical determine scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

[112] Johns, J. L. (1996). A concept analysis of trust. *Journal of Advanced Nursing*.

[113] Johnson, C. (2003). Inverting the control ratio: Human control of large, autonomous teams. In *Workshop Humans Multi-Agent Systems*.

[114] Jones, C. A. (1997). Unmanned aerial vehicles (UAVs) an assessment of historical operations and future possibilities. *Technical report*, Air Command and Staff College USAF.

[115] Jovanovic, M. & Starcevic, D. (2008). Software architecture for ground control station for unmanned aerial vehicle. In *10th International Conference on Computer Modeling and Simulation*.

[116] Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing*, *10*(4), 409–430.

[117] Kaber, D. B., Wright, M. C., Prinzel, L. J., & Clamann, M. P. (2005). Adaptive automation of human-machine system information-processing functions. *Journal of Human Factors*, *47*(4), 730–741. arXiv:http://hfs.sagepub.com/content/47/4/730.full.pdf+html, doi:10.1518/001872005775570989.

[118] Karl, L. R., Petty, M., & Shneiderman, B. (1993). Speech versus mouse commands for word processing: An empirical evaluation. *International Journal of Man-Machine Studies*, *39*(4), 667–687.

[119] Kazi, T. A., Stanton, N. A., Walker, G. H., & Young, M. S. (2007). Design driving: Driver's conceptual models and level of trust in adaptive cruise control. *International Journal of Vehicle Design*, *45*(3).

[120] Kim, T. & Hinds, P. (2006). Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *Robot and Human Interactive Communication, 2006. ROMAN 2006.*

[121] Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, eduring questions. *50*, 569–598.

[122] Krasner, G. E. & Pope, S. T. (1988). A cookbook for using the model-view-controller user interface paradigm in smalltalk-80. *Journal of Object Oriented Programming*, pp. 26–49.

[123] Langley Air Force Base, V. MQ-1B predator accident report released (2013), April. Retrieved from `http://www.acc.af.mil/news/story.asp?id=123343632`.

[124] Larzelere, R. E. & Huston, T. L. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and Family*, *42*(3), 595–604.

[125] Lee, J. D. (2008). Review of a pivotal human factors article: "humans and automation: Use, misuse, disuse, abuse". *Human Factors*, *50*(3), 404–410.

[126] Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

[127] Legras, F. & Coppin, G. (2007). Autonomy spectrum for a multiple UAVs system. In *COGIS' 07 - COgnitive systems with Interactive Sensors*, Stanford University, Palo Alto, CA, USA.

[128] Legras, F., Glad, A., Simonin, O., & Charpillet, F. (2008). Authority sharing in a swarm of UAVs: Simulation and experiments with operators. In S. Carpin, I. Noda, E. Pagello, M. Reggiani, & O. von Stryk (Eds.), *Simulation, Modeling, and Programming for Autonomous Robots*, volume 5325 of *Lecture Notes in Computer Science*, pp. 293–304, Springer Berlin Heidelberg, ISBN 978-3-540-89075-1.

[129] Lemon, O., Bracy, A., Gruenstein, A., & Peters, S. (2001). the WITAS multi-modal dialogue system i. In *Interspeech*.

[130] Lemon, O., Cavedon, L., Gruenstein, A., & Peters, S. (2002). Collaborative dialogue for controlling autonomous systems. *Technical Report FS-02-03*, American Association for Aritificial Intelligence.

[131] Li, L., Heymsfield, G., Carswell, J., Schaubert, D. H., McLinden, M. L., Creticos, J., Perrine, M., Cervantes, M. C. J. I., Vega, M., Guimond, S., Tian, L., & Emory, A. (2016). The NASA high-altitude imaging wind and rain airborne profiler.

[132] Liu, D., Wasson, R., & Vincenzi, D. A. (2009). Effects of system automation management strategies and multi-mission operator-to-vehicle ratio on operator performance in UAV systems. *Journal of Intelligent Robotics Systems*, *54*, 795–810.

[133] Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *Trust and Autonomous Systems: the 2013 AAAI Spring Symposium*.

[134] Lyons, J. B. & Havig, P. R. (2014). *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environment*, chapter Transparency in a Human-Machine Context: Approach for Fostering Shared Awareness/Intent, pp. 181–190. Springer.

[135] Mackintosh, M.-A., Lozito, S., McGann, A., & Logsdon, E. (1999). Designing procedures for controller-pilot data link communcation: Effects of textual data link on information transfer. In *1999 World Aviation Conference*.

[136] Madsen, M. & Gregor, S. (2000). Measuring human-computer trust. In *11th Australasian Conference on Information Systems*.

[137] Manktelow, K. & Jones, J. (1987). *Principles from the Psychology of Thinking and Mental Models*. Chichester, England: Wiley.

[138] Marcus Watson, W. J. R. & Sanderson, P. (2000). Ecological interface design for anaesthesia monitoring. *Australiasian Journal of Information Systems*, *7*(2).

[139] Marquez, J. J. & Cummings, M. L. (2008). Design and evaluation of path planning decision support for planetary surface exploration. *Journal of Aerospace Computing, Information, and Communication*, *5*(3), 57–71.

[140] Master, R., Jiang, X., Khasawneh, M. T., Bowling, S. R., larry Grimes, Gramopadhye, A. K., & Melloy, B. J. (2005). Measurement of trust over time in hybrid inspection systems. *Human Factors and Ergonomics in Manufacturing*, *15*(2), 177–196.

[141] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *20*(3), July.

[142] Mcfadyen, A. (2015). *Visual Control for Automated Aircraft Collision Avoidance Systems* (Doctoral dissertation).

[143] Mehrabian, A. & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, *31*(3), 248–252.

[144] Mehrabian, A. & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, *6*(1), 109–114.

[145] Mercado, J. E., Chen, J. Y. C., Barnes, M., Rupp, M. A., Barber, D., & Procci, K. (2015). Effects of agent transparency on multi-robot management effectiveness. *Technical report*, US Army Research Laboratory.

[146] Mercier, S., Dehais, F., Lesire, C., & Tessier, C. (2008). Resource as basic concecpts for authority sharing. In *Proceedings of Humans Operating Unmanned Systems (HOMOUS)*.

[147] Michaloski, J., Birla, S., & Yen, J. (2000). Software models for standardizing the human-machine interface connection to a machine controller. In *Proccedings of the World Automation Congress*.

[148] Miller, C., Funk, H., Wu, P., Goldman, R., Meisner, J., & Chapman, M. (2005). The playbook approach to adaptive automation. In *Human Factors and Ergonomics Society 49th Annual Meeting*, pp. 15–19.

[149] Miller, C. A. (2014). Delegation and transparency: Coordinating interactions so information exchange is no surprise. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, pp. 191–202, Springer.

[150] Miller, C. A., Goldman, R. P., Funk, H. B., Wu, P., & Pate, B. B. (2004). A playbook approach to variable autonomy control: Application for control of multiple, heterogeneous unmanned air vehicles. In *American Helicopter Society 60th Annual Forum*.

[151] Miller, C. A. & Parasuraman, R. (2007). Designing for flexible interaction between human and automation: Delegation interfaces for supervisory control. *Human Factors*, *49*(1), 57–75.

[152] Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors affecting trust in market research relationships. *Journal of Marketing*, *57*(1), 81–101.

[153] Morphew, M. E. & Wickens, C. D. (1998). Pilot performance and workload using traffic displays to support free flight. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

[154] Mouloua, M., Gilson, R., Kring, J., & Hancock, P. (2001). Workload, situation awareness, and teaming operations UAV/UCAV operations. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*.

[155] Muir, B. M. & Moray, N. (1996). Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460.

[156] Murphy, R. R. & Burke, J. L. (2010). The safe human-robot ratio. *Human-Robot Interactions in Future Military Operations, Human Factors in Defence. Ashgate*, pp. 31–49.

[157] Narayanan, S., Edala, N. R., Geist, J., Kumar, P. K., Ruff, H. A., Draper, M., & Haas, M. W. (1999). UMAST: A web-based architecture for modeling future uninhabited aerial vehicles. *Simulation: Transactions of The Society for Modeling and Simulation International*, *73, 1*, 29–39.

[158] Natarajan, G. (2001). Ground control stations for unmanned air vehicles. *Defence Science Journal*, *51 (3)*, 229–237.

[159] Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi:10.1037/0033-295X.84.3.231.

[160] Ollera, A. & Maza, I. (2007). *Multiple Heterogeneous Unmanned Aerial Vehicles*. Springer-Verlag Berlin Heidelberg 2007.

[161] Olson, W. A. & Sarter, N. B. (1999). Supporting informed consent in human-machine collaboration: The role of conflict type, time pressure, and display design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

[162] Olson, W. A. & Wuennenberg, M. G. (2001). Autonomoy based human-vehicle interface standards for remotely operated aircraft. In *20th Digital Avionics Systems Conference*, Daytona Beach, FL.

[163] Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2013). Human performance consequences of stages and levels of automation: An integratd meta-analysis. *Human Factors*, *56(3)*, 476–488.

[164] Oppermann, R. (1994). *Adaptive User Support: Ergonomic Design of Manual and Automatically Adaptable Software*. Lawrence Erlbaum Associates.

[165] Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Inproceeding of Unmanned Systems Technology XVI*, volume 9084.

[166] Page, J., Chi, T.-Z., & O'Neil, D. (2015). Self-organised swarms; a technology for aerospace. In *AIAC16: 16th Australian International Aerospace Congress*.

[167] Parasuraman, R. (1997). Human and automation: Use, misuse, disuse, abuse. *Journal of Human Factors*, *39*(2), 230–253.

[168] Parasuraman, R., Barnes, M., Cosenzo, K. A., & Mulgund, S. (2007). Adaptive automation for human-robot teaming in future command and control systems. *The International C2 Journal*, *1*(2), 43–68.

[169] Parasuraman, R. & Miller, C. (2006). *Delegation Interfaces for Human Supervision of Multiple Unmanned Vehicles: Theory, Experiments, and Practical Applications*, chapter 26, pp. 251–266.

[170] Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency". *The International Journal of Aviation Psychology*, *3(1)*(1), 1–23.

[171] Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *J. Human Factors and Ergonomics Soc.*, *38*(15), 665–679.

[172] Parasuraman, R. & Riley, V. (1997). Human and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

[173] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. A, Syst. Humans*, *30*(3), 286–297.

[174] Parasuraman, R., Sheridan, T. B., & 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), 140–160.

[175] Parasuraman, R. & Wickens, C. D. (2008). Human: Still vital after all these years of automation. *Human Factors*, *50*(3), 511–520.

[176] Perez-Rodriguez, D., Maza, I., Caballero, F., Scarlatti, D., Casado, E., & Ollero, A. (2012). A ground control station for a multi-uav surveillance system: Design and validation in field experiment. *Journal of Intelligent & Robotic Systems*, *69(1)*, 119–130.

[177] Phillips, E., Ososky, S., Grove, J., & Jentsch, F. (2011). From tools to teammates: Toward the development of appropriate mental models for intelligent robots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pp. 1491–1495.

[178] Piazza, T., Heller, H., & Fjeld, M. (2009). CERMIT: Co-located and remote collaborative system for emergency response management. In *SIGRAD*, p. 12.

[179] Purdy, E. M. (2008). The increasing role of robots in national security. *Defense AT&L*, *5*, 26–29.

[180] Quigley, M., Goodrich, M. A., & Beard, R. W. (2004). Semi-autonomous human-UAV interfaces for fixed-wing mini-UAVs. In *Proceedings of 2004 IEEE/RSJ International conference on Intelligent Robots and Systems*.

[181] Rao, A. S. & Georgeff, M. P. (1995). BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems*, pp. 312–319.

[182] Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, *13*(3), 257–266.

[183] Rasmussen, J. (1985). The role of hierarchical knowledge represent in decision and system management. *IEEE Transactions on Systems, Man, and Cybernetics*, *15*(2), 234–243.

[184] Rasmussen, J. (1986). *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. Elsevier Science Inc. New York, NY, USA.

[185] Rasmussen, J. (1998). Ecological interface design for complex systems: An example: SEAD - uav systems. *Technical report*, Air Force Research Laboratory, Human Effectiveness Directorate.

[186] Reenskaug, T. (May). The original MVC reports. Note. 1979.

[187] Regis, N., Dehais, F., Rachelson, E., Thooris, C., Pizziol, S., Causse, M., & Tessier, C. (2014). Formal detection of attention tunneling in human operator-automation interactions. *IEEE Transactions on Human-Machine Systems*, *44(3)*, 326–336.

[188] Reid (1988). *The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload*, chapter 8, pp. 185–214. Elsevier Science Publishers B.V. (North-Holland).

[189] Reid, G. B. & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology*, *52*(185–218).

[190] Reid, G. B., Potter, S. S., & Bressler, J. R. (1989). Subjective workload assessment technique (SWAT): A user's guide. *Interim report for period june 1986 to october 1988*, Human Reliability Associates.

[191] Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112.

[192] Riley, J. M. & Endsley, M. R. (2004). The hunt for situation awareness: Human-robot interaction in search and rescue. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, pp. 693–697.

[193] Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665.

[194] Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology An International Review*, *53*(1), 61–86. doi:10.1111/j.1464-0597.2004.00161.x.

[195] Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence: Teleoperators & Virtual Environments*, *11*(4), 335 – 351. doi:10.1162/105474602760204264.

[196] Rusnock, C. F. & Geiger, C. D. (2013). The impact of adaptive automation invoking thresholds on cognitive workload and situation awareness. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting*.

[197] Saget, S., Legras, F., & Coppin, G. (2008). Cooperative interface of a swarm of UAVs. In *Humans Operating Unmanned Systems*.

[198] Sanders, T. L., Wixon, T., Schafer, E., & Chen, J. Y. C. (2014). 2014 ieee international inter-disciplinary conference on cognitive methods in situation awareness and decision support (cogsima). In *The Influence of Modality and Transparency on Trust in Human Robot Interaction*.

[199] Santamaria, E., Segor, F., & Tchouchenkov, I. (2013). Rapid aerial mapping with multiple heterogeneous unmanned vehicles. In *Proceedings of the 10th International ISCRAM Conference*.

[200] Santamaria, E., Segor, F., Tchouchenkov, I., & Schönbein, R. (2013). Rapid aerial mapping with multiple heterogeneous unmanned vehicles. *International Journal on Advances in Systems and Measurements*, *6*, 384–393.

[201] Sarter, N. B. & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, *1*(1), 45–57. doi:10.1207/s15327108ijap0101_4.

[202] Sarter, N. B. & Woods, D. D. (1995). How in the world did we ever get into that mode? mode error and awareness in supervisory control. *Human Factors*, *37*(1), 5–19.

[203] Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). *Handbook of Human Factors and Ergonomic, Second Edition*, chapter Automation Surprises, pp. 1–25. Wiley.

[204] Scerbo, M. (2001). Adaptive automation. In W. Karwowski (Ed.), *International Encyclopedia of Ergonomics and Human Factors*.

[205] Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y. C., & Hancock, P. A. (2014). A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction. *Technical report*, Army Research Laboratory.

[206] Scholtz, J. C., Antonishek, B., & Young, J. D. (2005). Implementation of a situation awareness assessment tool for evaluation of human-robot interfaces. *IEEE Transactions on Systems, Man, and Cybernetics - Part A; Systems and Humans*, *35(4)*, 450–459.

[207] Sellner, B., Heger, F. W., Hiatt, L. M., Simmons, R., & Singh, S. (2006). Coordinated multiagent teams and sliding autonomy for large-scale assembly. In *Proceedings of the IEEE*, volume 94.

[208] Seppelt, B. D. & Lee, J. D. (2007). Making adaptive cruise control ACC limits visible. *International Journal of Human-Computer Studies*, *65*(3), 192–205.

[209] Sheridan, T. B. (1988). *Handbook of Human-Computer Interaction*, chapter Task Allocation and Supervisory Control. Elsevier Science Publishers.

[210] Sheridan, T. B. (2002). *Human and Automation: System Design and Research Issues*. John Wiley and Sons, Inc. New York, NY, USA.

[211] Sheridan, T. B. & Verplanck, W. L. (1978). Human and computer control of undersea teleoperators. *Technical report*, Massachusetts Institute of Technology.

[212] Shively, R., Battiste, V., Matsumoto, J., Pepition, D., Bortolussi, M., & Hart, S. G. (1987). In flight evaluation of pilot workload measures for rotorcraft research. In *In Proceedings of the Fourth Symposium on Aviation Psychology*, pp. 637–643, Department of Aviation, Ohio State University, Columbus, OH.

[213] Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, *43*(9), 63–65.

[214] Singh, P. W. (2010). War of the robots. *Scientific American*, *303*, 56–63.

[215] Taylor, R. M. (1990). Situation awareness rating technique (SART). In *Situation Awareness in Aerospace Operations*, pp. 3/1–3/17, Neuilly-Sur-Seine, France: NATO – Advisory Group for Aerospace Research and Development.

[216] Thales, A. (????). Pty ltd NESACC EUROCAT-X system.

[217] Thornton, J., Grace, D., Spillard, C., Konefal, T., & Tozer, T. C. (2001). Broadband communications via high-altitude platofmrs: the european helinet programme. *Electronics & Communication Engineering Journal*, *13*(3), 138–144.

[218] Tsang, P. S. & Velaquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381.

[219] Tso, K. S., Tharp, G. K., Zhang, W., & Tai, A. T. (1999). A multi-agent operator interface for unmanned aerial vehicles. In *Digital Avionics Systems Conference*, volume 2, pp. 6.A.4–1 – 6.A.4–8, St Louis, MO.

[220] Uggirala, A., Gramopadhye, A. K., Melloy, B. J., & Toler, J. E. (2004). Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, *34*, 175–186.

[221] van Ginkel, H. T., de Vries, M. F., Koeners, J. J., & Theunissen, E. (2006). Flexible authority allocation in unmanned aerial vehicles. In *Proccedings of the Human Factors and Ergonomics Society 50th Annual Meeting*.

[222] Venturino, M., Hamilton, W. L., & Dvorchak, S. R. (1990). Performance-based measures of merit for tactical situation awareness. In *Situation Awareness in Aerospace Operations*, pp. 4/1–4/5, Neuilly-Sur-Seine, France: NATO – Advisory Group for Aerospace Research and Development.

[223] Vicente, K. J. (1999). *Cognitive Work Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

[224] Vicente, K. J. & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(4), 589–606.

[225] Villaren, T., Madier, C., Legras, F., Leal, A., Kovàcs, B., & Coppin, G. (2010). Toward a method for context-dependent allocation of functions. In *Proceedings of the 2nd Conference on Human Operating Unmanned Systems (HUMOUS'10)*.

[226] Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*, *51*(3), 281–291.

[227] Warren, M., Mejias, L., Kok, J., Yang, X., Gonzalez, L. F., & Upcroft, B. (2015). An automated emergency lanland system for fixed-wing aircraft: Planning and control.

[228] Wheeler, M., Schrick, B., Whitacre, W., Campbell, M., Rysdyk, R., & Wise, R. (2006). Cooperative tracking of multiple targets by a team of autonomous UAVs. In *25th Digital Avionics Systems Conference*.

[229] Whitehead, A. N. (1927). *Symbolism, Its Meaning and Effect*. New York: Macmillan.

[230] Wickens, C. D. (1987). Information processing, decision making, and cognition. In Amsterdam:Elsevier (Ed.), *Cognitive Engineering in the Design of Human-Computer Interaction and Expert Systems*.

[231] Wiener, E. L. (1988). *Human factors in aviation*, chapter Cockpit Automation, pp. 433–461. San Diego, CA, US: Academic Press.

[232] Wierwille, W. W. & Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceedings of the 27th Annual Meeting of the Human Factors Society*, pp. 129–133, Santa Monica: Human Factors Ergonomics Society.

[233] Williams, R., Singer, B., Feitshans, G., Rowe, A., & Burns, R. S. (2005). A prototype UAV control station interface for automated aerial refuleing. *Technical report*, Air Force Research Laboratory.

[234] Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*.

Résumé

L'intérêt d'utiliser des groupes de plateformes aériennes autonomes et hétérogènes, pour des missions telles que le Search and Rescue ou la surveillance de l'environnement, est désormais reconnu et avéré. Cependant, les opérations engageant plusieurs drones sont toujours limitées par le nombre d'opérateurs nécessaires à leur contrôle, et il est nécessaire d'inverser le ratio nombre d'opérateurs / nombre de drones. L'objectif de cette thèse est de contribuer à l'amélioration des performances cognitives de tels opérateurs, en traitant de la charge cognitive (CW), la conscience de situation (SA) et de la confiance dans les automatismes.

Les contributions de la thèse concernent la gestion d'un modèle fonctionnel du drone et la mise en place de différentes niveaux de transparence informationnelle qui étendent l'approche traditionnelle PAC (Presentation-Abstraction-Control) pour la conception d'interface homme-machine. Le résultat de cette approche est une amélioration de la gestion du niveau d'autonomie système.

Le modèle fonctionnel du drone est divisé en deux parties concernant respectivement la nature de la fonction et le degré d'agrégation de l'information. La présentation de ces informations est structurée de façon hiérarchique de façon à minimiser l'effort cognitif d'un opérateur devant contrôler plusieurs drones simultanément.

Les concepts proposés dans la thèse ont été validés à travers différentes expérimentations, visant respectivement à mesurer l'effet de la nouvelle représentation fonctionnelle ainsi que du niveau d'autonomie sur la charge cognitive, sur la confiance. Les résultats montrent ainsi une amélioration en termes de charge cognitive, en termes de conscience de situation et en termes de temps de réponse à des événements inattendus.

Mots-clés : Interaction homme-machine, Drones, Coopération homme-machine, Transparence informationnelle

Abstract

The benefits of using multiple heterogeneous Unmanned Aerial Vehicles (UAVs) to perform tasks such as search and rescue and wildlife monitoring are now well recognised. However, multiple UAV operation is currently limited by the current Operator to UAV management ratio of n:1, where more than one human operator is needed to manage one UAV. Effective management of multiple heterogeneous UAVs can be achieved by inverting the management ratio to 1:n.

One challenge that arose is the human operator's cognitive performance, and improving the performance is the goal of this thesis. This thesis addresses the three main aspects of the cognitive challenges associated with one operator managing multiple UAVs: Operator Cognitive Workload (CW), Situation Awareness (SA), and Automation Trust.

Novel contributions are made through increasing the UAV's functional subsystem and autonomy capability transparency by leveraging the traditional the Presentation-Abstraction-Control (PAC) design paradigm and proposing an improved interface implementation. This results in an increase in the systems' autonomy transparency.

A Functional Capability Framework (FCF) is proposed to categorise a UAV's functional subsystem capabilities, and increase the UAV's autonomy capability transparency. These are tested by communicating the UAV's autonomy status to the human operator.

The FCF organises the UAV's functional subsystems into two dimensions; its nature of functionality and it's level of information aggregation abstraction. The framework presents the UAV subsystem information in a categorical and hierarchical manner such that the cognitive performance of the human operator is increased when managing multiple UAVs.

The communication of the UAV's autonomy capability information is crucial for the human operator to understand their assets, improving their cognitive performance when managing multiple heterogeneous UAVs. This was achieved by creating a graphical and textual representation of the UAV information, where a textual message box with messages stating the sender, its decisions and intentions is displayed.

The improved PAC approach consists of the FCF and autonomy transparency amd it was verified with three experiments: 1) Evaluation of the operator's cognitive performance when using the FCF, 2) Evaluation of the operator's cognitive performance through graphical representation of the UAV's autonomy, and 3) Evaluation of the operator's cognitive performance through natural language information exchange. Results collected from these experiments yielded an improvement to the operators' cognitive performance when using the improved PAC model compared to the traditional PAC approach. This suggests that by increasing a UAV's capability transparency in a multiple heterogeneous UAV system, the human operator demonstrated a reduced CW, an improved SA, and an improved reaction time.

Keywords : Human-MAchine Interaction, UAV, Human-Machine cooperation, Information transparency

TELECOM Bretagne

n° d'ordre : 2016telb0423

Télécom Bretagne

Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3

Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00