UNIVERSITY OF BIRMINGHAM

# COPING WITH UNRELIABLE AUTOMATION: CONTENT/FORMAT/FORM IN THE DESIGN OF HUMAN-AUTOMATION SYSTEMS

by

**NATAN SORIN MORAR**

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

School of Engineering
Electronic, Electrical and Systems Engineering
College of Engineering and Physical Sciences
University of Birmingham
April 2018

# ABSTRACT

The research presented in this thesis is funded by the European Union and addresses the relationship between people and automated decision support in the context of Traffic Management. Given that automation might not always be 100% reliable, the first research question to be addressed is what effect does automation reliability have on human decision making? User trials contribute to addressing the question of, how can user interfaces be designed to cope with the effects of different levels of automation reliability. The thesis is developed around the concept of Content (the users' information requirements), Format (the paradigm of interaction and communication protocols) and Form (how information is presented to the users).

Results demonstrate that, even in the absence of explicit feedback, users are sensitive to automation reliability and can adapt their information search and decision making strategies accordingly. The user's decision on whether or not to seek further information cannot be attributed only to information availability or accessing costs, but the visual appearance of the user interface can have a higher influence on user behaviour. These observations and conclusions led to the refinement of the Content/Format/Form concept to a broader sociotechnical design framework.

# ACKNOWLEDGEMENTS

Thanks to Chris Baber, my supervisor, for his infinite patience, encouragement and support that came in addition to his invaluable guidance.

Thanks to my wife for her understanding of my reduced interaction, interspersed with thoughts relating to Human-Automation Systems, whilst writing the Thesis.

Thanks to my parents for trying very hard to teach me the lessons of responsibility, acceptance and determination, which helped me to complete this work.

Last, but not least, thanks to SPEEDD for the interesting use-cases and knowledge gathered and the project partners (Alexander Artikis, George Paliouras, Anastasios Skarlatidis, Fabiana Fournier, Alexander Kofman, Inna Skarbovsky, Marius Schmitt, John Lygeros, Chithrupa Ramesh, Assaf Schuster, Daniel Keren, Moshe Gabel, Sandra Starke, Andrew Howes, Neil Cooke, Alain Y. Kibangou, Carlos Canudas, Federica Garin, Ivo Correia – I do apologise if I left anybody out) for their professionalism, cooperation and hospitality. Thanks also to the EU Commission for the financial support of the project.

# CONTRIBUTIONS

The work undergone in this thesis is part of a larger European Research Project (SPEEDD). However, this thesis presents the author's own original work. Where previously published studies with multiple authors are presented (Chapter 3 – Chapter 5), the work (design of the study, running the study, analysing the data, writing the report/paper) has been done by the main author (Natan Morar) and input has been received from the other authors. In the case of the experiment in Chapter 3, help has been received from Sandra Starke for running the study with participants and guidance has been received for doing the analysis.

# TABLE OF CONTENTS

# TABLE OF FIGURES

# TABLE OF TABLES

# LIST OF ABBREVIATIONS

AoF – Allocation of Function

CCTV – Closed-Circuit Television

CWA – Cognitive Work Analysis

CFF – Content/Format/Form

DM – Decision Making

EID – Ecological Interface Design

HA – Human-Automation

HAS – Human-Automation Systems

HTA – Hierarchical Task Analysis

ID – Identification

SME – Subject Matter Experts

SUS – System Usability Scale

TM – Traffic Management

UI – User Interface

VMS – Variable Message Signs

# CHAPTER 1 INTRODUCTION

This chapter outlines the motivation of the work undergone as part of the PhD. It shows how it fits in with previous research and, more importantly, how it advances knowledge in the area of Human-Automation Systems (HAS). The PhD is centred around investigating the issues related to the communication between humans and computers in the context of complex human-automation systems. The state of the research done prior to this work is presented, followed by the advancements which this PhD brings to the state-of-the-art and the contributions to the body of knowledge in the area of Human-Automation systems.

## 1.1 Research Questions

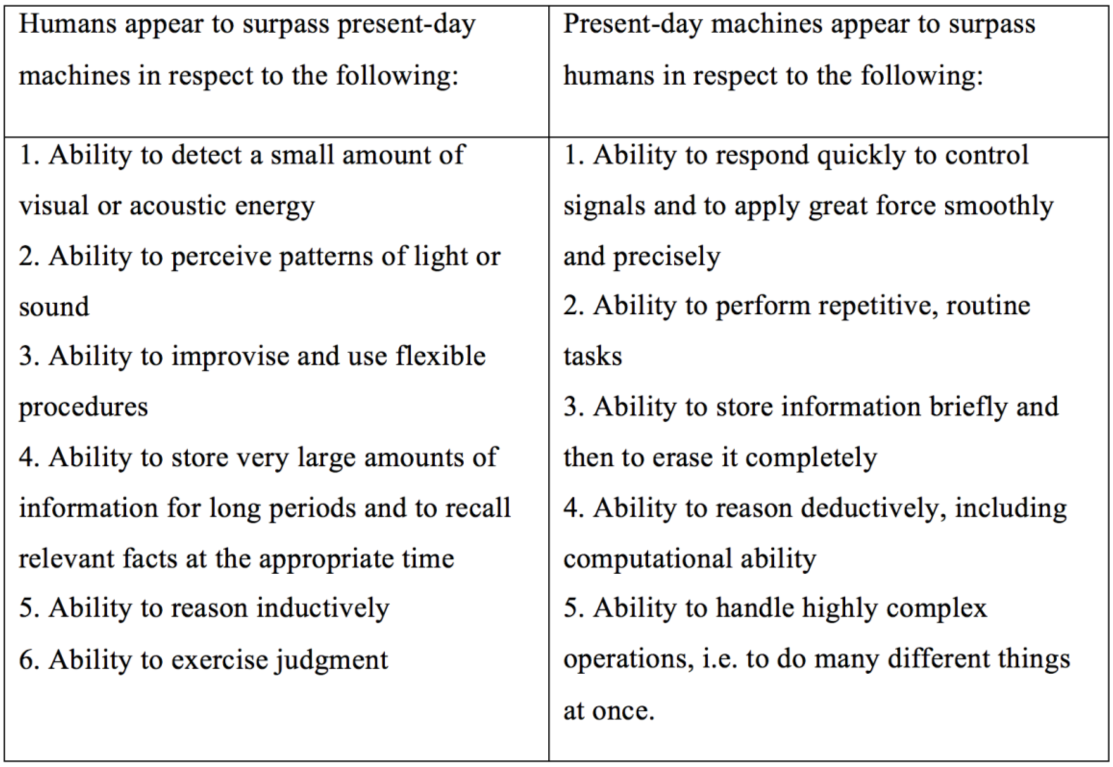
1. What effects does automation reliability have on human decision making?
2. How can we design user interfaces to help users cope with these effects?

## 1.2 Motivation

Automation has seen a tremendous increase in adoption over the last few years (Onnasch et al., 2014). Bainbridge, (1983), Greengard (2009), Parasuraman and Riley (1997), to name a few, have illustrated the pitfalls of extreme automation. Computers operate based on the model of the world that has been programmed into them. However, one must remember that models are an approximation and the model is not the world with which the computers interact. In many cases, the model is good enough to predict the behaviour of the system that automation is aiming to control. Nevertheless, there are countless variables that could interfere with the proper functioning of that system. Even though the computer may be able to deal with some of those, having a finite number of inputs (sensors), it cannot yet respond to all, even assuming some sort of advanced on-line learning and prediction capabilities. Despite that, in the case of Big Data, where capabilities of automation are largely increased, the ability to draw knowledge and respond to novel situations is still limited due to the challenges presented by data capture, availability, analysis and visualisation (Ahrens et al., 2011). This is, perhaps, largely due to the fact that computers are increasingly "*CPU-heavy but I/O-poor*" (Philip Chen and Zhang, 2014). This is to say that computing speeds have massively increased, while the modes of interaction with it have remained relatively the same for the past few decades (screen, mouse, keyboard). This further poses the problem of how the generated knowledge can be acted on and effectively communicated to humans.

Furthermore, it seems like there is a debate of who or what would be best suited for performing specific tasks between humans and automation. Probably, these are the same kind of arguments that Fitts was faced with back in 1951 when he came up with the HABA-MABA list (humans are better at, machines are better at – see Table 1.1) (*Human engineering for an effective air-navigation and traffic-control system*, 1951). The important issue is not a measuring of force between humans and computers, but rather the question of how the two would best complement each other (Bainbridge, 1983; Hoc, 2000; Hoc and Debernard, 2002; Parasuraman and Riley, 1997). How could they share their strengths and put them to good use at solving complex problems together? By assuming that computers are the solution to every problem, one immediately excludes the possibility of a human being extending the computer's reach in the real world, possibly acting like a sensor or an actuator for the entire system. As part of this PhD, the DIR-CE (Grenoble) Traffic Management control room has been studied. During observations, we witnessed the automatic obstacle detection system triggering multiple alerts. It turned out that the vehicle stopped on the side of the road was performing scheduled maintenance work. Even though the traffic operators knew about this in advance, there was no means by which they could inform the automated system.

Equally, by taking the opposite stance, that humans would outperform automation in any situation, one is not considering the potential benefits of the computer augmenting human cognition by offering the ability to search large databases, quickly implement complex algorithms and detect patterns from multiple large data streams. So, rather than being drawn into debates of whether humans are better than computers at doing specific jobs, this present work is looking to identify how humans and automation could work together in order to make use of the strengths of both.



| Humans appear to surpass present-day machines in respect to the following: | Present-day machines appear to surpass humans in respect to the following: |
|---|---|
| 1. Ability to detect a small amount of visual or acoustic energy<br>2. Ability to perceive patterns of light or sound<br>3. Ability to improvise and use flexible procedures<br>4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time<br>5. Ability to reason inductively<br>6. Ability to exercise judgment | 1. Ability to respond quickly to control signals and to apply great force smoothly and precisely<br>2. Ability to perform repetitive, routine tasks<br>3. Ability to store information briefly and then to erase it completely<br>4. Ability to reason deductively, including computational ability<br>5. Ability to handle highly complex operations, i.e. to do many different things at once. |

**Table 1.1 - Fitts' List (*Human engineering for an effective air-navigation and traffic-control system*, 1951)**

If we were to reiterate Fitts' List today, in the age of Big Data, perhaps it would look a bit different. For instance, computers have surpassed humans in the amount of data they can store and access, as well as the number of variables they can handle at any one time (virtually unlimited for computers vs around four for humans (Halford et al., 2005)). Moreover, it can be argued that, thanks to Big Data, computers have also gained the ability to reason inductively. However, humans still hold an advantage over computers when encountering novel situations (Lee, 2008), thanks to their ability to adapt and improvise. So, while the situation has changed since Fitts first formulated the famous HABA-MABA list, humans still have a place alongside complex automation, their joint work proving beneficial in the face of complex and uncertain scenarios (Parasuraman and Wickens, 2008). Therefore, it is important to investigate how this collaboration between humans and automation can be made more effective and efficient.

## 1.3 Interacting with Imperfect Automation

Researchers have investigated the implications of having humans work alongside automation (Hancock and Scallen, 1996; Hoc, 2001; Hollnagel, 1987; Lee and See, 2004; Parasuraman and Riley, 1997). The potential benefits that they have uncovered stand as motivation for the work undertaken in this PhD. When considering the design of a human-automation (HA) system, the immediate concern is with how the interaction between the two agents should be managed. In other words, how should the two agents exchange information? In order to answer such a question, one must consider how one chooses to define such a system. A human-automation system can be seen as construct comprising of two independent agents working on separate tasks, for the purpose of solving a shared problem. An example of such a view is the Mueller et al. (2011) Visual Analytics Process (see Figure 1.1). Alternatively, a human-automation system can be regarded as a unity, where the two agents are bound together, describing some sort of symbiotic relationship, as Licklider (1960) envisaged. Licklider's idea of man-computer symbiosis is somewhat loosely defined, but could potentially fit the former view, where the human and the computer work closely together as separate agents on a shared task, provided that communication (i.e. information sharing) is established with ease. Of course, when Licklider put forth his idea (1960s), it took a computer scientist working for a considerably wealthy institution to operate a computer. However, today, the availability of and accessibility to computers is no longer an issue and research in this direction is therefore, not only possible but also due. So then, the question that remains to be answered is how this collaboration between the human and the computer in human-automation system should be designed in order to, so to speak, blur the lines between the two agents, having them operate as an effective unit.

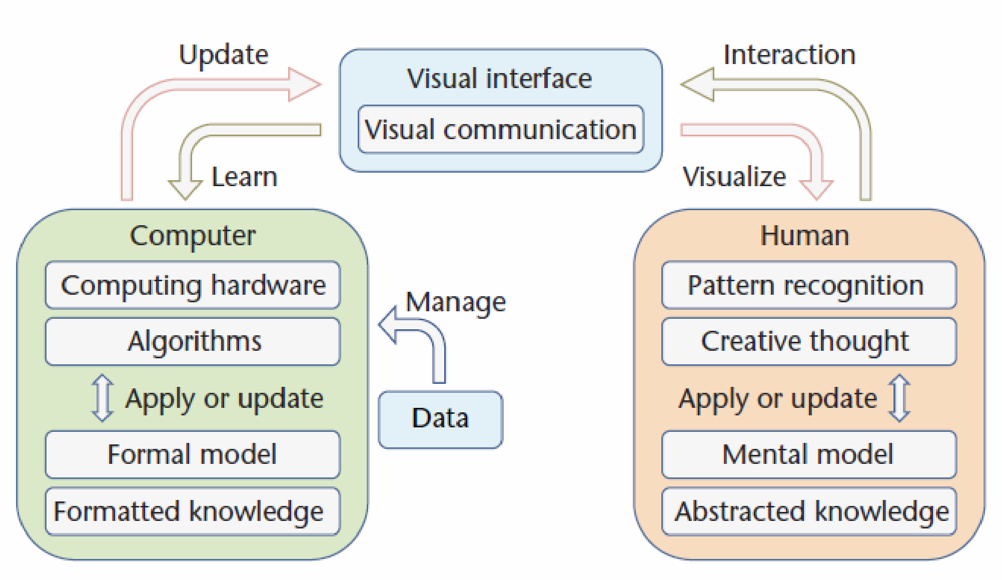


**Figure 1.1 - The Visual Analytics Process (Mueller et al., 2011)**

The key points that require investigation in view of the correct operation of the Human-Automation system are: communication, task allocation, authority and responsibility. In this context, authority refers to which of the two agents ‘has the final say’ with regard to a specific action. Assuming that computers and humans both work in parallel at solving a particular problem, there may be some instances where the two generate, not only different, but conflicting solutions. This is not unlikely, as they rely on different information sources, may have different underlying assumptions about the world they interact with and different models of how the system they are trying to control works (Morar et al., 2015a; CHAPTER 3). The question of how one could settle these conflicts arises. Flemisch et al. (2011) see this as a matter of balancing the key factors of authority of control, ability to respond, and responsibility (see Figure 1.5). Perhaps the issue of ability can be settled by following the guidelines of the Fitts’ List (see Table 1.1), with some adaptations depending on the work domain in question and the state of the art. Through Figure 1.2, Flemisch et al. (2011) propose that authority for making a decision should not be higher than the ability to make that decision, that responsibility should be directly proportional to the level of control a particular agent has and that a higher degree of control should be given to the agent with higher ability and authority for a given task.

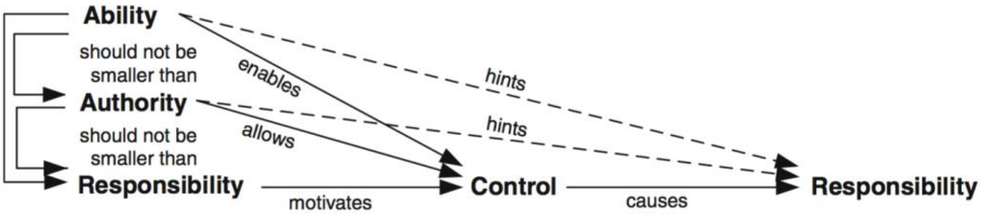


**Figure 1.2 - Relations between ability, authority, control and responsibility (Flemisch et al., 2011)**

However, one prominent issue in work domains which have recently undergone an increase in automation is the fact that, while the authority of control shifted from humans to the computer, responsibility for correct action remains still in the hands of the human agent (Lyons and Stokes, 2011). Moreover, the outputs of automation are rarely associated with explanations as to how they were computed, thus justifying them to the human operator. This puts the human in a very difficult situation. He may be forced into taking responsibility for an action performed by what is, essentially, a ‘black-box’ system. It may be argued, however, that even though this is the case, the solution would be to give the human operator the ‘final say’, that is, the ultimate authority of either following the computer’s output or his own derived answer. This issue is explored in CHAPTER 4. Nevertheless, the human may not be given the means to form a judgement of the situation or to appraise the computerised output, as a result of being deprived of such key factors as, for example, contextual information (Bainbridge, 1983).

Inagaki (2003) suggests that authority in human-automation systems should always sit with the human. However, one cannot assume that giving the human the final authority over automation would be the best approach in all situations. Take, for example, the recent Germanwings disaster (“Germanwings crash,” 2015). Is it safe then to assume that the human will make the ‘right’ decision, whatever ‘right’ is defined as, irrespective of the situation? This example suggest that this may not be the case. However, the assumption that the computer is able to always make the correct decision is just as moot for many reasons, discussed in the first part of this section.

Therefore, in order to allow for shifts in authority and thus, responsibility, one must provide for a clear understanding of the momentary situation to be acted upon and also for an unobstructed exchange of information, such as the appearance of unexpected factors, working assumptions and possible steps to arriving to a particular solution.

## 1.4 Allocation of Function

Allocation of Function in human-automation systems is closely linked to the issue of ability, control and responsibility. Allocation of Function (AoF) is a Human Factors method which describes the ‘who-does-what’ of a multi-agent system (Hancock and Scallen, 1996). More specifically, considering a list of functions attributed to a system, AoF specifies to which of the agents the control over of each function is assigned.

Over the past decades several guiding criteria for task allocation have been put forth. One of the most famous being Fitts list or “HABA-MABA” (Humans Are Better At – Machines Are Better At) (*Human engineering for an effective air-navigation and traffic-control system*, 1951) (see  Table 1.1). However, this criteria has been criticised as it is considered to be useful only to the extent that humans are compared with automation (Helander et al., 1997) and does not offer any resolution in cases where both humans can computers can perform at a similar standard.

A more modern criteria for task allocation can be seen in Figure 1.3. Some tasks (such as analysing large databases, polling sensors, etc.) are evidently better performed by automation, other tasks (such as formulating hypotheses, insight gaining, etc.) better match human capabilities (Morar et al., 2015a). Figure 1.3 also illustrates the existence of a task space where both agents would perform equally well. This further complicates the allocation of function issue. The question that arises is to which of the agents the functions associated with this task space should be assigned.

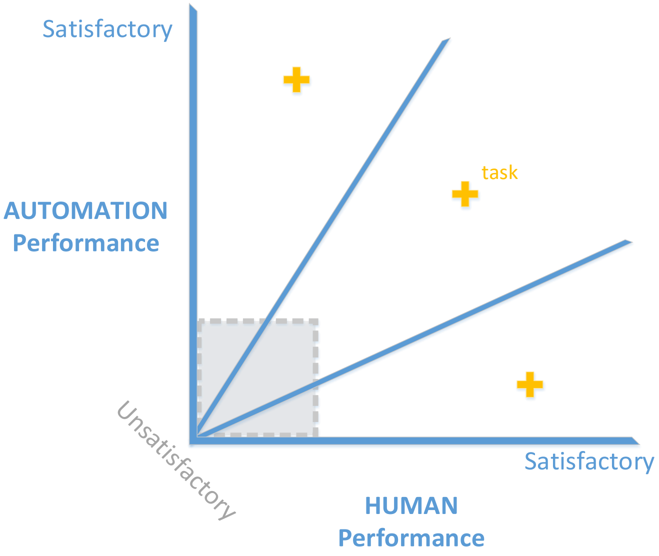


**Figure 1.3 - Criteria for Task Allocation (adapted from Sheridan (2000))**

As a potential solution to this issue, some researchers have proposed the idea of task trading and task sharing (Figure 1.4) (Sheridan, 2002). Task trading means that during process of solving a task, control shifts from the human to computer (or the other way around). Now, this can be triggered automatically (as a result of an increase in workload, by the occurrence of an unexpected situation, etc.), or simply manually (as a result of one of the agents explicitly requesting control over, or help with a task) – relates more to top (automation performance – highly satisfactory, human performance - highly unsatisfactory) and bottom areas (automation performance – highly unsatisfactory, human performance - highly satisfactory) of Figure 1.3.

Task sharing refers to the scenario where the computer (automation) and the user work on the same task simultaneously. This can either happen when the task can be split into subtasks that do not require sequential solving or when the two agents collaborate continuously on a task. Sharing would probably work better in area in the middle of Figure 1.3 (where automation and humans perform equally well at solving the tasks).

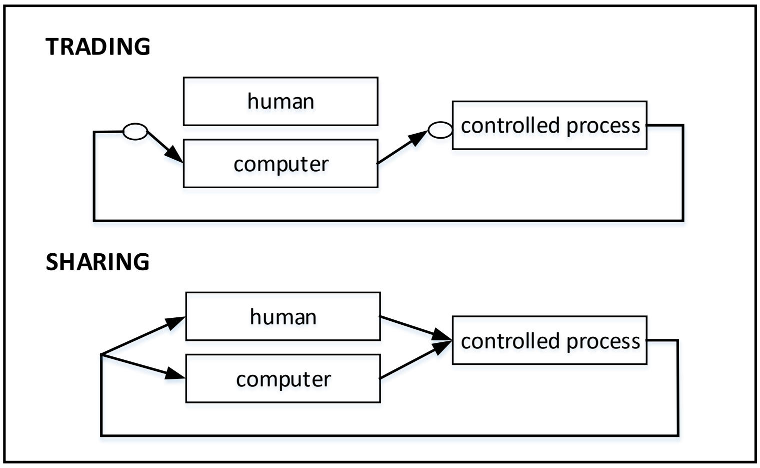


**Figure 1.4 - Trading and sharing tasks (reproduced from Sheridan) [8, Figure 3.5]**

Past research (Byrne and Parasuraman, 1996; Hancock and Scallen, 1996; Hoc, 2000; Hoc and Debernard, 2002; Hollnagel, 2001, 1987; Johnson et al., 2014; Lee and Moray, 1992; Parasuraman et al., 1996; Parasuraman and Riley, 1997) addressed the possibility of implementing an Allocation of Function that changes. This variety of AoF, where the control would dynamically shift from one agent to another (i.e. computer to human and vice-versa) was named Dynamic (or, Adaptive) Allocation of Function. The implications of adopting this dynamic AoF (DAoF) in the design of human-automation systems have shown the approach to be a welcome alternative to complex systems design. The use of a DAoF has shown to reduce complacency and fatigue, increase situation awareness, lead to better management of trust, and to increase the overall reliability of the human-computer system (Johnson et al., 2014, 2014; Lee, 2008; Lee and See, 2004; Lee and Moray, 1992; Parasuraman et al., 2009, 1996; Parasuraman and Manzey, 2010). It has also been found that adopting this dynamic shift in task control can lead an increase in spotting automation errors and overall to better operator skill retention (Byrne and Parasuraman, 1996; Greef et al., 2009; Johnson et al., 2014; Parasuraman et al., 2009, 1996). Moreover, "cooperation between human operators and autonomous machines in dynamic situations implies need for dynamic allocation of activities between the agents" (Hoc and Debernard, 2002).

Adaptive allocation of function can be employed either explicitly (when operators specifically delegate tasks to the computer) or implicitly (when tasks are delegated automatically based on metrics such as workload) (Vanderhaegen et al., 1994). It has been investigated how this shift in task control could be triggered by psychophysiological cues (Rani et al., 2007; Sims et al., 2002; Vanderhaegen et al., 1994). However, this tends to only one side of the issue, that is, increases in mental workload or changes in the human's psychophysiological state. This achieves an implicit communication in one direction, from the human (user) to the computer (automation). However, one can imagine tasks being delegated from the computer to the human on data bus overload or when sensor errors are encountered, for example.

Although Vanderhaegen et al. (1994) found a clear performance increase in the implicit allocation of function mode as compared to the explicit mode, the human operators were less appreciative of the implicit mode. "They reported that they were very anxious to keep control over the entire situation" (Hoc and Debernard, 2002). This, perhaps, was because that although control and authority shifted to the computer in situations of high workload, the responsibility for correct operation of the entire system remained with the human. This creates a problem, as the human is held accountable for actions of the system not under his control. For example, the 1996 Washington Metro accident happened because of inappropriate management of authority (Parasuraman and Riley, 1997). The automated speed control system received erroneous data from track speed sensors that were covered with ice and snow and even though the operator became aware of the danger and requested manual control, he was refused it, resulting in a crash with another train at the end of the line.

Hoc and Debernard (2002) propose that a solution would be for the human to hold Authority in delegating tasks and for the computer to intervene in case it finds a potential error with the solution given by the operator. They call this Dynamic Function Delegation. This seems like a good compromise between the two extremes of full automation and no automation as it aims to solve the confusion of “who’s in charge”. This would, perhaps, best fit the augmentative form of cooperation, in which the human is the task coordinator, and, therefore, has authority of the system and responsibility for its proper functioning (Schmidt et al., 1991). However, this approach is limited because it presupposes that responsibility will always sit with the human, even though it was found that human operators tend to see themselves less responsible for the tasks performed by the automation (Hoc and Lemoine, 1998) (see also CHAPTER 4). Moreover, while there is some value in having authority always sitting with the human (Inagaki, 2003), it can cannot be treated as independent from responsibility (Dekker, 2002; Flemisch et al., 2011; Woods, 1985; Woods and Cook, 2002). This issue becomes more apparent when dealing with examples of the nature of the Germanwings disaster (Willsher, 2015). Perhaps, a resolution of this problem could be brought about by the sharing of authority and responsibility between the human and automation.

AoF relates directly to the authority of control. One can imagine the responsibility for control actions taken to follow the dynamic shifts in function allocation. In this way, most responsibility for an action sits with the agent that holds more control over it, thus being coherent with the Flemisch et al. (2011) guidelines (see Figure 1.5). The experiment presented in CHAPTER 4 looks at how humans understand responsibility when faced with imperfect automation. In trying to balance responsibility, authority, ability and control when designing Human-Automation systems, one encounters the following challenges: how can one evaluate which agent is more able and, thus, more suited to control a particular situation; should this evaluation be performed by the human or by automation; and, finally, what happens when automation performs unreliably in situation where it has the most control? For the scope of this thesis, we will look at how humans interact with imperfect automation. Specifically, we will investigate whether humans can accurately judge automation reliability in uncertain situations and in the absence of performance feedback (CHAPTER 3; CHAPTER 4; CHAPTER 5), thus mimicking the probabilistic nature of the world we live in. Provided that they can, this would suggest that Inagaki's (2003) approach of giving the humans the final authority might be more desirable than having authority dynamically shift between the agents. Moreover, this would also mean that humans will rely on automation only when appropriate, i.e. only in cases where it is highly reliable.

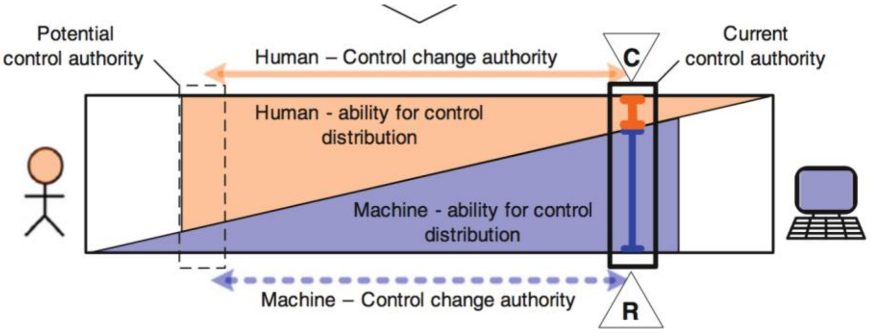

Potential
control authority
Human – Control change authority
C
Current
control authority
Human - ability for control
distribution
Machine - ability for control
distribution
Machine – Control change authority
R

**Figure 1.5 - Balance of responsibility and authority of control (Flemisch et al., 2011)**

## 1.5 Trust and Reliance on Automation

Trust has been considered as one of the main elements that determine reliance on automation. This is shown by the large body of literature that examined trust in the context of automation (Dzindolet et al., 2003; Lee and See, 2004; Lee and Moray, 1992; Parasuraman and Riley, 1997). It has been suggested that the key to understanding automation misuse and disuse might rely on understanding which factors cause over-trust and under-trust (Lee, 2008; Lee and See, 2004).

Trust is defined as the willingness to be made vulnerable to another party’s actions with the expectation of positive outcomes (Mayer et al., 1995). Any HA system comprises at least two agents. For the system as a whole to achieve its goals, each agent needs to perform its tasks correctly. Moreover, for the system to function without interruption or replication of effort, each agent needs to trust that the other will perform its tasks as expected. Even though many factors have been found to influence reliance (workload, boredom, expertise, situation awareness, etc.) (Cummings et al., 2013; Lee and Moray, 1994; Masalonis et al., 1999; Parasuraman and Manzey, 2010; Sheridan, 2002), trust is seen as a prime factor (Parasuraman and Riley, 1997). Most accidents are considered to have been caused by either under- or over-reliance on automation (Greengard, 2009) both of which are forms of misuse (Parasuraman and Riley, 1997).

Many terms have been used interchangeably with trust: cooperation, confidence and predictability, to name a few (Mayer et al., 1995). Ajzen and Fishbein (1980) developed a framework that can help distinguish between the different terms. They proposed that beliefs be considered as the information base. Beliefs are influenced by the person's past experiences and the information available at the time of making the decision. Beliefs, in turn, influence attitudes and attitudes form intentions. Intentions, ultimately determine behaviours. As a parallel to this framework (Lee and See, 2004), proposed that reliance on automation is seen as the behaviour and trust as the attitude (Figure 1.6). However, there are many other factors that can affect reliance, such as self-confidence, workload, boredom, complacency, etc. (Cummings et al., 2013; Lee and Moray, 1994; Masalonis et al., 1999; Parasuraman and Manzey, 2010).

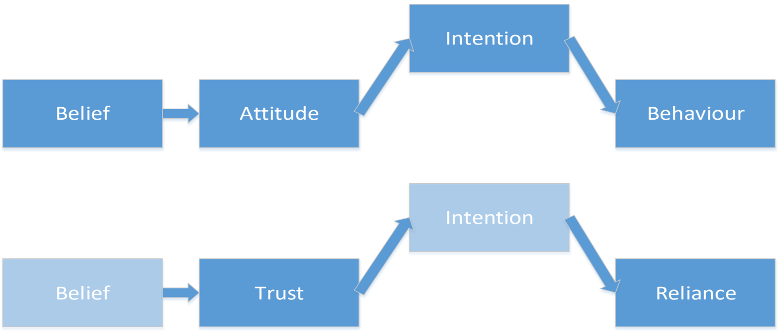


**Figure 1.6 - Trust as Attitude**

In this work, I do not talk about trust directly, but I am more interested in reliance for number of reasons. Firstly, reliance is the key factor which ultimately shows whether and how often automation is used. Secondly, measuring subjective trust through questionnaires proves rather distracting, invasive and disruptive to users (Kaniarasu et al., 2012). Finally, users' confidence in automated responses can be inferred from how they choose to interact with the system (i.e. do users follow the computer advice or not, do they look for more information) and this translates to reliance.

While trust in automation is an important issue in the design of HA systems, it is a research field in itself and its study is far beyond the scope of this thesis. In their review of trust in automation, Hoff and Bashir (2015) propose that the relationship between trust and reliance is not directly linked, but that other factors influence the strength of this relationship, namely: complexity of automation, novelty of situation, ability to compare automated performance to manual and the operator's degree of decisional freedom (see Figure 1.7). They suggest that trust plays a role in human behaviour in uncertain situations, when automation is more complex, when users cannot evaluate automation reliability, when they are not given the opportunity to formulate their own judgement and when their authority is lower than that of automation. In other words, humans are 'forced' to rely on automation when they do not have enough information to veto it. Hoff and Bashir (2015) call these factors *environmental conditions,* however it can be argued that all, except 'novelty of situation', are inherent in the design of the HA system (complexity of automation – automation design; operator's ability to compare automated performance to manual and operator's degree of decisional freedom – UI design). In order to be able to use reliance as a proxy for trust, work has been done in the experiments in order to ensure a strong relationship between trust and reliance. Specifically, i) users have been given the possibility to make their own decisions (access to data), ii) they have been given the authority over the final decision, iii) novelty of situation and iv) complexity of automation are kept constant.

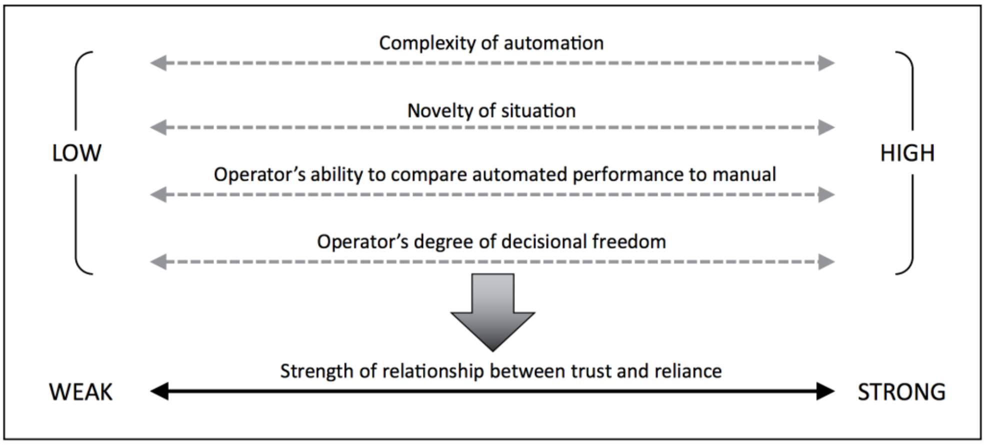

Complexity of automation
Novelty of situation
LOW
HIGH
Operator's ability to compare automated performance to manual
Operator's degree of decisional freedom
Strength of relationship between trust and reliance
WEAK
STRONG

**Figure 1.7 - Factors influencing the relationship between trust and reliance [from (Hoff and Bashir, 2015)]**

The measure of reliance, in this present work, is defined as the percentage of times the user does what the computer suggests. In the experiments presented in the later chapters (CHAPTER 4; CHAPTER 5), this measure of reliance is called 'decision match'.

## 1.6 User Interface Design and Evaluation

The second matter that the present work is addressing is, how can user interfaces be designed to ensure appropriate reliance on automation? In other words, how do different designs affect user behaviour under different automation reliability levels?

Research has shown that UI design (Kammerer and Gerjets, 2010; Kim and Moon, 1998), and the way in which automated outputs are communicated to the human have an effect on their trust in automation (de Visser et al., 2012). This points to the fact automation reliance cannot be discussed independently from the user interface, which is the main means of interaction between the human and the computer (Figure 1.1). It is important, then, to asses how UI design can affect user behaviour, specifically, how we can design UIs that support and reflect our goals regarding AoF, authority, responsibility and reliance.

Research has been undertaken in many areas regarding automation use. Some works have focused on the Allocation of Function aspect (Hancock and Scallen, 1996; Hoc and Debernard, 2002; Parasuraman et al., 1996; Sheridan, 2000), others on producing interfaces in line with the work domain (Borst et al., 2017; Burns et al., 2011; Flach et al., 1998; Vicente, 1999; Vicente and Rasmussen, 1992; Zhang and Norman, 1994). Some have focused on how humans can help overcome automation failure (Meyer et al., 2003; Parasuraman et al., 1996; Parasuraman and Manzey, 2010; Rasmussen and Vicente, 1989) and others have looked at which factors influence humans' trust and reliance on automation (de Visser et al., 2012; Lee and Moray, 1994; Madhavan and Wiegmann, 2007; Mayer et al., 1995; Merritt et al., 2013; Muir, 1987; Parasuraman and Manzey, 2010, 2010; Wickens et al., 2015). However, to the researcher's best knowledge, there is no systematic means of integrating this large body of work so that it can inform the design of HA systems.

In the design of generic User Interfaces, the display of the required information (physical information regarding the status of system components) is prioritised, with little thought given to how they are going to be used (Ham and Yoon, 2001). These interfaces increased operators' workload by requiring them to search and integrate information (Vicente and Rasmussen, 1990), while "[a]n effectively designed display reduces operators' cognitive loads and helps them to cope with complexity in dynamic systems" (Ham and Yoon, 2001). The user interface is not only the single means of communication between humans and automation (Figure 1.1), but it is also a representation of the system running in the background. It has been shown that when this representation matches the mental model of the operator interacting with it, this leads to better performance in terms of both decision time and accuracy in spotting automation failure (Ham and Yoon, 2001; Jamieson and Vicente, 2001;

McIlroy and Stanton, 2015; Vicente et al., 1995). UIs which support operators control goals and match their mental model of the system are called ‘Ecological Interfaces’. Vicente and Rasmussen (1992) have developed ‘Ecological Interface Design’ (EID), which is a framework for designing Ecological UIs.

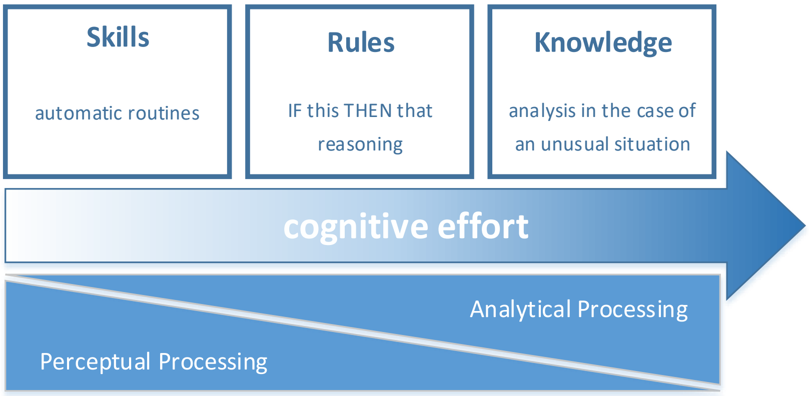
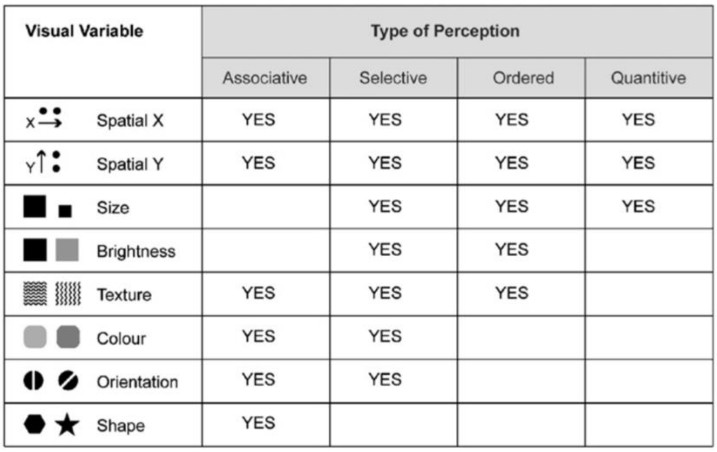


**Figure 1.8 - SRK Framework**

The ecological approach to UI design proposes that not only the constraints of the technological system and the human should be reflected in the design, but also the constraints that arise from the job description and the environment in which the interface is to be used. Ecological Interface Design, developed from Cognitive Work Analysis (CWA) (Vicente, 1999), is based on the SRK (skills-rules-knowledge) taxonomy (Vicente and Rasmussen, 1992). CWA is a methodology that consists of a number of steps that aid in the analysis of a socio-technological system’s purpose and functions, the agents’ (human and automation) abilities and responsibilities, strategies and workflow dictated by regulations of the work domain and constraints inherent to agents. The outputs of this analysis help build what is called an Abstraction Hierarchy (a framework that relates properties of the integrated work domain to the design of user interfaces (Vicente and Rasmussen, 1992). The SRK taxonomy is a framework that helps to relate the way in which information is presented to the different processing mechanisms of human operators (Vicente, 1999). Figure 1.8 shows the distinction between the three manners of processing and the cognitive effort associated with information processing. The SRK framework is based on Gibson (2014) notion of direct perception, which states that, as tasks move from requiring a knowledge-based approach to a rules- and, finally, skill-based approach, they require a decreasing cognitive effort. This is also in line with how

Rasmussen described the tendency for learned behaviour to move from being knowledge-based towards being skill-based (Sheridan, 2002).

In an attempt to extend the application of EID, Upton and Doherty (2008) integrated Data Scale Analysis (Stevens, 1946) and Visual Scale Matching (Bertin, 1983) in the design process of User Interfaces. Stevens (1946) proposed that all data fall into one of the following four categories: nominal, ordinal, interval and ratio. The addition of Visual Scale Matching (Table 1.2) was motivated, partly, by research showing that correctly matching data to visual variables leads to better performance of some cognitive tasks (Zhang and Norman, 1994). The User Interface Design Methodology that Upton and Doherty (2008) proposed can be seen in Figure 1.9.



| Visual Variable | Type of Perception | | | |
|---|---|---|---|---|
| | Associative | Selective | Ordered | Quantitive |
| Spatial X | YES | YES | YES | YES |
| Spatial Y | YES | YES | YES | YES |
| Size | | YES | YES | YES |
| Brightness | | YES | YES | |
| Texture | YES | YES | YES | |
| Colour | YES | YES | | |
| Orientation | YES | YES | | |
| Shape | YES | | | |

**Table 1.2 - Visual Scale Matching (Upton and Doherty, 2008)**

The addition of Data Scale Analysis and Visual Scale Matching brings the designer a step closer, from the analysis of the work domain, which outputs the information requirements (IR), to the realisation of an actual UI. More specifically, these extra considerations of how raw data should be treated (i.e. showed to the user) are constraints that ensure users accurately extract the information presented to them. The methodology still leaves the Design Space as a rather mysterious and infinite ‘desert’ in which the User Interface designer is left to work his ‘magic’.

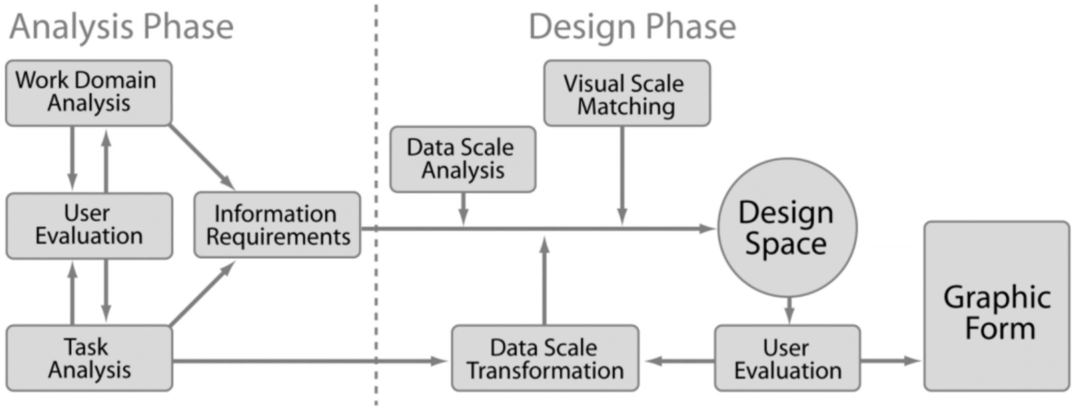


**Figure 1.9 – User Interface Design Methodology**

Bennett et al. (2012) suggest that, in the design process of User Interfaces, designers should aim at solving the issues of 'correspondence' and 'coherence'. The term 'correspondence' is used in reference to the link between the work domain and the user interface, and translates into the information content of the interface. This notion is consistent with the Upton and Doherty (2008) methodology and refers to the left-hand side of Figure 1.9. The 'coherence problem' addresses the mapping of the visual interface to the mental model of the human operator. Coherence deals with the issue of how information sources are displayed to the user and is closely related to human visual perception. Upton and Doherty (2008) begin to address this issue by the addition of Data Scale Analysis and Visual Scale Matching.

However, when designing UIs, we need a way to work out the effect of design choices on user behaviour. User Evaluation is limited to comparing user performance when using different designs (Ham and Yoon, 2001; Jamieson and Vicente, 2001; McIlroy and Stanton, 2015; Vicente et al., 1995), without having a means of working out which changes in UI design will lead to the user's change in behaviour; in other words, without a means of quantifying differences between interfaces. The aspects of a UI which can be modified need to be identified and categorised, thus reducing the domain and ambiguity of the design space, making it clearer. Each UI and UI component can be thought of as having three dimensions: Content (the informational load, i.e. what information is shown), Format (the means of interaction with it, i.e. how can the user interact with the UI) and Form (the way in which it is displayed, i.e. graphically, textually, as absolute/relative values etc., and where it is placed). A first step would be to specify the particularities of UI components in terms of

Content, Format and Form. This makes the designer aware of what he is changing on a UI and whether this change affects multiple aspects of the UI. For example, we may want to change an information source from a textual to a graphical form and this may impact on the position of the information source as well. However, the way in which the user can interact with it may also change and this matter relates to Format.

Furthermore, there is a considerable body of literature which has looked into one aspect or another of UI design (Ahn et al., 2011; Bennett and Flach, 2011; Cook and Thomas, 2005; Cossalter et al., 2011; Ellis and Dix, 2006; Griethe and Schumann, 2006; Kammerer and Gerjets, 2010; Kim and Moon, 1998; Rovira et al., 2014; Wanner et al., 2015). Thinking of UIs in terms of Content, Format and Form allows for the inclusion of this large knowledge-base into the process of design. This gives the designer a taxonomy in terms of which he can classify and make use of the previous findings in an informed and tractable manner.

Content links directly to Bennett et al.'s (2012) notion of 'correspondence', while Form relates to the issue of 'coherence'. However, there is the additional aspect of Format, which defines and describes user interaction with the interface. Perhaps, Bennett et al. (2012) would regard Format to be related to 'coherence', saying that "user's tasks are defined by that domain rather than by the visual characteristics of the display itself". However, one can argue that user interaction with the UI is rather an emerging property of the interplay between the work domain, the information content of the interface and the way in which this information is displayed. One can identify aspects of a UI which can influence interaction, such as the placement of UI components, company politics reflected in automation design (AoF, the issue of authority, responsibility and control), the action required for control as defined by the work domain (define operational bounds, set absolute values, manage alerts, etc.) and the action required for control as defined by the UI component (i.e. move slider, type in value, drag element, etc.). In what follows, the proposed CFF taxonomy is described in more detail.

### 1.6.1 Content

Content refers to the information requirements of the user-environment. This, naturally, depends on the type of decisions that need to be made, however, Content is also dictated by less obvious factors such as company politics, legislation and user preferences. For example, the control room was equipped with very large wall-mounted screens showing multiple live video feeds of the Grenoble ring road. The operators did not use them, but instead they preferred to look at one camera view at a time on their desk-mounted screens (Kibangou et al., 2015). Cognitive Work Analysis (CWA) (Vicente, 1999) has been used in the scope of the SPEEDD project in order to identify requirements for the information content. More details can be found in CHAPTER 2 and SPEEDD report D5.4 (Baber et al., 2014). More simply put, a consideration of Content aims to answer the question of *what* data should be shown to the user.

### 1.6.2 Format

The second dimension, that of Format, refers to the protocols of communication between the human and automation, or the paradigm of interaction. Format looks at the actions which users can perform in order to control, set bounds on automated operations and determine outcomes using a UI and how he can achieve these actions. User interaction with automation is defined and determined by AoF, authority, responsibility and control.

Investigating the influence of the order of response of the two agents to a specific flagged issue and the factor which triggers a response from either of the agents are also important matters that relate to the Format of the UI (explored in CHAPTER 4). Another important consideration around Format could be whether interaction between the agents should be continuous, or only prompted by the appearance of error. Debernard et al. (2002) have investigated the application of the latter paradigm. Furthermore, should the human have ultimate authority over the functioning of the system, or should authority be shared? This also brings to question the problem of who is responsible for improper functioning of the system. These issues have been investigated by Dekker (2002), Inagaki (2003) and Woods and Cook (2002), while the study presented in CHAPTER 4 discusses the matter further, showing that users may feel less responsible for the correct operations of the system in high automation reliability scenarios.

In terms of UI design, however, these questions translate into “how do users understand their role in relation to automation and how do changes in design alter their understanding of how they should interact with the automated system?" Being aware of which UI components and aspects have a bearing on operators' interaction with automation, better informs the interface designer of how to establish a coherent relationship between goals regarding AoF, authority, responsibility and control and operators' understanding of their role and position in the system.

### 1.6.3 Form

CWA (Vicente, 1999) provides the designer with the information requirements, which define the Content of the UI, while interaction modes are given by the consideration of Format. So far, the designer knows what data to display, and how users should interact with it, however there is no indication of how the data should look and where it should be displayed. Form is the third and final dimension of UI design and it refers to the way in which information is displayed to the user.

CWA does not inform UI designers of how information should be placed on the interface (i.e. should multiple information sources be integrated, should they be grouped together, etc.), nor does is specify how each information bit should look. In aid of these issues come the Proximity Compatibility Principle (PCP).

The concept of 'task proximity' (Wickens and Carswell, 1995) states that information sources which need to be used by the operator for a specific task should be spatially grouped together or integrated. Multiple studies reported that consideration of PCP leads to superior operator performance (C. Melody Carswell, 1992; Carswell and Wickens, 1987; Wickens and Carswell, 1995) and that layout influences human interaction with UIs (Kammerer and Gerjets, 2010). An assumption that could be made is that differences in information accessing costs arising from changes in layout (including higher/lower 'task proximity') would lead to differences in how operators access and use the information displayed to them in order to complete the same task. This matter has been investigated in CHAPTER 5.

The notion of 'display proximity', along with, Principles of Ecological Interface Design (EID) (Burns et al., 2011; Flach et al., 1998; Gibson, 2014; McIlroy and Stanton, 2015; Rasmussen and Vicente, 1989; Vicente and Rasmussen, 1992) can be used for producing UIs that match operators' understanding of the system they need to control. Display proximity states that visual objects (i.e. display components) which have similar appearance will be processed together. Therefore, physical proximity of information sources (or display components) is not the only matter affecting user integration of information sources and that visual appearance can also influence user behaviour in term of decision time and accuracy. Moreover, UI design (Kim and Moon, 1998) and the way in which automated outputs are communicated to users (de Visser et al., 2012) can affect their trust in automation and, thus, reliance on it. Form raises the question of how the chosen visual representation of the required information and its placement relative to other UI components affects user behaviour.

The SRK framework (Skills-Rules-Knowledge)(Vicente and Rasmussen, 1992), on which the concept of Ecological Interface Design is based, also gives promising input regarding Form. It shows that, as cognitive effort required for extracting information and decision making is reduced, that is, as analytical processes shift to perceptual processes (see Figure 1.8), human performance in terms of accuracy of response and decision time is increased (Flach et al., 1998; McIlroy and Stanton, 2015; Rasmussen and Vicente, 1989). This implies that the Form in which information is communicated plays a big role in the overall performance of the HA system. While EID is very good at producing a list of Information Requirements and emphasising the importance of the UI matching the mental model of human operators, it does not provide a clear methodology for arriving at an 'Ecological Display'. Upton and Doherty (2008) have extended EID through the introduction of Data Scale Analysis and Visual Scale Matching (Table 1.2), which say of what type the data are and what are the possible visual representations of those data, respectively. This takes the designer a step closer to an actual UI.

## 1.7 Content/Format/Form – Conceptual Example

### 1.7.1 Version 1

Let's consider a relatively simple system designed for indoor ambient temperature control. Now, let's imagine how would a user interface for such a system look like. In terms of Content, the UI would need to display the temperature value in degrees Celsius. In terms of user functions, operators could be expected to be able to turn on and off a heating system. In the dimension of Form this would, possibly, translate into two buttons for turning the heating system on and off, respectively and the display of the temperature value as a number. In terms of Format, the user would be able to click the two buttons to control the heating system. See Figure 1.10, below.

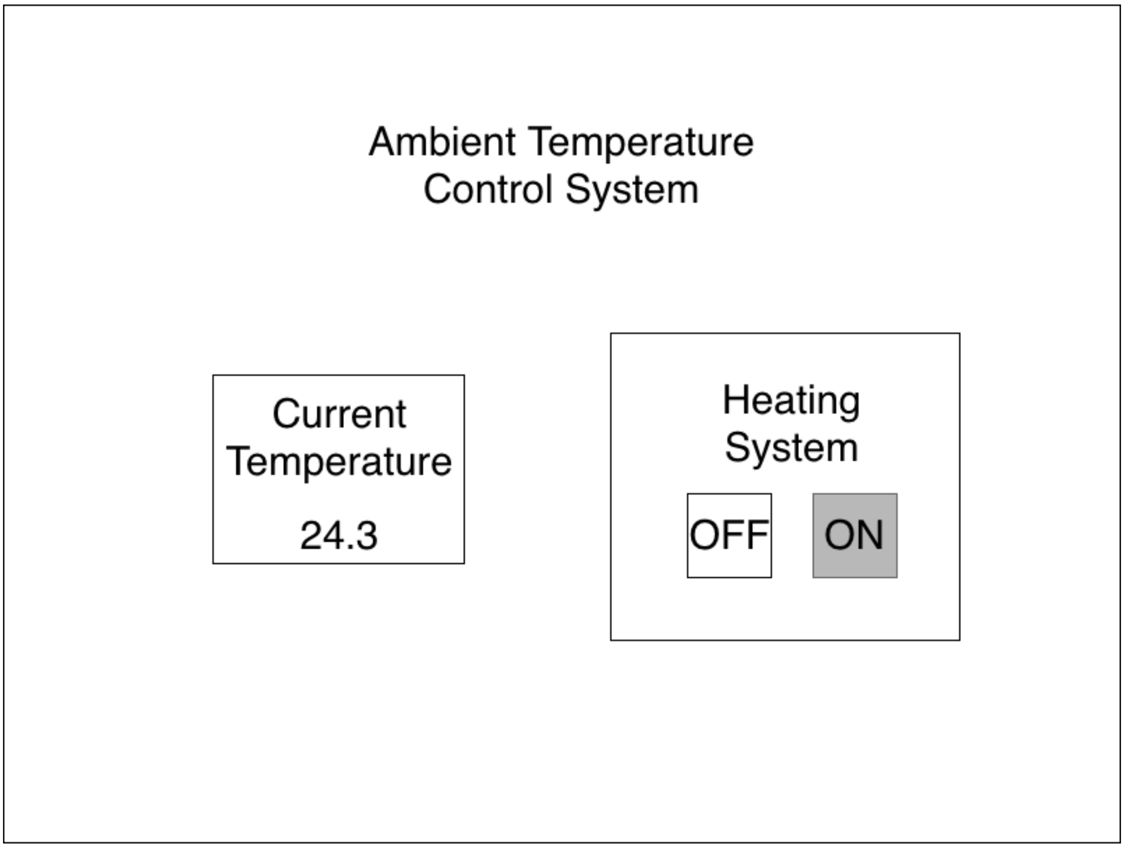


**Figure 1.10 - UI v1**

## 1.7.2 Version 2

Let’s assume now that the system’s complexity was increased by the addition of a cooling system. This would affect UI Content as it would the display of the status of the cooling system in addition to that of the heating system and the ambient temperature value. This relatively simple change has affected what the user is able/required to do (from turning the heating on and off to keeping a stable temperature by using both heating and cooling systems) and has increased the number of possibilities in the dimensions of Format and Form. Following from the previous UI version (keeping Form and Format constant, i.e. on/off buttons), a new interface could look like the one in Figure 1.11.

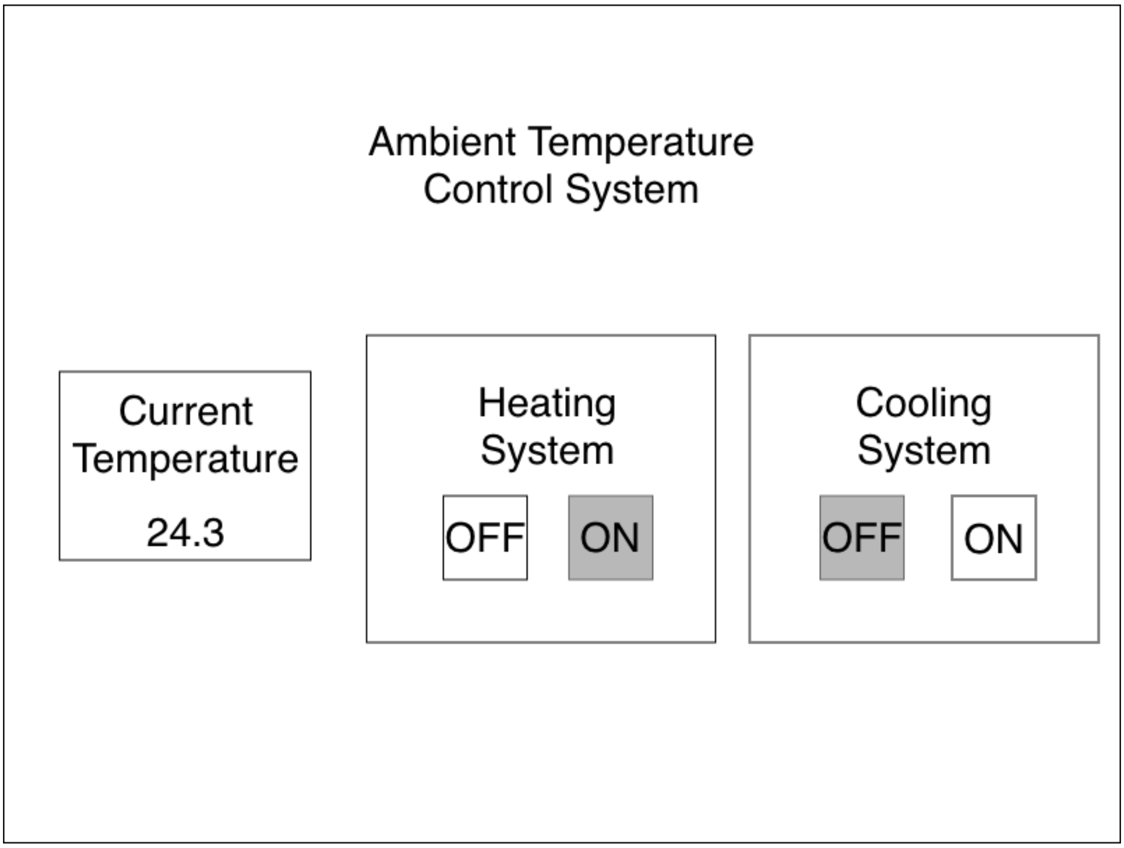


**Figure 1.11 - UI v2**

### 1.7.3 Version 3

Adding automation that decides whether to turn the heating or cooling system on or off depending on target temperature and current ambient temperature would further change the UI of this control system. Users of this system might be expected to merely set a target temperature based on the current temperature and whether they feel hot or cold. Thus, one instantiation of this UI in the dimension of Format could be limited to the ability to increase or decrease the target temperature by clicking one of two buttons. In this scenario, Content may be limited to the display of the current and target temperatures, along with buttons that increase and decrease the target temperature, respectively. In terms of Form, temperatures can be shown as numerical values and the actions of increasing and decreasing the target temperature may be offered by two +/- buttons lateral to the displayed value. See figure Figure 1.12, below.

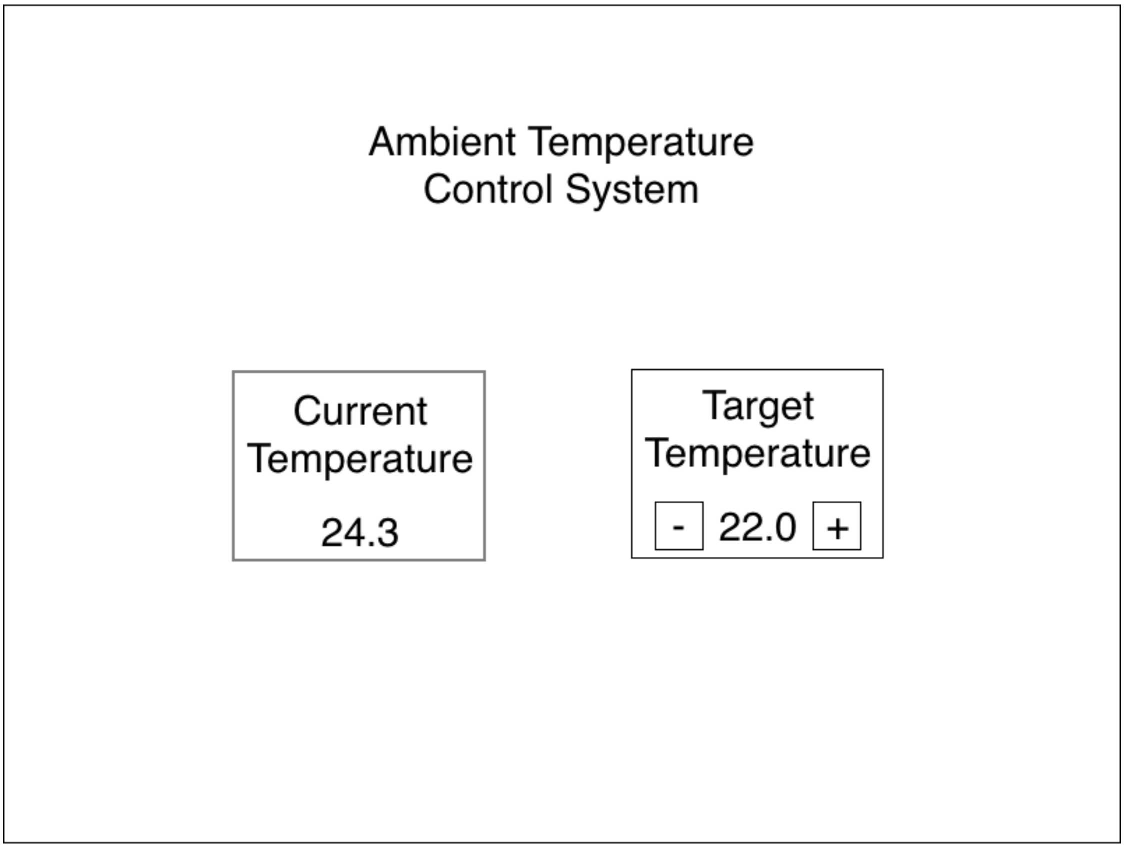


**Figure 1.12 - UI v3**

Alternative UI Formats could be given by presenting the user with an input box or a slider for changing target temperature, instead of the two buttons lateral to the value. For these changes, Content would stay the same, but Format and Form would change.

### 1.7.4 Version 4

Increasing the level of automation even further, such that the target temperature is decided upon by the automated system (so that optimum operation parameters are ensured) can change the function and the UI of the human-automation system even further. The user may no longer need to control the ambient temperature at all, and his job would change from performing a control task to performing a monitoring task. In terms of Format, the display would no longer need to support user interaction with the heating or the cooling system, nor would it need to allow for changes in target temperature. These changes also affect Content, as no buttons would need to be displayed on the UI.

Moreover, the user would need to be able to determine whether the automation is performing well (change in Content). Therefore, he would need to check whether the automation control decision (turn heating/cooling system on/off) is correct in relation to current and target ambient temperatures. One version of the UI could look like this (Figure 1.13):

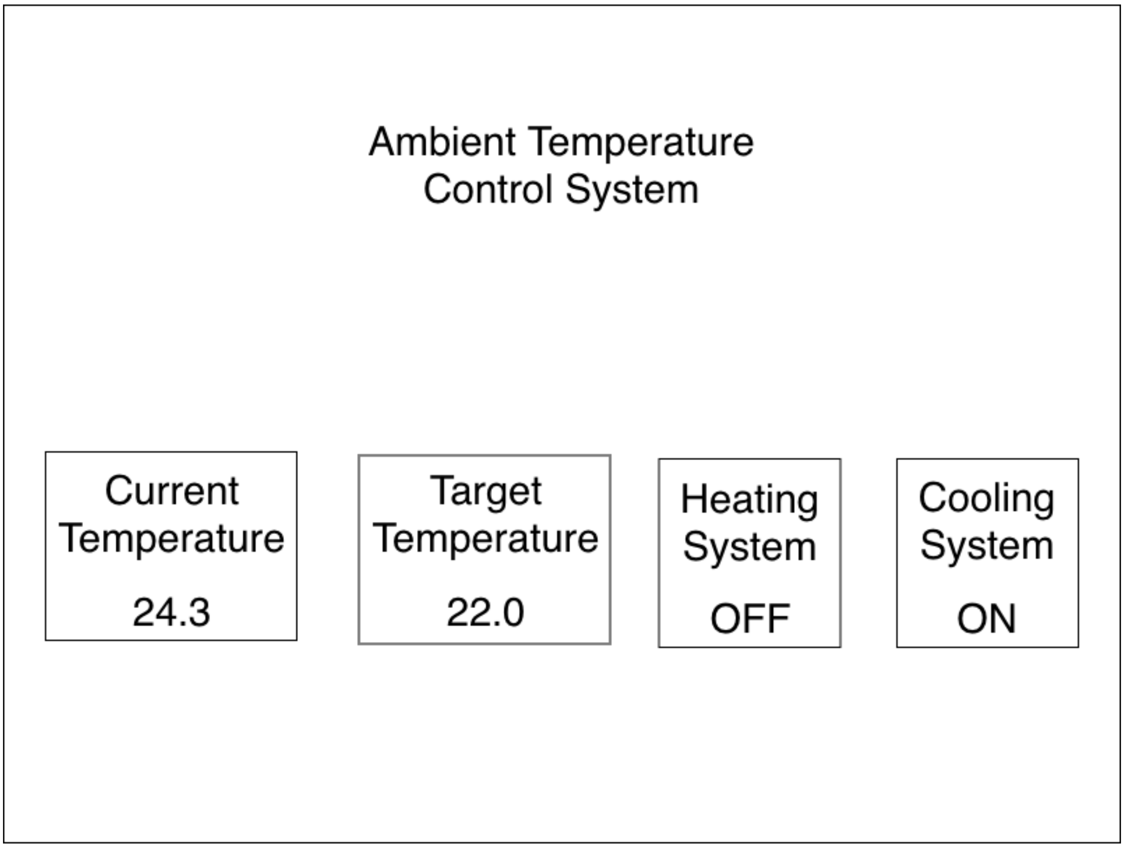
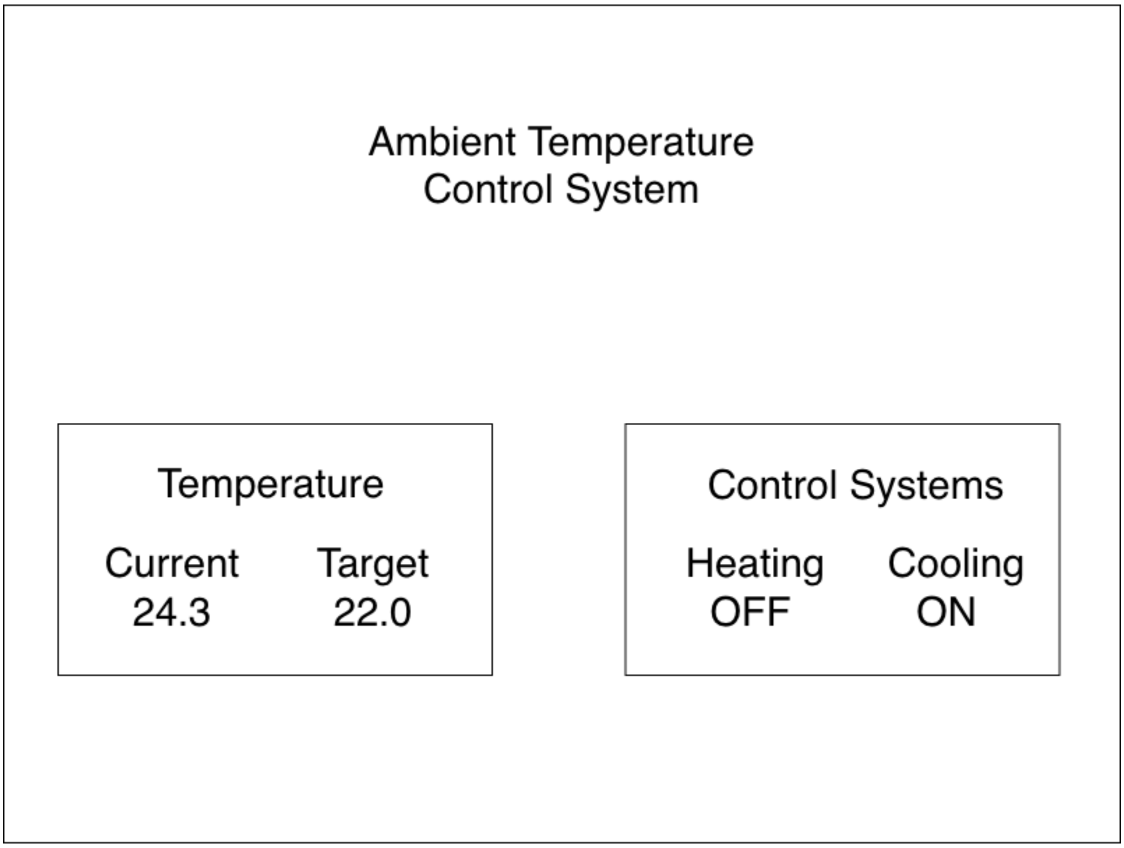


**Figure 1.13 - UI v4**

### 1.7.5 Version 5

In terms of Content, the UI in Figure 1.13 provides all the information required by the user to appropriately spot errors in automation performance. By applying the Proximity Compatibility Principle (PCP), a change in the dimension of Form that may improve user performance is the grouping together of alike components (temperatures and control systems) (Figure 1.14).



Ambient Temperature
Control System

Temperature

Current 24.3

Target 22.0

Control Systems

Heating OFF

Cooling ON

**Figure 1.14 - UI v5**

### 1.7.6 Version 6

Furthermore, the above Control System component can be integrated so as to support non-integrative processing, thus reducing cognitive effort and, therefore, decision time even further (see Figure 1.15).

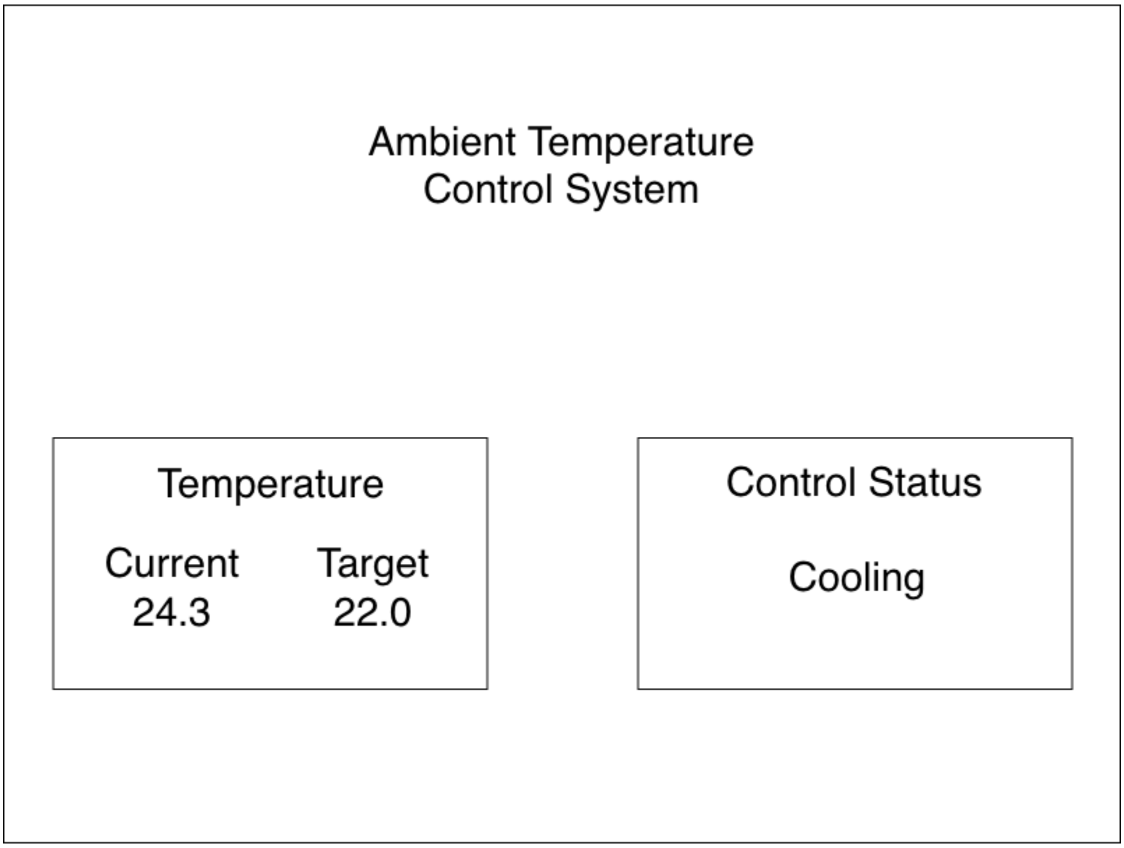


**Figure 1.15 - UI v6**

## 1.7.7 Version 7

Moreover, the Form of UI v6 (Figure 1.5) is purely textual. Textual information may take much longer than graphical information to decode. Therefore, while UI versions 4-6 (Figure 1.13 - Figure 1.16) may satisfy the functional requirements, they may still not deliver the best results in terms of decision time, for the task at hand. The UI in Figure 1.15 could be further improved in the dimension of Form by drawing on the knowledge provided by SRK framework. SRK shows that cognitive effort and, thus, decision time is reduced by as tasks move from analytical to perceptual processing, i.e. from rules to skills (Figure 1.8). The UIs shown until now, draw heavily on user skills as the user need to perform a series of if-then reasoning processes (i.e. if target temperature is lower than current temperature then the correct action would be to turn cooling on and heating off). An improved version of this UI could be (Figure 1.16):

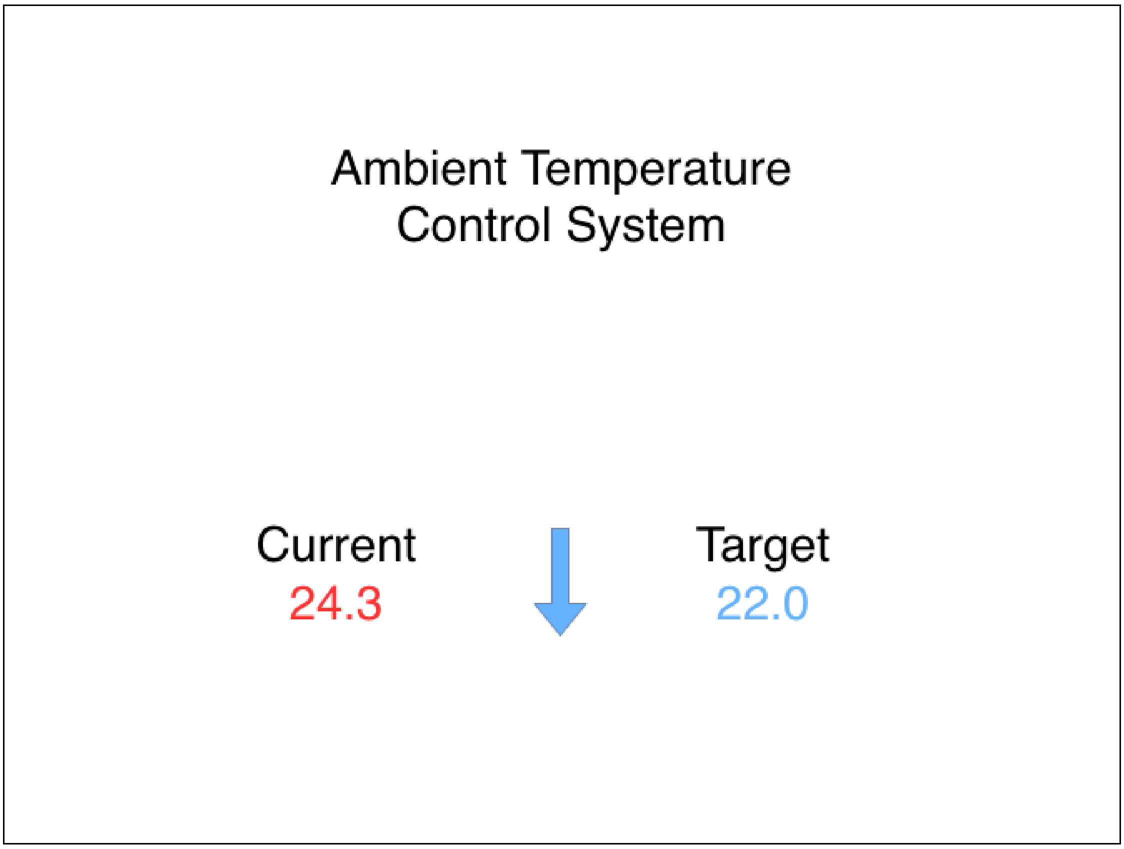


**Figure 1.16 - UI v7**

The colour of the current temperature value shows its status relative to the target temperature. Red signifies that the current temperature is higher than the target. The blue arrow pointing down signifies that cooling is turned on, i.e. action to reduce the current temperature is being taken. The fact that the colour of the arrow matches the colour of the target temperature value, confirms to the user that the correct action is being taken by the automation. Stability of the system would be illustrated by having both temperature values shown in green and the absence of the arrow. All of the above (Figure 1.13 - Figure 1.16, i.e. the application of PCP and SRK) mark changes in Form, whilst Content and Format remaining constant.

### 1.7.8 Version 8

However, from a Human Factors perspective, UI v7 (Figure 1.16) is far from being the ideal one as it takes the human operator out of the control loop, leaving him unable to intervene in case of an automation error. To amend this issue, one can add a button below the arrow to toggle the heating/cooling systems. This change in Format, also changes the UI Content and Form (addition of the toggle button). However, this Format change can be implemented without any further change in

Content or Form from Figure 1.16. The UI can simple allow the user to override the automated decision by clicking on the arrow. This would cause the arrow to change orientation and the underlying heating/cooling systems to change status, accordingly.

### 1.7.9 Format, Form, Function – Disambiguation

Let us reconsider the thermostat discussed above (1.7.1-1.7.8), where the user can set the desired temperature. The UI of the thermostat may have two buttons (labelled '+' and '-') to change the desired temperature, so that if the user wants to lower the temperature, he clicks on the '-' button and vice-versa. If the Form of the user interface was changed from the two on/off buttons to a text input box, then the Format would also change requiring the user to type in the desired temperature value. In both cases, the function of the thermostat would still be to allow the user to set the desired temperature, but the way in which the user accomplishes this goal is different (i.e. clicking buttons vs typing in a value).

To reiterate, Form relates to how UI components look, while Format relates to how users interact with them. In this scenario, it is not evident how Format differs from Form, as changing the Form of the UI (on/off buttons to text input) also determines user interaction (Format). However, let us consider Version 8 (1.7.8). Here, UI Form is the same as Version 7, while Format is different (arrow now clickable) due to the change in function in Version 8 to allows users to override automated decision. Therefore, it can be seen that Format is neither equivalent to Form nor equivalent to function, as we have seen instances when Format changes independently from both. However, Format can be affected by changes in both Form and Function.

### 1.7.10 Summary

From the previous design exercise, we have seen that:

- Content, Format and Form may interact. In Version 2, the addition of a Cooling System (Content) which the user was required to control, also impacted on UI Format, as the UI had to provide a means for this control (on/off buttons).
- Function may affect Content, Format and Form (Version 4)
- Form may change independently of Content and Format (Version 4 (Figure 1.13) - Version 7 (Figure 1.16))

- Format of the UI can change even when the defined user function stays the same (Version 8 and Version 3)
- Format and Form can change without altering UI Content (Version 3, last paragraph)
- Format can be changed without interfering with Content or Form (last example in Version 8, last example in Version 3)

## 1.8 Fitting in Past Research with the Content Format Form (CFF) Taxonomy

CWA and EID provide a methodology for answering the question of what information should be displayed to the user. Where it should be placed or how it should look has to do with the positioning of information (Wickens and Carswell, 1995) displays (UIs), the ease of extraction of actionable information (Rasmussen, 1983) and the type of perception (Bennett and Flach, 2011) required to extract the encoded information associated with them. The question of how UIs allow users to achieve their goals relates more closely to the to notions of Authority, Responsibility and AoF. Answering these questions aims to help designers produce user interfaces that reflect the mental model of their users, integrate seamlessly with the work domain and achieve a high task fidelity, while at the same time reducing the cognitive effort spent in order to extract information. Vicente and Rasmussen (1992) have undertaken a comprehensive literature review of the area EID, showing promising results for its application.

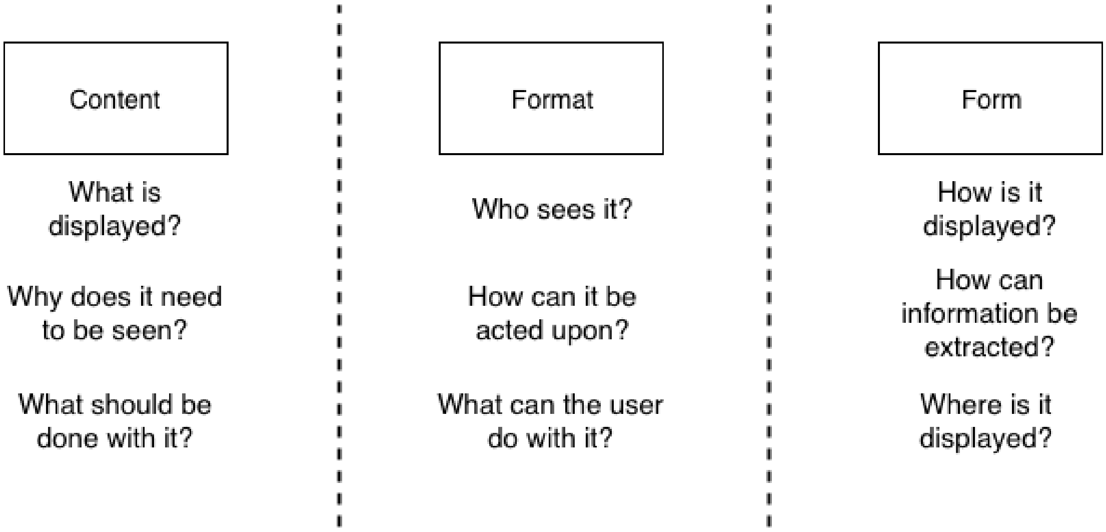

Content
What is displayed?
Why does it need to be seen?
What should be done with it?
Format
Who sees it?
How can it be acted upon?
What can the user do with it?
Form
How is it displayed?
How can information be extracted?
Where is it displayed?

**Figure 1.17 - The “What”, “Where”, and “How” of UI Design**

Figure 1.17 illustrates what questions UI designers might ask when considering the Content, Format and Form aspects. What is rather under-researched is the effect on human behaviour of changes in Content, Format or Form of the UI.

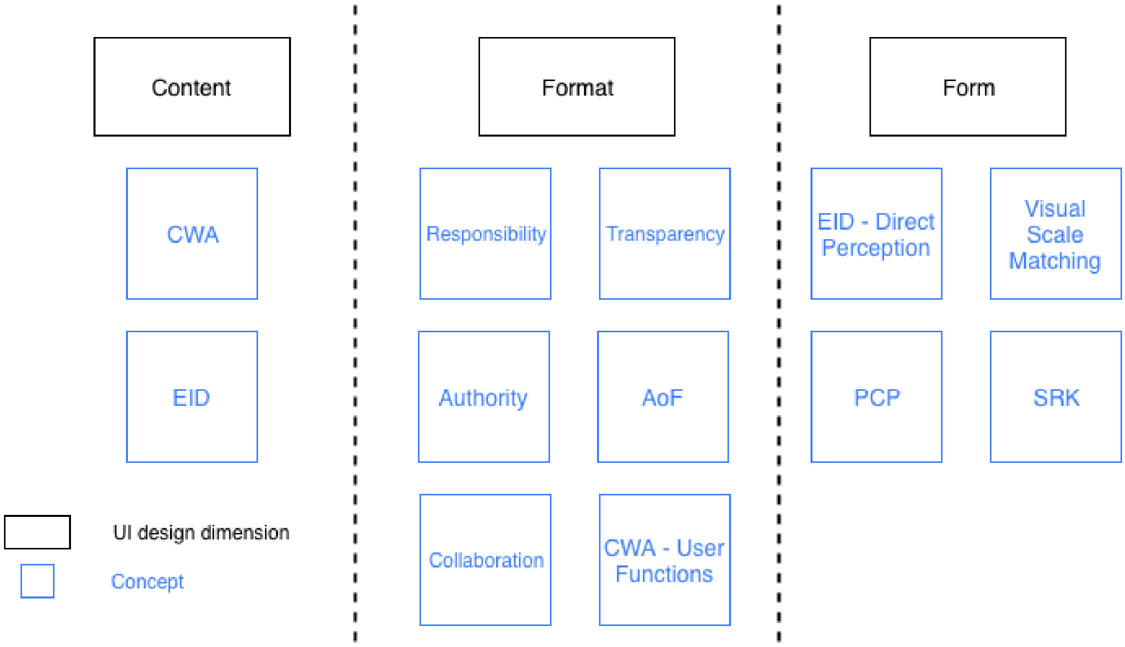


**Figure 1.18 - Relationship between dimensions of interface design and theoretical concepts from past research**

Through careful consideration of the three dimensions of display design (Content, Format and Form) we can extend the methodology proposed by Upton and Doherty (2008), by allowing for changes in display design to be tracked, thus informing both the design and the evaluation process of visual displays (see Figure 1.19). The categorisation of visual variables as either pertaining to Content, Format, or Form leads to a more clearly defined design space and allows for the elaboration of more controlled experiments for the purpose of evaluation of one or more versions of a display/interface to automation. Moreover, the CFF (Content/Format/Form) taxonomy allows for the consideration of past research providing input regarding design procedures, guidelines and other aspects related to UIs and HAS in the design process in tractable manner (see Figure 1.18).

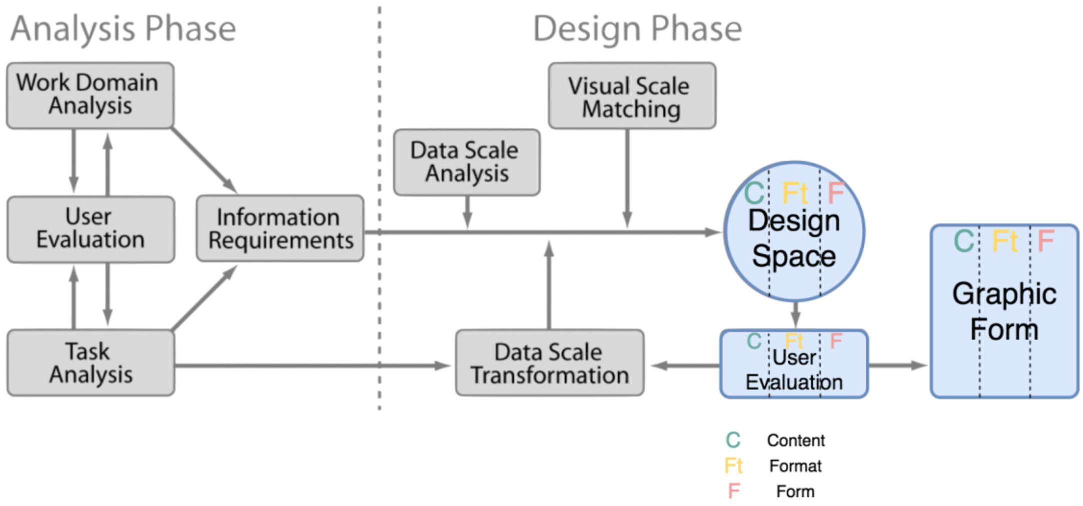


**Figure 1.19 – Extended User Interface Design Methodology**

## 1.9 Summary

Most sectors of human activity have seen a great increase in automation in the past few years. Even though automation is taking over more and more of the tasks formerly performed by humans, there still is a place for a human in the loop. Researchers have pointed out that the potential pitfalls of extreme automation can be avoided by having a human overlooking and/or working alongside automation.

This chapter presented the issues that arise when humans and automation work together at solving tasks. The notions of trust, reliability and AoF have been introduced and the importance of appropriate reliance has been emphasised. It has also been shown that human reliance on automation is influenced, not only by automation reliability, but also by the design of the UI, which sits at the boundary between humans and machines and serves as the means of communication between them.

Existing UI design methodologies are very good at defining the information requirements, however they do not tell the designer how these information sources should be transformed to visual display components. Thinking of displays in terms of Content, Format and Form can help in the design and testing/evaluation process, thus further advancing existing design methodologies. Moreover, it is shown how this approach can be used to keep track of design changes and their effects on aspects of Human-Machine Systems, such as, automation reliance and human/system performance. Isolating the aspects of display components in terms of Content, Format and Form ensures that changes in the design of a User Interface is done in a tractable manner, with awareness of the impact on user behaviour in the design stage and with the ability to pin-point differences in human behaviour to changes in design, in the process of evaluation.

The scope of this PhD is limited to a small number of variables in the dimensions of Format (order of response, transparency) and Form (display proximity, graphical vs textual display of information). The effects of these manipulations of the display were investigated in the context of varying automation reliability. The effects of varying UI Content have not been investigated due to the large body of literature which stresses on the Information Requirements being satisfied and because that this has been achieved in the design of the SPEEDD UIs by performing CWA and by following EID guidelines (see CHAPTER 2). The questions that are explored as part of this work are:

1. What effects does automation reliability have on human decision making?
2. How can we design user interfaces to help users cope with these effects?

These questions are investigated in the context of the EU Project SPEEDD, which is introduced in the following chapter.

# CHAPTER 2 UI DESIGN PROCESS AND EVALUATION

The work presented in this thesis is based on real-world use-case of Traffic Management, as defined in the European Project SPEEDD. This chapter illustrates the relationship between the work undergone in this PhD and the SPEEDD Project. The SPEEDD Project is succinctly introduced, after which The Traffic Management use-case is presented. Moreover, the design process of the SPEEDD UIs is presented, along with the evaluation methods for these UIs.

*Parts of section 2.3 have been published in [1](Morar et al., 2015a). Sections 4.1 and 6 in [1] reproduced.*

## 2.1 Introduction

This research is funded by the European project SPEEDD[1] (Scalable ProactivE Event-Driven Decision-making) which aims to bring fully integrated big data solutions to the areas of Traffic Management and Credit Card Fraud Investigation. Work on the European project was undertaken by partners from:

- National Centre of Scientific Research 'Demokritos' (Athens, Greece) – on-line and off-line machine learning, technical development and architecture integration
- IBM Research (Haifa, Israel) – architecture design and implementation
- ETH Zurich (Switzerland) – Control Theory approach to managing traffic, technical development and architecture integration
- Technion-Israel Institute of Technology (Haifa, Israel) – architecture scalability
- CNRS (Centre National de la Recherche Scientifique, Grenoble, France) – developing new approaches to traffic management

[1] European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 619435

- FeedZai, Consultoria e Inovação Tecnológica, S.A. (Lisbon, Portugal) – providing access to data and expert knowledge in fraud investigation
- University of Birmingham (Birmingham, UK) – development and evaluation of Visual Analytics systems for the two use-cases, technical development and architecture integration

Our work at the University of Birmingham involved developing and testing the user interfaces, along with the back-end integration with the systems that the consortium produced, as well as evaluating the performance of the overall human-automation systems.

The technology developed as part of the project had the trifold purpose of advancing the state-of-the-art in terms of event processing, producing a reusable architecture that one can implement in any heavily data-driven domain and of adding value to the domains of traffic management and credit card fraud investigation. These goals were achieved by producing and integrating automation that makes use of readily available data to compute assessments that better inform operators/analysts in the process of decision making. The architecture is designed so that it can take advantage of the high volume and high velocity of data coming through, being able to produce both automated control signals and user recommendations. For the scope of this thesis, the discussion will be limited to the Traffic Management use-case.

## 2.2 Human-Machine Systems in the Context of SPEEDD – Traffic Management

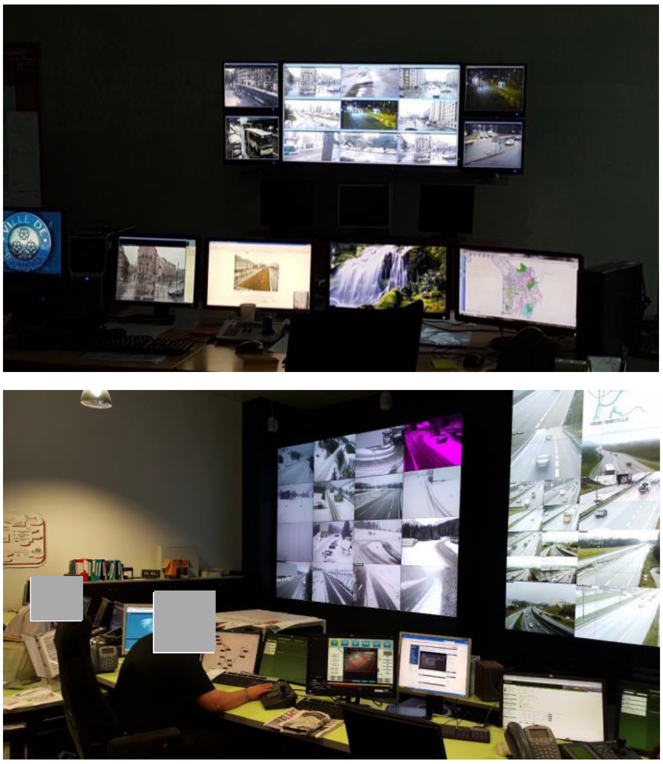


**Figure 2.1 - DIR-CE TM Control Room**

Data collection infrastructures in cities has allowed for Road Traffic Management (TM) to extend from congestion management and speed control to pollution monitoring or multimodal transport management (Batty, 2013; Townsend, 2013). Data in these systems can be captured from a range of data sources, including in-vehicle Satellite Navigation (SatNav) devices, road-side Closed-Circuit Television (CCTV), sensors in the road, and voice communications (via radio from roadside personnel or emergency services). As such, 'big data' collected from the various data sources in Road Traffic Management present an important challenge to humans in the loop. Even assuming that the sensors have modest sampling rates and low bandwidth, there is still potential for the volume of data to become overwhelming for the human operator.

From observations of the DIR-CE traffic management control room in France, operator decision making was into two broad categories. The first concerns the management of traffic flow. Road Traffic Control operators can use Variable Message Signs (VMS) to manage speed limits in a bid to reduce risk and increase traffic flow. Given the variability in conditions which can influence traffic behaviour, the role of the human operator is to judge when and how to use VMS. Ideally, the operator would make changes to the signage in anticipation of problems, but it is often the case that, rather than being proactive, current operations tend to be reactive. This is partly an issue of the nature of data that are available to operators, with limited capability to make direct predictions of future state. The SPEEDD project demonstrates that it is possible to make congestion predictions several minutes in advance of congestion occurring, which could be sufficient time for the operator to modify VMS.

The second category of decision concerns the management of traffic activity through the control of intersections, e.g., in terms of controlling the sequences at which traffic lights operate. In cases where control of traffic signals is automated, the role of the operator is to ensure that the appropriate sequences are being applied and to monitor traffic activity in order to intervene as necessary (e.g., in case of accidents). Combining these two categories of road traffic management decision making could allow congestion on major routes to be managed using traffic signals which control ingress and egress on these routes. In this case, automated control would require real-time data on traffic activity in order to manage traffic light schedules.

At the beginning of the SPEEDD Project, the DIR-CE traffic managers did not do any adjustments to ramp metering rates. These were operating according to schedules set in advance and the only control they had was whether to turn traffic lights off (i.e. usually late at night) or leave them operating. Apart from this binary level of control of traffic lights, operators were able to select from a list of messages to display on VMSs. In terms of automation, they had an obstacle detection algorithm running on the CCTV feed. This system triggers alerts whenever it can detect pedestrians, cyclists or stopped vehicles in the road or on the side of the Grenoble ring road. However, this automation was not fully integrated with their system, leading to false alarms being triggered in the case of scheduled maintenance on the road, for example. The way that operators used this automation was to verify the alert the first time it triggered at a specific location by looking at the CCTV feed at that location.

SPEEDD was looking to add automation to control ramp metering rates of inbound ramps (ramps leading traffic onto the ring road). Based on data gathered from sensors buried in the road, which were already available to the operators, metrics such as ramp occupancy, main road density, average vehicle speeds, average distance vehicles could be computed. The challenge, however, was to see how could operators make sense of the data. As previously mentioned, these data were available to them but they were not using them. Moreover, as management of traffic lights was previously an on/off problem, how could automated fine tuning of ramp rates be integrated in their work? Should they be able to completely override automated control values, set boundaries on the control space, or merely monitor their status in order to spot errors in operation?

From interviews with, and observations of, operators in a road traffic control room (Starke et al., 2017), a descriptive model has been developed, using Cognitive Work Analysis, of how operators combine information gathering with making a decision. This undertaking was a joint effort of the whole team at University of Birmingham and can be found in (Baber et al., 2014b). I do not report this process in the thesis.

## 2.3 Design of the SPEEDD UIs

*Parts of this section have been published in [1](Morar et al., 2015a). Sections 4.1 and 6 in [1] are reproduced in 2.3.1 and 2.3.2.*

The user interfaces for the two use-cases underwent a very similar development process which involved a study of the work environments (CWA)(Vicente, 1999) (when possible) and analysis of the tasks that are completed on a daily basis along with the procedures for completing them (Baber et al., 2014). The design of the UIs also took into consideration the requirements and limitations of the underlying technical systems that support these tasks and, finally, previous research on appropriate visual representations of data (Upton and Doherty, 2008) and UI design (Gibson, 2014; Rasmussen and Vicente, 1989; Wickens and Carswell, 1995). These three factors, more specifically, the organisational, technical and perceptual characteristics have guided the development of the interfaces and informed different aspects of it. Figure 1.19 illustrates the design methodology used in order to design the SPEEDD user interfaces. This section also documents what caused the changes in UI design and discusses them in terms of Content, Format and Form. To remind the reader, Content refers to the information displayed on the screen, Format relates to layout of information sources and the way in which interaction with the information and control of the system in question is achieved, and finally, Form is the way in which the information is displayed to the end user.

Traditionally, the design of user interfaces is done by technical teams as an augmentation (or rather, afterthought) (Few, 2013), or terminal to the underlying automation developed (technology-centred approach). This very often leads to disuse of the entire system (Parasuraman and Riley, 1997) and, where that is not an option, to an unnecessary increase in complexity of the end-user’s work (e.g. increased workload). The SPEEDD user interfaces have undergone an incremental design process in which both social and technological aspects of the work environment have been taken into consideration and have informed the final prototype designs, presented in this chapter.

Traditional UI design is merely interested in Form. EID adds and stresses the importance of appropriate representation of Content. The user interfaces that have been developed as part of the SPEEDD project are not concerned merely with the visualisation of the information content for the operator, but also of AoF, reliability, responsibility, authority. These aspects relate to Format and the consideration of this dimension of UIs, in addition to Content and Form, is what makes this work different from other interface design work.

The design process began with a study of the work environment which provided us with an understanding of tasks traffic operators deal with on a daily basis, the available resources and usual procedures they follow, which determined the informational requirements of the UI, or more specifically, its content. SPEEDD deliverable D5.4 (Baber et al., 2014) shows how this study was conducted and describes the data gathering process.

In order to understand the nature of the domain and the decision making that Road Traffic Operators are required to perform, we visited Road Traffic Control Rooms. This provided an initial perspective on operators’ work and an opportunity to record it using Hierarchical Task Analysis (HTA) (Stanton, 2006). From this description, one can begin to discern possible strategies that operators could apply in their selection of information. A study was conducted in which eye-movement data

(Figure 2.2), using Tobii glasses with infra-red markers on monitors to track gaze (sampled at 30Hz frequency), were collected in the working control room and these data were used to define information search strategies (Starke et al., 2017). It was clear that the strategies were influenced by the operators' experience and by the availability of information. However, the strategies were also influenced by the priorities set by National policy and local ordinances (in terms of traffic regulations). This study of the DIR-CE (Direction Interdépartementale des Routes Centre-Est) Control Room, along with previous research (Folds et al., 1993), has allowed us to formulate the requirements for the TM use-case.

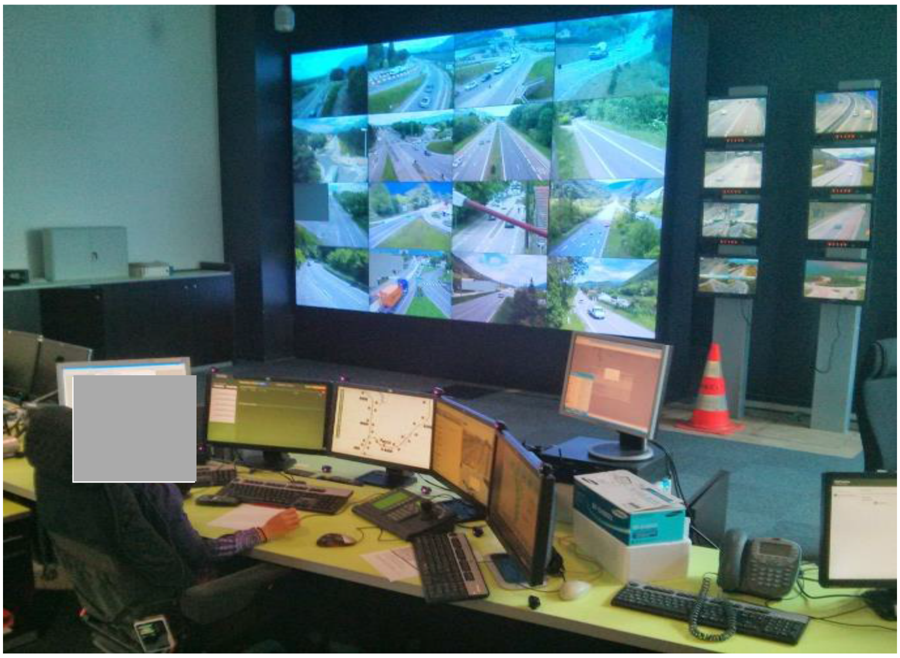


**Figure 2.2 - Collecting eye-tracking data in traffic control room**

## 2.3.1 Requirements for the Traffic Management Use-Case

- To ensure minimal congestion in the road network
- To ensure minimal risk to road users
- To enable minimal journey times for road users
- To ensure informed road users
- To support maintained infrastructure
- To encourage compliant road users
- To support immediate response to incidents
- To produce an auditable record of activity

### 2.3.2 Initial Layout

Following the CWA, an initial layout of for the User Interface was produced. This can be seen in Figure 2.3. It contains 8 regions. The following list outlines some of the options that are being considered in the design. Items in the list marked * correspond to existing information sources in the control room.

1. Road status (traffic conditions): This could also compare current traffic conditions with the same time last week, or predicted traffic conditions and likely trends;
2. Values / trends / forecasts: this component could provide operators with views of the predicted traffic, or driver behaviour, to allow comparison between alternative courses of action;
3. Road user goals: this UI component could indicate information which might be relevant to road user activity, for instance, alternative routes which drivers might take if there is congestion;
4. Driver behaviour and compliance: this UI component could indicate how road users are behaving. This could include average speed in each lane or average distance between vehicles;
5. CCTV content / control*: this UI component would present the images from the selected CCTV camera to the operator, and allow the CCTV camera to be controlled;
6. Control activity, signage content*: this would show the actions that the operator is able to perform and the content which could be presented on variable message signs;
7. Log, open tasks, scheduled events*: this would show the log of the current incident that the operator is working on, together with open tasks or any scheduled events that need to be dealt with;
8. Map of road network*: displayed as a map of the ring road (either a schematic as in the current design or a more detailed map of Grenoble and the road network), with key Objects indicated, e.g., CCTV and sign locations, junction (ramps) etc. This could also be used to display the location of incidents, such as congestion.

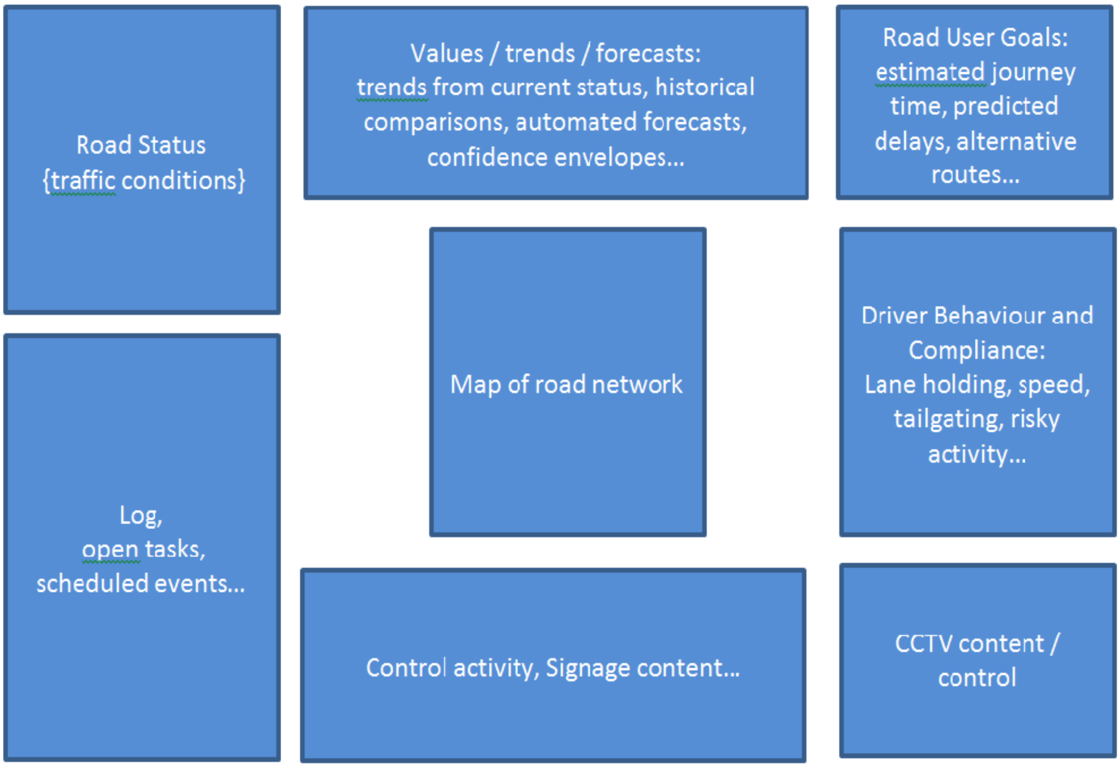

Road Status
{traffic conditions}
Values / trends / forecasts:
trends from current status, historical comparisons, automated forecasts, confidence envelopes...
Road User Goals:
estimated journey time, predicted delays, alternative routes...
Map of road network
Driver Behaviour and Compliance:
Lane holding, speed, tailgating, risky activity...
Log,
open tasks,
scheduled events...
Control activity, Signage content...
CCTV content / control

**Figure 2.3 - Schematic User Interface Layout for TM**

One would assume that transferring this conceptual design to an actual working prototype is straight-forward and implies the mere placement of the information in individual boxes in Figure 2.3 on a screen. This is also the point where research in the area of Ecological Interface Design (EID) stops. After the analysis of the work domain is achieved, information requirements are defined a user interface has to be then designed according to the identified requirements and there is no methodology for achieving this.

### 2.3.3 Applying Visual Scale Matching

CWA establishes the information requirements of the UI. Visual Scale Matching can be further applied in order find the visual representation requirements of each information source/data stream.

**Table 2.1 - Visual Representation Requirements for Information Sources TM**

| | **Visual Representation - Requirements** |
|---|---|

| Data to be Visualised | Associative | Selective | Ordered | Quantitative |
|---|---|---|---|---|
| **Density** | Yes | | Yes | |
| **Speed** | Yes | | Yes | |
| **Ramp Rate** | Yes | | Yes | non-mandatory |
| **Ramp Occupancy** | Yes | | Yes | |
| **Ramp Overflow** | Yes | Yes | | |
| **Predicted Ramp Overflow** | Yes | Yes | | |
| **Congestion** | Yes | Yes | | non-mandatory |
| **Predicted Congestion** | Yes | Yes | | |

By comparing the requirements shown in Table 2.1 above with the Visual Scale Matching presented by Upton and Doherty (2008), I have been able to define the most appropriate visual encoding for each data stream/information source (Table 2.2).

**Table 2.2 - Visual Encoding of Information Sources for TM**

| Data to be Visualised | Visual Encoding |
|---|---|
| **Density** | Spatial, Size, Colour |
| **Speed** | Spatial, Size, Colour, Brightness |
| **Ramp Rate** | Spatial, Size, Colour, Brightness |
| **Ramp Occupancy** | Spatial, Size, Colour, Brightness |
| **Ramp Overflow** | Spatial, Brightness |
| **Predicted Ramp Overflow** | Spatial, Brightness |
| **Congestion** | Spatial, Size, Colour, Brightness |
| **Predicted Congestion** | Spatial, Colour, Brightness |

## 2.4 Summary of Design Process

Key to the development of HA Systems is an appreciation of how HA operates in a working environment in which other actors will share information with each other, or will interact with systems outside the core HA system. This means that it important to appreciate the Socio-Technical Infrastructure in which the technology will be used. This chapter has attempted to relate information need to information visualization. The latter is concerned by *how* the available information is presented, whereas the former shows *what* information shall be presented. It is proposed that the link between them can be the CFF taxonomy.

In this chapter, so far, we have seen how the outputs of Cognitive Work Analysis and principles of Ecological Interface Design are used in order to design the User Interface for the SPEEDD project’s Road Traffic Management use-case. Understanding operator/analyst tasks and information requirements (in terms of a

Socio-Technical Systems) allows us to develop concepts for User Interface designs which reflect the job of the operator.

The extended version of the UI Design methodology proposed by Upton and Doherty (2008) (Figure 1.19) was used and the changes in User Interfaces were discussed in terms of Content, Format and Form. The first step was to perform CWA. The output of this analysis are the user requirements and information requirements. Using the IR, a sketch of the UI layout was produced. The next step in the Upton and Doherty (2008) method was to perform Data Scale Analysis (categorising the data into nominal, ordinal, interval and ratio) and Visual Scale Matching (a method which specifies the most appropriate visual representation of information based on data type and dimensionality). Evaluation of the User Interfaces was performed, both in terms of the social domain (user interviews, subjective evaluation) and in terms of the technological domain (architectural constraints, component functionality). Changes that resulted from this evaluation and which lead to the final designs, were tracked and categorised in terms of Content, Format and Form (see APPENDIX I).

This approach is different from standard methods of UI design because it gives attention to the dimension of Format, in addition to Content and Form. Furthermore, CFF provides a taxonomy with the help of which the designer can keep track not only of the changes in design, but also of their magnitude and the impact of those changes on user behaviour (see next section). In terms of how this approach can inform HA design, the categorisation of UI design changes in terms of CFF can help the designer understand the effect of each UI element on user behaviour, as opposed to merely the effect of an entire UI. This could greatly simplify the design and evaluation process by providing more systematic approach to UI and UI component evaluation.

## 2.5 Experimental Evaluation

### 2.5.1 Introduction

SME evaluation of the User Interfaces helped solve some of the inconsistencies between requirements of the work environment and technology. They ensured that 'coherence' between the architecture and the work environment is achieved. However, issues relating to how the use of this technology will effect user performance should be explored through experimental evaluation.

The DIR-CE operators do not have a means of varying ramp metering rates. The ramps function on a schedule, i.e. their behaviour is scripted based on the day of the week and time. These scripts are fixed and operators have no means of fine-tuning ramp metering operation. However, because of the introduction of the automated ramp metering system, the goals (derived from CWA) are shared between the two entities – the operator and the automation. The task of spotting errors in the data and analysis outputs of the ramp metering system is added to the operator's role. This raises the question of how well the operators can detect cases where automation produces incorrect answers, either due to corrupt data, or due to errors in computation. This issue is explored in CHAPTER 3. Will increasing transparency of automation make the operator better at spotting errors (see CHAPTER 4)? A further issue that rises from the introduction of automation is how disagreements between the two agents should be managed. Who should have the final say (i.e. who should have the final authority) and should the agents be left to compute answers for a particular problem independently, or should they be allowed take each other's answer into account (CHAPTER 4)? While the previous question related to Format, the final question relates to the Form of the interface between user and automation. How will different UI designs (i.e. UIs that support integrative vs non-integrative processing) affect operator performance in establishing correct ramp metering rates and spotting automation errors (see CHAPTER 5)?

### 2.5.2 Quantifying the Impact of UI Design on Human Performance

This section presents an overview of the experiments which were run as part of this PhD. The motivations for the experiments along with a brief summary of their design are given and discussed. Furthermore, some measures for quantifying changes in UI design are proposed and discussed in terms of the dimensions of Content, Format and Form, the consideration of which is proposed as an extension of the existing methodology for UI Design and Evaluation.

Four experiments have been designed for the purpose of demonstrating the applicability and utility of considering the three dimensions of User Interfaces proposed in this PhD. In other words, they show the advantage of thinking about UIs in terms of Information Content, Format and Form. Again, to remind the reader, Content refers to the information requirements necessary for performing the tasks which should be supported by the UI. Format refers to the visual variables that define and constrain the interaction, negotiation and information exchange between the human operator and automation. Form is concerned with the way in which information is presented to the user.

The User Interface Design Methodology proposed by Upton and Doherty (2008) can be extended by the consideration of the Content/Format/Form (CFF) Taxonomy (Figure 1.19). Doing this, could bring benefits not only in terms of design, but also to the evaluation process. In terms of design, it provides a much clearly defined Design Space and a more easily trackable design evolution. In terms of evaluation, thanks to the ability to describe User Interfaces in a more specific manner, it allows for changes in user behaviour to be linked to changes in Interface design.

As we are investigating the effects of UI design in the context of Human-Machine, or Human-Automation Systems we cannot ignore the issue of automation reliability. Past research has shown that automation reliability influences user behaviour, potentially leading to over-reliance, complacency, boredom, skill loss, etc. (Bahner et al., 2008; Johnson et al., 2014; Parasuraman and Manzey, 2010; Parasuraman and Riley, 1997; Woods and Dekker, 2000). Therefore, as part of the experimental design, as well as the changes in UI components, one of the independent variables will be automation reliability. Apart from the avoidance of such effects as complacency and over-reliance, for example, this should also ensure that implicit attitudes towards automation (Merritt et al., 2013) are controlled for.

The experiments presented are developed in the context of the SPEEDD EU Project and are designed around one of the use-cases considered by the consortium, specifically, Traffic Management.

#### 2.5.2.1 *Objective Measures*

In order to quantify the effect of changing UI components on human behaviour we need to define measures of performance. For the purpose of the experiments presented in this thesis the following metrics have been chosen:

- decision correctness/accuracy: the percentage of correct decisions given by the operator
- decision time: the total time elapsed from the moment an issue is presented to the operator until a final decision is given
- decision match: percentage of decisions given by the user that are the same as the computer suggestion/answer – indirect measure of reliance
- solution source: whether the user sees himself as the source of the decision – indirect measure of perceived automation reliability
- dwell times (eye-tracking): measure of information extraction (see CHAPTER 3 section 3.1.2.4)
- switch count (eye-tracking): measure of information search
- % viewing time: proportion of time spent of a particular information source
- modals opened: average number of windows containing extra information opened – measure of information search

When considering the design of User Interfaces, there is virtually an unlimited space in which one operates and from which one can pick and choose UI components. This poses a problem for evaluation, as well, in that, not only information requirements need to be considered but also where this information is placed and in what manner it is shown to the user. Mapping the relationship between UI Design Dimensions and evaluation metrics allow for a better understanding of the where in the UI could lie the change which produced a certain effect in human behaviour. Table 2.3 below shows the relationship between these metrics and the dimensions of Content, Format and Form.

**Table 2.3 - Relationship between the defined metrics and UI Dimensions**

| Metric | Effects | | |
|---|---|---|---|
| | **Content** | **Format** | **Form** |
| **decision correctness** | The absence of an important information source may reduce the likelihood of making a correct decision. | Making it difficult to communicate an answer to the computer, incorrect placement of information sources (task proximity) could lead to an incorrect or less accurate result. For example, using a slider instead of a textual input to set a 3- or higher-digit number. | The way in which data are presented (i.e. relying on inappropriate type of perception, inappropriate display proximity, may make it impossible to extract the information, which negatively affects correctness |
| **decision time** | The absence of an important information source may increase the time of making a decision because the user is searching for the missing link. | Inappropriate input modes, or sub-optimal placement of information sources could increase decision time | The way in which data are presented may make it difficult to extract the information, which may increase decision time |
| **decision match** | Match (reliance on automation) may be higher due to impossibility or higher effort to make an informed decision in case of absence of an important information source. | Inappropriate input modes could increase reliance on automation, resulting in higher match levels | Improper presentation or sub-optimal placement of data may lead increase reliance on automation due to inability or difficulty to extract information |

| | | | |
|---|---|---|---|
| **solution source** | Too much information/ too high information content can lead the user to select the computer as a solution source more often. | Difficulty of interaction for the purpose of Task Delegation (for example), or low transparency can lead to the user selecting himself as the solution source more often. | Difficulty of extracting the information can lead the user to rely more on the automated response (select the computer as the solution source). |
| **dwell times** | Higher dwell times may hint to the user's expectation of extra information to be available in the area dwelled upon. | Higher dwell times on areas of the UI related to communication/interaction with the computer may be linked with an inappropriate mode of interaction | Higher dwell times on information sources may indicate improper presentation of data (i.e. using graphical representation where textual is more appropriate, for example, using the area of a circle to indicate a single number) |
| **switch count** | Higher switch count could be due to the search for unavailable information | Higher switch count may be due to the fact that the possibility of interaction with automation is unclear to the human. Perhaps the computer does not ask for assistance appropriately, using a mode that humans would understand. | Higher switch count may indicate that the way information is presented makes it hard to remember. Alternatively, it may mean that it is not obvious to the user how to decode the information from the chosen representation on the screen. |

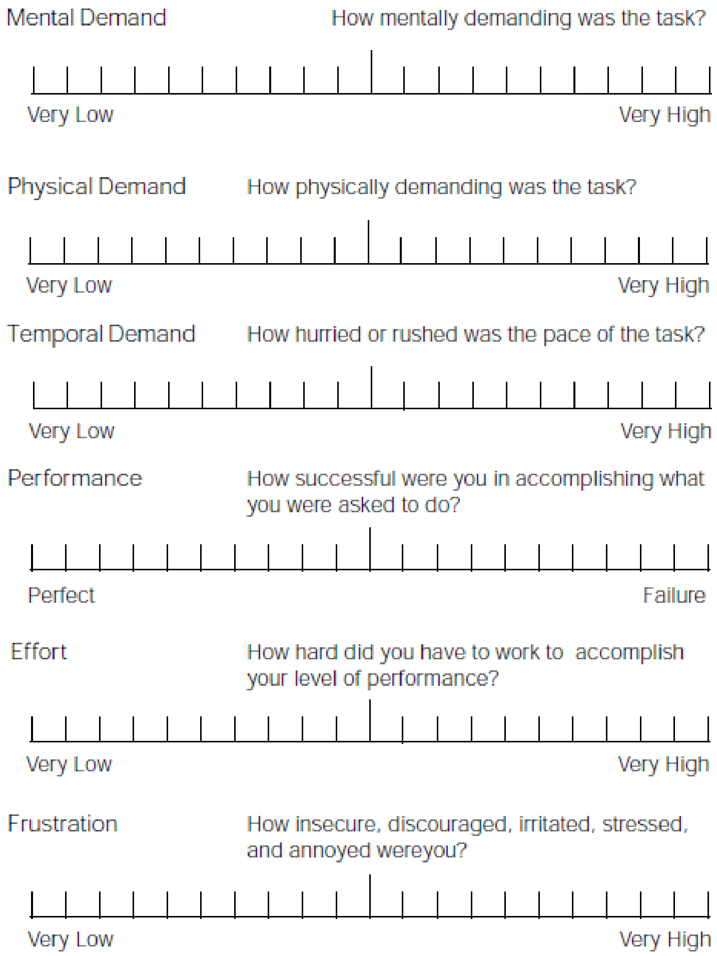
| | | | |
|---|---|---|---|
| | | | Higher switch count may indicate low task/display proximity |
| **% viewing time** | Higher % viewing time may indicate that some information the user is looking for is absent. | Higher values may indicate a difficulty in acting on data (control tasks). | Higher values may indicate hard to decode information.<br>Higher values may indicate important information that should be made available on screen (increase task proximity) |

Where two different user interfaces were employed (CHAPTER 5), user cognitive workload and subjective system usability were measured for each of the UIs and participants.

### *2.5.2.2 Subjective Measures*

#### 2.5.2.2.1 Workload

While there are many ways to measure the cognitive effort (workload) that people experience in performing mentally demanding tasks, a popular set of measures rely on participants providing subjective estimates of their workload. These measures can be surprising robust, sensitive to changes in demands and correlate well with physiological measures. One commonly used subjective workload measure is the NASA TLX (Task Load Index) (Hart and Staveland, 1988). This is a rating scale with six workload dimensions. It can be administered in either a computer or paper based format. The rating scales are presented as questions that the participants scores on a scale of 1 (low) to 20 (high). The questions relate to mental demand, physical demand, temporal demand, effort, performance and frustration (Figure 2.4).



Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

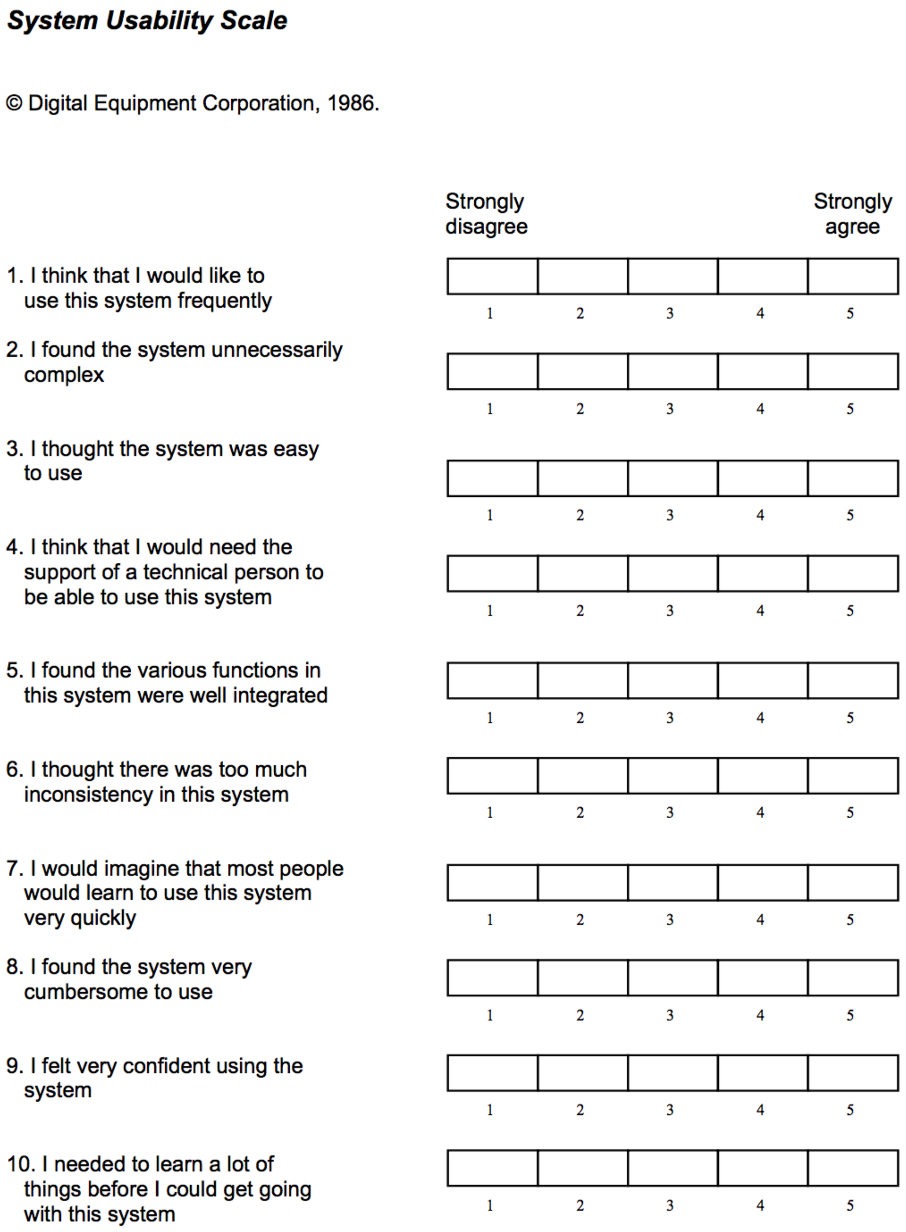
Very Low Very High

**Figure 2.4 - NASA TLX rating form**

**[http://humansystems.arc.nasa.gov/groups/tlx/paperpencil.html]**

### 2.5.2.2.2 Usability

Usability was measured using the System Usability Scale (Brooke, 1996). SUS is a ten-item Likert Scale (score from 1-5, ranging from “Strongly disagree” to “Strongly agree”) which is used to evaluate subjective usability. It assesses subjective effectiveness, efficiency of and user satisfaction with the system. A score above 68 indicates above-average usability, while anything below 68 is considered to be below average. A paper-based version of the questionnaire (which can be seen in Figure 2.5) was employed in the study presented in CHAPTER 5.



***System Usability Scale***

© Digital Equipment Corporation, 1986.

| | Strongly disagree | | | | Strongly agree |
|---|---|---|---|---|---|
| 1. I think that I would like to use this system frequently | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 2. I found the system unnecessarily complex | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 3. I thought the system was easy to use | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 4. I think that I would need the support of a technical person to be able to use this system | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 5. I found the various functions in this system were well integrated | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 6. I thought there was too much inconsistency in this system | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 7. I would imagine that most people would learn to use this system very quickly | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 8. I found the system very cumbersome to use | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 9. I felt very confident using the system | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| 10. I needed to learn a lot of things before I could get going with this system | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |

**Figure 2.5 - System Usability Scale [https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html]**

## 2.5.3 Summary of Experiments

### *2.5.3.1 Experiment 1 – Chapter 3: A baseline for Human-Automation Joint Decision Making and Implications for UI Design*

The first experiment (presented in CHAPTER 3) is designed around the Traffic Management use-case. It revolves around the task of ramp rate and density monitoring. This task was not one that operators were performing at the time, however, it is one that they were in the process of adopting and, at the time of writing this thesis, it is a task which they are currently performing in some form. Part of the objectives of the SPEEDD project involved the design of a system which could aid them in performing this task of ramp metering and the investigation of how it is used and how it could be improved.

This experiment employs the first prototype of the User Interface to the SPEEDD Traffic Management system. It was developed in order to get a baseline of how expert traffic managers would use the system and to compare their performance with that of non-expert university students. Moreover, the issue of automation error was explored as an incongruence between information sources and the computer suggestion. The hypothesis was that users would expect the computer to fail on the more complex task of deciding the correct change to a ramp rate, rather than on the simpler task of ensuring the correct data source is used in the computation. The effects were discussed in the context of the Proximity Compatibility Principle (Wickens and Carswell, 1995) and Joint Decision Making (Bahrami et al., 2010).

A version of this experiment has been published in the Annual Meeting of the Human Factors and Ergonomics Society (2015) (Morar et al., 2015b).

#### *2.5.3.2 Experiment 2 – Chapter 4: Format*

The second experiment (CHAPTER 4) is also designed around the Traffic Management use-case and the ramp metering task. The UI with which users interacted in this experiment was the UI of the first SPEEDD Traffic Management prototype. The study was motivated, in part, by the findings of the first experiment. Apart from the issue of varying computer reliability, two aspects related to Format are additionally investigated. More specifically, it has been tested whether showing the computer suggestion prior to the user's proposition of the course of action makes a difference to his performance. Furthermore, the experiment also investigates the issue of Transparency and the effect on the user's behaviour of requiring justifications for the proposed course of action. It was hypothesised that in the high reliability condition, the user would copy the computer's answer even in cases where it would be wrong. Moreover it was expected that the addition of the extra task of justifying the proposed action would increase user performance in terms of correct decisions. The results of this experiment are discussed in terms of Joint Decision

Making (Bahrami et al., 2010) and the notions of Authority, Responsibility, Transparency and Allocation of Function.

A version of this experiment has been published in the Annual Meeting of the Human Factors and Ergonomics Society (2017) (Natan Morar and Chris Baber, 2017).

#### *2.5.3.3 Experiment 3 – Chapter 5: Form and Format*

Experiment 3 (CHAPTER 5) is designed around the task of ramp metering in Traffic Management. Two versions of the last iteration of the SPEEDD system for Traffic Management were tested. The differences between them are given by the consideration of task proximity (PCP) in UI2 and, thus, can be confined to the dimension of Format. The hypothesis was that user decision time would be higher when using UI1 (lower degree of integration) and lower when using UI2 (higher degree of integration). Decision correctness was hypothesised to remain constant considering that the information available for decision making was constant. Differences in user behaviour when using the two UIs are discussed in terms of PCP and Joint Decision Making.

A version of this experiment has been accepted in IEEE Transactions on Human-Machine Systems (2018).

### 2.5.4 Summary of UI Integration and Evaluation Processes

The technical and user requirements inform, guide and constrain the design of the User Interfaces. In order to ensure that an optimal balance between the requirements of the technical and social domains is achieved in the instantiations of the UIs, the different designs need to be evaluated by Subject-Matter Experts (SMEs) in TM and technology. This ensures that the technical and user requirements and met and support each other, specifically, that discrepancies in terms of their goals are resolved and that they translate into a coherent User Interface. This means that after each User Interface version, evaluation is conducted and both user and technical requirements are updated.

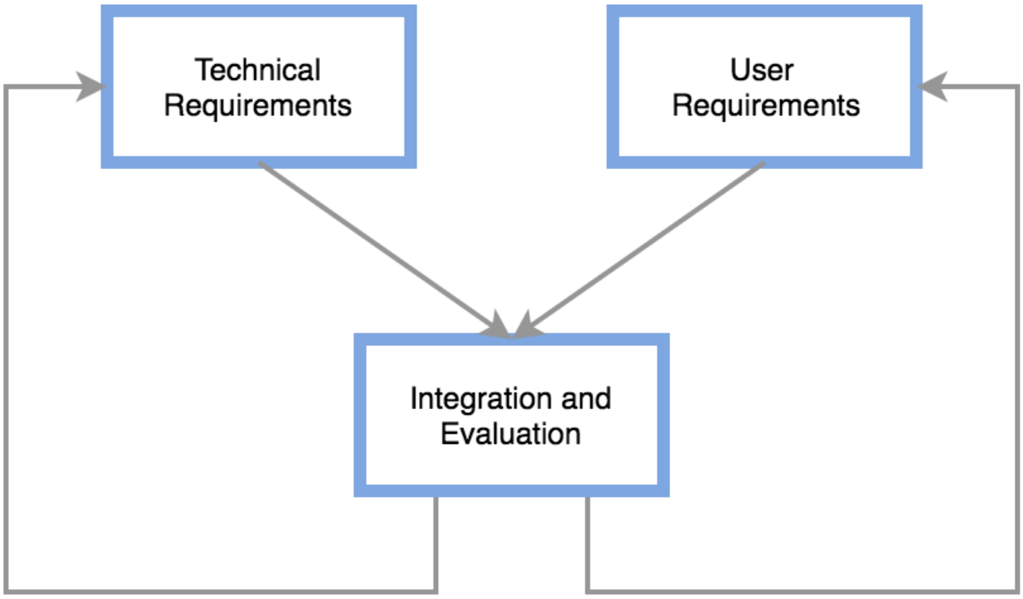


**Figure 2.6 - Integration and Evaluation Process**

Evaluation, in the scope of this thesis, has three parts to it. Firstly, interviews with SMEs were conducted for each of the UI versions. SMEs comprised of both technology experts and domain experts for both use-cases. Secondly, for major changes in UIs, the System Usability Scale (SUS) questionnaire was used as part of interviews with domain experts. Moreover, SMEs recommendations along with design considerations were subjected to experimental evaluation (see section titled Experimental Evaluation). This three-fold evaluation process informed the design of the SPEEDD User Interfaces. In APPENDIX I changes for each UI design iteration are shown.

# CHAPTER 3 A BASELINE FOR HUMAN-AUTOMATION DECISION MAKING AND IMPLICATIONS FOR UI DESIGN

Human monitoring of systems in which sensors provide data to automated decision support algorithms create interesting challenges for Human Factors. This study explores whether people are able to detect two types of automation failure: when decisions do not fit the data presented to the operator, and when data from different information sources do not agree. For those students that performed at a level of ≥ 97% correct (‘high performers), checking for both types of failure seemed easy. For those students that performed at a level of ≤ 95% correct (‘low performers’), checking for erroneous recommendations seems straightforward, but checking for information agreement seemed to be omitted. One suggestion is that the non-experts expended more effort on checking recommendation and ignored the need to check congruence across UI components. The implication is that the ‘worth’ of the displayed information for one task (decision check) outweighed its worth for the simpler task (congruence check) for the non-experts.

*A version of the experiment presented in this chapter has been published in [1] (Morar et al., 2015b). Additionally, section 3.7.2 from the SPEEDD Report D8.3 (Garin et al., 2015) is reproduced in this chapter.*

### 3.1.1 Summary

A simulated traffic management task was used to investigate the effect of automation failure on operator decision. 'Failure' could either arise from an erroneous recommendation or from disagreement between elements in the UI (i.e. errors in computations or corrupt/incomplete data). The study showed that participants were able to spot erroneous recommendations well, but non-experts that performed less well ('low performers') tended to miss disagreements between information sources. This leads to a higher incidence of false alarms in decisions for the non-experts. I propose that this indicates differences in the manner in which experts and non-experts might define the 'worth' of information in a UI component.

In this experiment, the relationship between the reliability of automated decision support and operator response is addressed. In order to articulate the research question, I was interested not only in varying reliability of the automation but also in modifying congruence of the visually displayed information. To this end, the experiment reported in this chapter is motivated by the Proximity Compatibility Principle (Carswell and Wickens, 1987; Carswell, 1992; Wickens and Carswell, 1995) and by the Bahrami et al. (2010) experiment of Joint Decision Making.

We manipulated the reliability of the displayed information (in terms of the 'automated' decision and in terms of the congruence between UI content) in order to see how this affected user's perception of the UI component diagnosticity. One might expect expert performance to involve an initial scan of the UI to ensure congruence and then a focus on the UI components (panels) which allow them to make a judgment on the correctness of the automated decision. In this strategy, all UI components have high worth in the first scan, but as this can be determined quickly, one would expect limited gaze duration on these. Two UI components (panels 1 and 3) have high worth on the second scan, therefore, one would expect higher gaze duration on these.

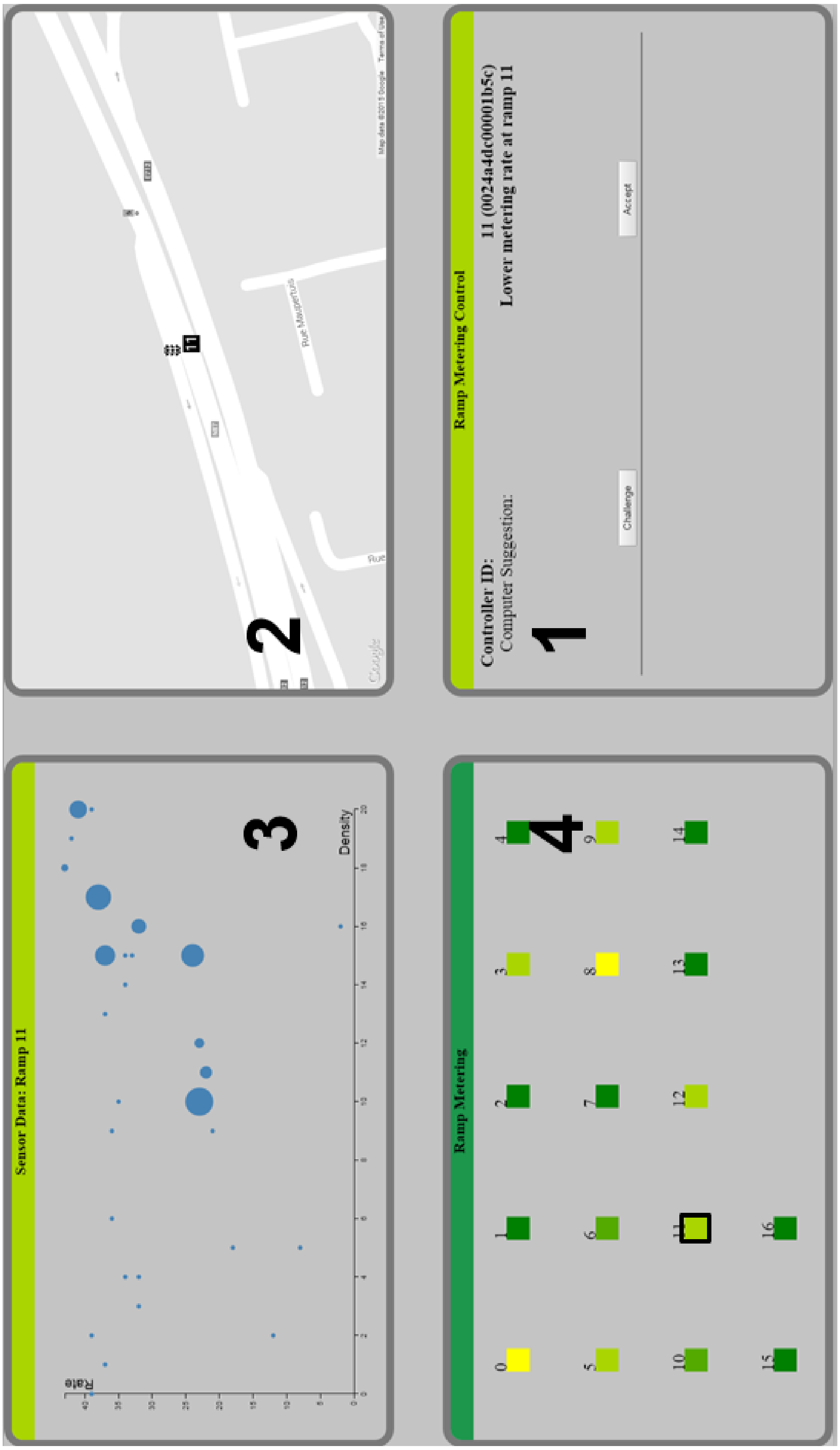
It is hypothesised that users would expect the computer to fail on the more complex task of deciding the correct change to ramp metering, rather than on the simpler task of ensuring the correct data source is used. Moreover, it is believed that expert performance would match student performance due to the novelty of this task for traffic managers.

## 3.1.2 Method

### *3.1.2.1 Scenario and User Interface*

This study employed a user interface (UI) mimicking a road traffic management task. The purpose of the UI was to enable an 'operator' to monitor, and potentially intervene in, computerised road traffic control decisions. These computerised decisions relate to ramp control (i.e., changing the rate at which traffic lights on a junction change in order to allow vehicles to join a main road). The scenario was derived from the operations of a real-world road traffic management facility (DIR Centre Est, Grenoble, France). This study was conducted as part of the SPEEDD European project, which aims to bring event forecasting to traffic management. The experiment presented in this chapter is a preliminary study which investigates how operators might respond to different levels of reliability in the system. As the ramp metering algorithms would run on data collected from sensors embedded in the road, there are potential problems which might arise from sensors failing, or data being lost or corrupted during transmission. While these problems might be dealt with by exception handling, it is possible that the recommendation could be based on erroneous data. Further, it is possible that the processing time of the algorithms could result in discrepancy between the recommendation and other UI components, i.e., the UI could show data for the ramp which is the current focus of the system, but the automated decision could present results for a different ramp. Thus, the operator would need to decide whether the recommendation related to the ramp being displayed and whether the computer suggestion was correct. For this experiment, the operator would either 'accept' the recommendation or 'challenge' (i.e., reject) it.

A custom UI was created in JavaScript. The UI contained four panels in an equally spaced 2 by 2 grid layout (Figure 3.1). Panel #1 (bottom-right) contained the operator response buttons 'challenge' and 'accept' as well as details on the computer suggestion regarding traffic light settings. Panel #2 (top-right) presented a crop of the road network surrounding a queried ramp based on google maps. Panel #3 (top-left) showed a historical data graph with density on the ramp on the x-axis (number of cars waiting to pass traffic light) and rate of the ramp (number of cars passing per second) on the y-axis. The most recent data points are represented by the largest bubbles (circles with the largest diameter). Panel #4 (bottom-left) presented a schematic grid of 17 ramp meters mimicking part of the instrumented road section. For instructions on how use the UI, see section 3.1.2.3 - Experimental design and data collection, paragraph 4 and 5.33



Controller ID: 11 (0024a4dc00001b5c)

Computer Suggestion: Lower metering rate at ramp 11

**Figure 3.1 - User interface developed for this study, consisting of four panels**

### *3.1.2.2 Participants*

An initial study involved 3 (male) experts in road traffic control, based in DIR-CE Grenoble, France. Following this, an experiment was conducted involving 17 second-year BEng students (mean age 24 years; 4 female and 13 male). All participants provided informed consent to participate in the study. This study was approved by the University of Birmingham Ethics Panel (Reference Number ERN_13-0997).

### *3.1.2.3 Experimental design and data collection*

The UI was presented on a 22” monitor (1080p resolution). Details of each response were captured for each trial were trial ID, trial start and end time (in ms computer time) and the participant response (challenge or accept). These data were stored locally in comma separated variable (csv) format. In between each trial a white screen with a timer was shown. The timer allowed for synchronisation with an eye-tracker which was used on a subset of the experimental participants (see below). Start time corresponded to the participant clicking the timer on the white screen, while stop time corresponded to the participant clicking on either the ‘challenge’ or the ‘accept’ button.

Following an explanation of the aims of the experiment and of the function of the UI components, participants performed two practice trials, after which they were given the opportunity to ask any clarifying questions. After the practice, participants performed the study, which consisted of 32 trials and took approximately 2-5 minutes to complete. Finally, participants were given a questionnaire to fill out.

Trials were separated into four scenarios based on the following characteristics of the displayed information and computer suggestion: 1) Information sources agree, suggestion correct (TT); 2) Information agree, suggestion incorrect (TF); 3) Information sources disagree, suggestion correct (FT); and 4) Information sources disagree, suggestion incorrect (FF). Each scenario was presented 8 times, and the 32 trials were presented in random order. Participants were asked to “accept” the computer suggestion if and only if information sources agreed and the computer suggestion was correct (TT). Hence, 3/4 of trials had to be challenged and 1/4 had to be accepted.

In order to determine whether the information sources agreed, the participants were instructed to check if all four regions of interest (ROIs) referred to the same ramp number. To determine whether the computer suggestion is correct or not, the participants were instructed to check the graph in ROI 3 (top-left in Figure 3.1). The presence of the biggest bubbles in the bottom-right quadrant of ROI 3 (low rate, high density) indicated that the rate must be increased. The presence of the biggest bubbles in the top-left quadrant of ROI 3 (high rate, low density) meant that the rate must be decreased. The presence of the biggest bubbles in either the bottom-left or the top-right quadrant (low density, low rate and high rate, high density, respectively) meant that the rate must remain unchanged. So, for the trial in Figure 3.1, the correct response would be to challenge.

The rules defined for this experiment are not necessarily the ones used in real-life traffic management situations, but have been simplified for the purpose of this task while still being illustrative of the real scenario. The ecological validity of the task was confirmed by asking road traffic experts from DIR-CE to perform the experiment. The three experts responded correctly to 97%, 100% and 97% of the trials. This expert performance data served as the threshold for splitting student participants into a ‘high-performing’ and ‘low-performing’ group.

#### *3.1.2.4 Eye tracking*

For a subset of seven participants (five from the ‘low-performing’ group and two from the ‘high-performing’ group), eye tracking data were collected. This could not be performed for all participants due to calibration issues when wearing corrective lenses. A Tobii Glasses v.1 head-mounted eye-tracker was used to record the point of gaze at 30 Hz while engaging in the task. Point of gaze was then automatically mapped to the four ROIs using custom Matlab (The MathWorks, USA) scripts. Mapping was performed based on the position of 16 infrared markers attached around the monitor at equally spaced intervals.

#### *3.1.2.5 Data analysis*

*Decision times.* For each participant and each trial, decision times were calculated as the difference between start and stop time.

##### *3.1.2.5.1* **Signal detection. Signal detection metrics were calculated based on the classification of each response as follows:**

True positive (TP): Response = accept, information = agree, suggestion = correct

False positive (FP): Response = accept, information = disagree and / or suggestion = incorrect

True negative (TN): Response = challenge, information = disagree and / or suggestion = incorrect

False negative (FN): Response = challenge, information = agree, suggestion = correct

*Gaze data.* From the eye tracking data, scan paths (sequence of attended ROIs) and dwell times (duration rested on each ROI per visit) were calculated. For each participant and trial, the number of attended ROIs and maximum dwell time per attended panel were calculated. For the data analysed above, the independent variable was trial category, and the dependent variables were the derived metrics.

### 3.1.3 Results

#### *3.1.3.1 Correctness of responses depending on scenario*

Of 17 student participants, one participant did not engage in the task as instructed due to a misunderstanding, a fact confirmed by a subsequent discussion; he was hence excluded as a non-representative outlier. The remaining 16 participants had performances ranging from 69% to 100 % of trials being assessed correctly. Three student participants had performances similar to those of the experts from DIR-CE (two with 100%, one with 97% correct trials), with the remaining 13 students showing performance < 95%. The performance level of the traffic managers in Grenoble was used as a threshold for partitioning the students into two groups: a 'high-performing' group (students that had performances comparable to those of experts) and a 'low-performing' group (students with lower performances compared to experts).

Decision time data and the number of correct responses were analysed for these groups for the four scenarios (TT – information sources agree, automation correct, TF – information sources agree, automation incorrect, FT – information sources disagree, automation correct, FF – information sources disagree, automation incorrect).

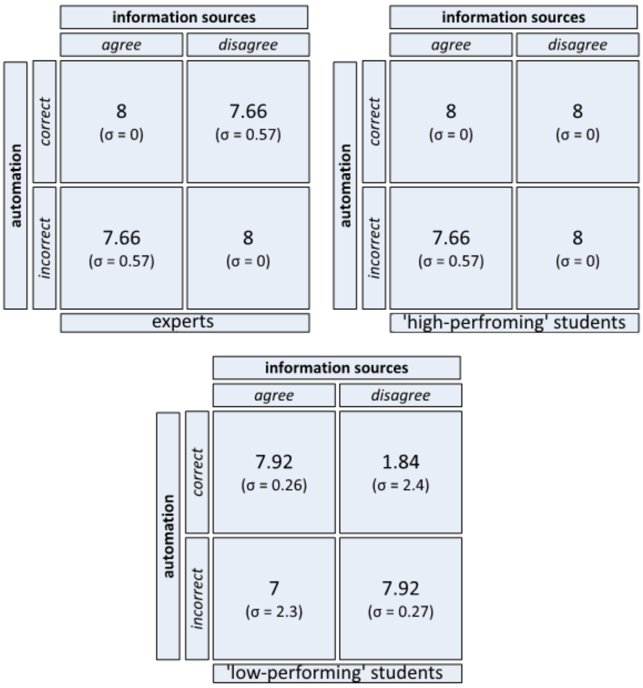


**Figure 3.2 - Mean correct responses in terms of scenario for each group**

Results for the different groups are shown in Figure 3.2. All groups easily identified cases where automation failed (an incorrect computer suggestion was given – TF and FF) or where automation was correct and information sourced agreed (TT). However, when information sources disagreed and automation was correct (FT), experts and ‘high-performing’ students responded correctly to all trials (σ = 0.57 and σ = 0, respectively), while ‘low-performing’ students responded correctly to only 1 out of 8 trials (σ = 2.4). Furthermore, the ‘low-performing’ student group presented a higher standard deviation (σ = 2.3) for the TF case.

### *3.1.3.2 Decision times*

Decision times per trial ranged from 1.4 s to 23.5 s across participants. The median decision time per participant across all trials ranged from 2.8 s to 10.3 s (median ± IQR 4.6 ± 1.9 s).

Decision times categorised by response category. To examine whether decision times varied between different response categories (Figure 3.3), the median decision time was calculated for each participant: a) in terms of four categories for all trials classified as TP, FP, TN and FN, b) in terms of response type and c) in terms of the 4 scenarios specified (see Method). A Kruskal-Wallis test was carried out to examine whether decision times differed between purely response types (challenge/accept) and between experimental design categories. There was no significant difference in decision time between response type (p = 0.931) or experimental design categories (p = 0.674). Furthermore, decision times seemed to be similar for all signal detection categories.

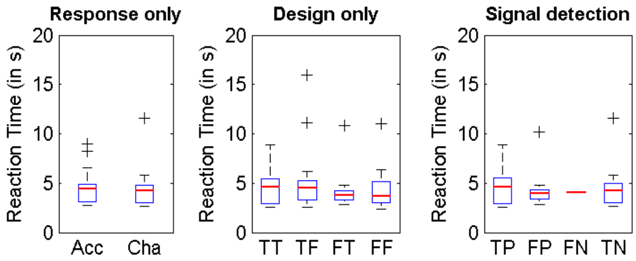


**Figure 3.3 - Boxplots for decision times different response categories**

### *3.1.3.3 Changes in decision times with elapsed trial.*

To examine whether there was a systematic trend for decision times to change as a function of elapsed trial, linear regression was performed for each participant with trial number as the independent and decision time as the dependent variable.

Results depended on the participant: on one hand, there was a significant linear association between decision time and elapsed trial number for 5 participants, albeit very shallow fitted slopes (range of fitted slopes: -0.08 to 0.06, range for $R^2$: 0.38 to 0.99; range for *p*: < 0.001 to 0.026). On the other hand, there was no significant association for 12 participants (range for $R^2$: 0.00 to 0.18; range for *p*: 0.059 to 0.445).

### *3.1.3.4 Gaze data*

Dwell times. The median dwell time was calculated for each participant and each panel for all trials classified as TP, FP, TN and FN. Results are shown in Figure 3.4, a. The two groups show similar median dwell times for ROIs 2, 3 and 4, however, on ROI 1, the ‘high-performing’ group dwells for a median of 0.8s, while the ‘low-performing’ group a median of 1.3s.

Number of attended panels. To examine whether the number of attended panels varied between different response categories, the median number of attended panels was calculated for each participant for all trials classified as TP, FP, TN and FN. Results are shown in Figure 3.4, b. There was not much difference between groups in terms of number of attended ROIs, medians ranging from 3.5 to 4.

Switch count. To examine whether the number of switches varied between different response categories, the median number of switches was calculated for each participant for all trials classified as TP, FP, TN and FN. Results are shown in Figure 3.4, c. While the 'low-performing' group switched between panels an average number of close to 5 times for each signal detection category, the 'high-performing' group switched between a median of 7 panels for the trials labelled TP and significantly lower (2) for those labelled FP.

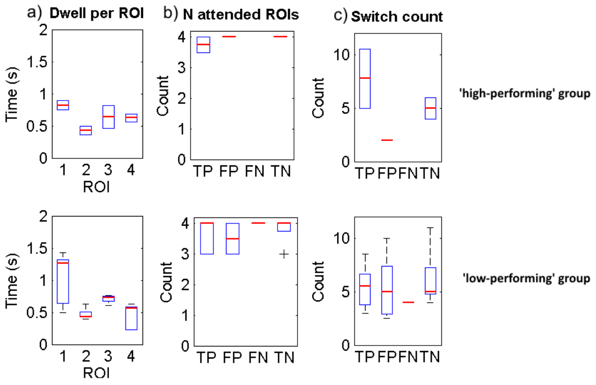


**Figure 3.4 - Dwell times per ROI (a), number of attended panels per response category (b) and switch count per response category(c) for both student groups**

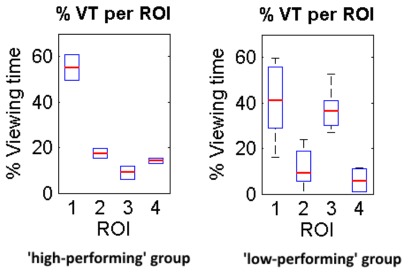


**Figure 3.5 - Percentage View Time per region of interest (ROI) for the 'high-performing' and 'low-performing' student groups**

### 3.1.4 Discussion

The maximum dwell time per attended panel shows that the maximum time spent on a panel was registered for ROI1 for both groups. However, this metric does not say what happened across all trials and participants, but that possibly the information in ROI1 might have been harder to decode at first for most participants. This offers some input in terms of UI Form (see Table 2.3). Perhaps presenting the computer suggestion in a more graphical way (e.g. as arrows pointing upwards or downwards for suggesting an increase or decrease, respectively and a horizontal line suggesting that the rate should not be changed) could make it faster to decode than in textual form.

The strategy adopted by the two groups was different. Figure 3.4 c suggests that the 'low-performing' group applied more or less the same strategy across all trials, the switch count being constant (5) for all signal detection categories. The 'high-performing' group, however, seems to have adapted their strategy depending on the trial at hand. They switched panels a larger number of times (7) for the TP case and a much smaller number of times for the FP case (2). There was no difference between the groups in terms of the median number of attended panels, both looking at all ROIs, for most trials. However, this does not imply that they have extracted and used the information present in the panels in their decision, but that their gaze simply passed over them. In terms of UI design, a higher switch count may indicate that the way in which information is presented (Form) may be hard to remember, or, alternatively, that the information presented in different panels require to be processed together and, thus, could benefit from integration (Format). Nevertheless, in order to gain an understanding of what information the different groups used, we look at % Viewing time per ROI.

While the task did not present a challenge to Subject Matter Experts, we note that the 'low-performing' students exhibited an interesting pattern in their response. Considering the results, we assume that all participants were able to use the graph component (top left of the screen, ROI 3) to apply the rules defined. Hence they correctly determined whether the computer suggestion was correct or false. However, the 'low-performing' students were confused by the FT condition, in which the automation was correct, but the information on the UI components disagreed). This suggests that they were not checking for component congruence, which was supported by eye-tracking data: the ‘high-performing’ student group attributed a similar percentage viewing time to ROI 2 to 4 (Figure 3.5). In contrast, the ‘low-performing’ group tended to spend a much larger proportion of their time looking at ROI 3 than ROIs 2 and 4 (which are used only to determine UI component congruence). ‘High-performing’ students seem to exhibit the same perceived importance for ROIs 2, 3 and 4. The discrepancy in percentage view time between ROI 1 and the other 3 ROIs is likely an artefact of the Form in which information is presented in ROI 1 (i.e. purely textual). Alternatively, this behaviour may be explained by the fact that this ROI was both the place where the computer recommendation was given and where the user had to give the final answer, thus having to return to this window after making each decision.

The low performance of 13 out of the 16 student participants in the experiment could be explained by the findings of the Bahrami et al. (2010) study. The considerably low sensitivity (i.e. reliability) of the automation may have been the reason for the poor accuracy of decisions of the low performing group. Bahrami explains that the mismatch in sensitivity (reliability) between the dyad members (i.e. human and automation) leads to worse joint performance than if the member showing the highest sensitivity were to approach the task alone. A criticism that could be brought to this is the inability of dyad members to communicate. However, more recent research found that interaction is not mandatory for the replication of the results of the Bahrami et. al. study (Bang et al., 2014; Koriat, 2012). In terms of how this finding informs the design of Human-Automation Systems, it might be better for the computer suggestion to be hidden in the case of low reliability (or, when the computer has low confidence in its decision) and prompt the user to give his response first. This may allow for a more careful consideration of data in lieu of the influence of the computer recommendation. Perhaps, after the user inputs his response, the computer suggestion could also be shown allowing for comparison. The experiment presented in CHAPTER 4 further looks into this matter.

In terms of task proximity, while all participants were presented with the same information, the ‘low-performing’ students were not able to judge the ‘worth’ of the UI components for congruence checking and focused their attention on the automation validation aspect of the task. It is possible that this might be an effect akin to change blindness in which relevant information is not attended to on the assumption that it is ‘given’ and does not require checking (Simons and Levin, 1997). Alternatively, a phenomenon termed ‘satisfaction of search’ is known from the medical literature, where diagnosticians terminate visual search after finding the first sign of pathology (Berbaum et al., 1994, 1990; Samuel et al., 1995). Similarly, participants may have terminated their search after completing the visual evaluation that computer suggestion and information held in ROI 3 agreed. Perhaps, the similar % Viewing time of ROIs 2, 3 and 4 of the ‘high-performing’ group indicates that these information bits have a similar perceived importance and that users could benefit from their integration.

In terms of UI design, it is important to consider not only how information can be presented to highlight its ‘worth’ but also how people might seek to extract information from UI components. The ‘low-performing’ group may have expected the automation to fail on tasks perceived as being more complex, leading to their attention being mainly focused on validating the computer suggestion. The findings presented in this chapter underline the importance of cueing operators using decision support software to make sure they are aware of the context (system state) in which they make decisions. One way to achieve this could be to prompt users to acknowledge if some components show different views (i.e. show data related to different ramps). Alternatively, placing together the UI components requiring integrative processing might increase overall decision correctness, by making it easier to spot inconsistencies in the input data.

# CHAPTER 4 FORMAT

In this chapter, automation bias in terms of joint decision making between humans and automation is explored. In an experiment, participants made decisions, and indicated the reason for their decisions, in a road traffic monitoring task with the aid of automation of varying reliability (i.e., 25% or 81%). Reliability level had a clear impact on the user's behaviour: at low reliability, participants ignored automated suggestions and relied on their own decision making, whereas in the high reliability condition, participants tended to accept the automation suggestion (even if this was incorrect). Overall, performance is higher as a result of the human intervention that would be expected from automation alone, i.e., accuracy is in the region of 87-96% on all conditions. Performance is affected by how much detail they are required to provide, but not by the order in which the human and automation give their answers. These results are considered in terms of a theory of joint decision making.

*A version of the experiment presented in this chapter has been published in (Natan Morar and Chris Baber, 2017).*

### 4.1.1 Introduction

In the previous experiment (CHAPTER 3), computer reliability was kept constant (at 25%). This did not allow the control of users' implicit attitudes towards automation (Merritt et al., 2013) (such as automation bias), which could lead to such effects as complacency conformance and boredom (Lee and See, 2004; Parasuraman and Riley, 1997). Varying the computer reliability will allow for testing whether the user is able to judge automation usefulness. Moreover, it will uncover any attitudes towards automation that existed prior to the experiment. This could be inferred from a relatively constant conformance to the computer suggestion or, conversely, a constant disregard of the computer recommendation.

#### *4.1.1.1 Joint Decision Making*

In a classic study of joint decision making, Bahrami et al. (2010) demonstrate the importance of information sharing and (more importantly) of weighting information by its reliability. In these experiments, participants were presented with a visual detection task (in which they had to spot a target against a background of distractors). For each decision, participants worked individually, then they shared the decision with another person, and then the two participants discussed the decision until they reached consensus. These experiments show that when two (human) decision makers have similar levels of reliability (or sensitivity) in a detection task, their combined performance is superior to that of either individual, providing they are able to communicate freely and indicate their confidence in their own decisions. However, when either person has lower reliability, then performance is much worse than that of either individual. The model that Bahrami et al. (2010) propose assumes that the pair of decision makers are Bayes optimal and exchange their level of confidence in their detection decisions. In a recent development of this approach, Koriat (2012) removed the requirement to discuss the decision, using the result of the most confident member of a pair makers (where confidence was measured using self-report). In this case, the initial findings of Bahrami et al. (2010) were replicated (i.e., relying on the performance of the most confident member of the pair leads to consistently superior performance), even in the absence of discussion. If the most confident member of the pair was, however, wrong, then performance deteriorates (because the least confident member accepts their partner's recommendation). This suggests that while Bahrami et al. (2010) saw their results, in part, as arising from the development of consensus through discussion, Koriat (2012) has demonstrated that the relationship between the report of an answer and the confidence of that person reporting the answer is key. In an interesting development of this work, Bang et al. (2014) show that the approach advocated by Koriat (2012) works well when participants are of 'nearly equal reliability' but when there are discrepancies then it is important to allow interaction. This seems to suggest that the approach taken needs to be adapted to suit differences in confidence and raises some questions about how human participants are able to evaluate the credibility of each other's rating of confidence and how should they relate this to actual performance.

Assume that the pair consists of a human and an automated recommender system. I am not aware that the 'optimally interacting' research area has considered what happens when one of a pair of decision makers is a computer. If either the computer or the human partner in this decision making dyad exhibits different reliability to their partner, will joint performance deteriorate (as shown in the Bahrami et al. (2010) and the Koriat (2012) studies)?

#### *4.1.1.2 Transparency and Recommender systems*

Recommender systems are software tools which aid people in the process of decision-making by providing suggestions for a specific action course or proposing solutions for an arisen problem (Ricci et al., 2011). The general idea is that automated reasoning on the data computes an answer and displays it to the user, for example, in the form of a recommendation for an action to be taken, or in the form of a detected event. 'Transparency' is a defining factor of a 'good' recommender system (Tintarev and Masthoff, 2012), i.e., the extent to which the computational process behind the recommendation is visible and clear to the human. It has been shown that increasing transparency of recommender systems, that is, making explanations available to the user along with recommendations improves decision performance . One way in which transparency can be increased is by presenting the confidence level associated with the computer suggestion. However, there are others ways in which computational processes can be made transparent to the human, for example by having the computer share its reasoning or justification for the presented recommendation. Based on previous research, one would expect that this increase in transparency would lead to better performance in terms of decision correctness, however it may also lead to an increased decision time due to the extra information the human needs to attend to.

#### *4.1.1.3 Automation Reliability and Human Performance*

Measuring overall performance while using varying levels of computer reliability, will enable testing of the Bahrami et al. (2010) conclusion that joint performance is better than that of just the highest performing individual, provided that they have similar sensitivities. In the case of automation, sensitivity is represented by its reliability level. From the previous experiment (CHAPTER 3), one would expect that expert performance is somewhere between 90-100%. If Bahrami's findings apply also in the case of a dyad composed of a human and a computer, then in the high reliability condition, one would see performances close to expert levels (95-100%) and in low reliability level performances would be lower than 95%.

Wickens and Dixon (2007) conclude their review of the impact of automation reliability on human DM with the finding that human performance with automation that is less than 70% reliable was often worse than having no automation, especially under conditions of high operator workload. For this experiment, two reliability levels were chosen: one above this margin at 81% and one much below it, at 25% so that a clear baseline for unreliable automation could be established.

Apart from varying reliability, two other independent variables have been introduced: turn and task. Task refers to whether the user and computer were required to give justifications for their answers in addition to their response, or not. Turn refers to the order in which the dyad members are required to give their response. This translates to trials where the user has to give his response prior to seeing the computer recommendation and trials where the computer recommendation appears before the user is prompted to give his response. An additional stage in decision making was added: the ability for the user to finally pick between the response he has given or the computer recommendation. Apart from providing the opportunity to test for conformance, turn allows us to see whether the user was able to adequately gauge automation reliability. For example, a highly conformant user would be expected to simply copy the computer's suggestion and to select its answer as the final response.

The inclusion of task as an independent variable in the experimental design was motivated by two considerations. First, having the computer include its reasoning could stand for an increased transparency, which was seen to influence trust in automation (Sinha and Swearingen, 2002; Tintarev and Masthoff, 2012, 2007), thus leading to potentially higher reliance. Secondly, form-filling was set up to simulate communication between the computer and the human as in the Bahrami et al. (2010) experiment. Moreover, form-filling (or, reporting) was a main task of the DIR-CE traffic managers.

The way the experimental design translated into changes in the design of the UIs is discussed in terms of the CFF (Content Format Form) Taxonomy.

We hypothesise that in the high reliability condition, users would copy the computer's answer even in cases where it would be wrong. Moreover it is expected that the addition of the extra task of justifying the proposed action would increase user performance in terms of correct decisions, in addition to increasing decision time.

## 4.1.2 Method

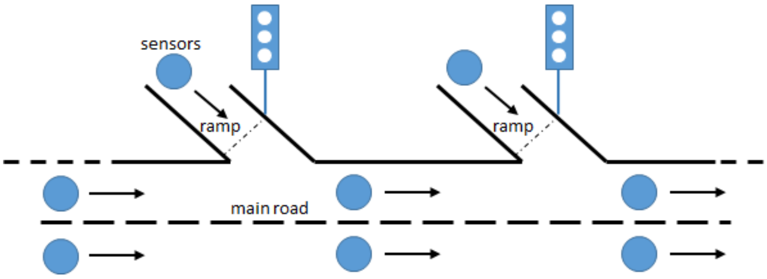


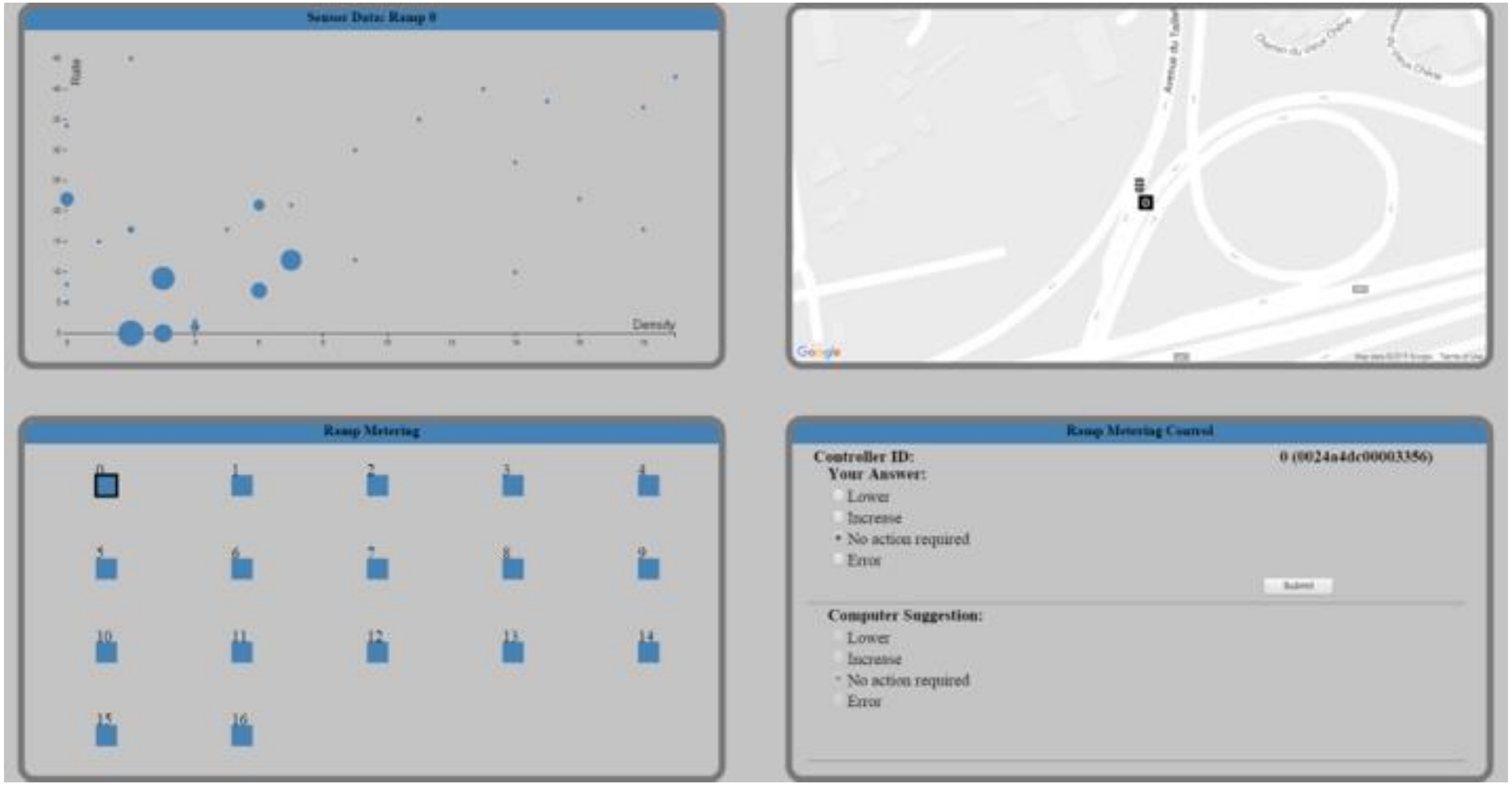
**Figure 4.1 - Experiment Scenario**

### *4.1.2.1 Experimental Task: Simulated Traffic Ramp Metering*

The experiment is based on Traffic Management operations and implements a scenario in which the human-automation system is monitoring the ramp rate (rate of change of traffic lights on inbound ramps). This is illustrated by Figure 4.1. In order to keep the task tractable in the laboratory setting, two simplifications were made: a) traffic densities in the main road are not considered and b) ramp rate refers to the number of cars that are able to enter the main road from the respective ramp.

The traffic management task was performed under different conditions of automated support. The reliability of the automated support was either low (25% correct) or high (81% correct). Reliability was defined by two factors: (i.) whether the identity of the ‘ramp’ was the same in all windows (to simulate a sensor malfunction), or (ii.) whether the computer suggestion was correct or not (to simulate a reasoning failure). Ideally, participants should recognise that one of these failures has occurred and respond accordingly.

The task was also performed under different conditions of operator activity. In some trials, the participants were required to select a decision option (Figure 4.2), and in other trials the participant also had to select a reason for a decision (Figure 4.3). The automated support would display its suggested decision and reason either before the user response, i.e., the computer suggestion field would be filled in before the user made a response, or this would appear after the user made a response. The idea was to simulate an automated suggestion and to see if this affected the user’s response. In this instance, the provision of a reason for the decision is intended to simulate the sharing of information in the Bahrami et al. (2010) study.



**Controller ID:** 0 (0024a4dc00003356)

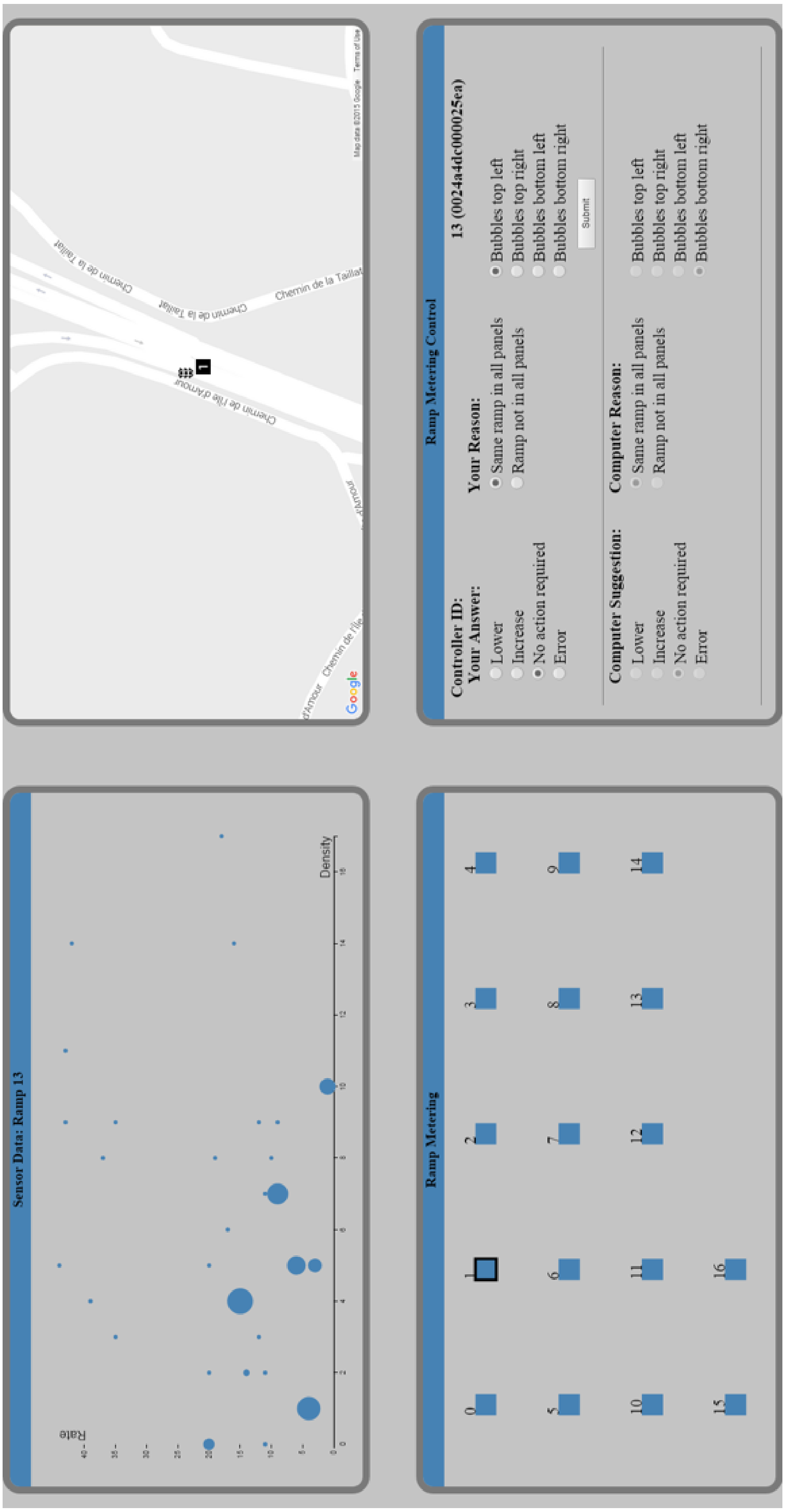
**Your Answer:**

○ Lower
○ Increase
● No action required
○ Error

**Computer Suggestion:**

○ Lower
○ Increase
● No action required
○ Error

**Figure 4.2 - User Interface for Decision Only condition**



Controller ID: 13 (0024a4dc000025ea)

Your Answer:
○ Lower
○ Increase
● No action required
○ Error

Your Reason:
● Same ramp in all panels
○ Ramp not in all panels

● Bubbles top left
○ Bubbles top right
○ Bubbles bottom left
○ Bubbles bottom right

Computer Suggestion:
○ Lower
○ Increase
● No action required
○ Error

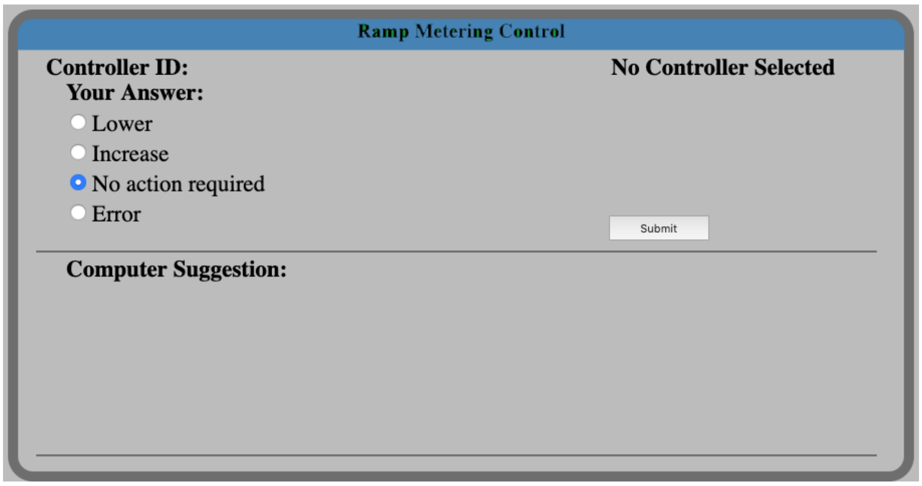
Computer Reason:
● Same ramp in all panels
○ Ramp not in all panels

○ Bubbles top left
○ Bubbles top right
○ Bubbles bottom left
● Bubbles bottom right

**Figure 4.3 - User Interface for Complete Form (Explanation) and Make Decision condition**

### 4.1.2.2 *User Interface and Interaction*

Two different versions of the user interface were employed in this study, the distinguishing factor between them being the window in bottom right corner of the screen (Figure 4.3). In this window users can see the computer's recommendation and submit their decision. The window on the top right is a road map that shows the ramp, the flow and density data for a ramp (top left corner) and the selected ramp (bottom left). In both situations there are two possibilities; first, the user needs to respond before the computer gives its recommendation, followed by which a final decision is required to be made by the user, of whether to stick with his own answer or follow the computers suggestion. Alternatively, the computer recommendation is presented before the user gives a response. In this case, after users enter their own response, they make a final decision. The UI was presented on a 22" screen of 1920 x 1080 resolution. Interaction with the UI consisted of selecting radio buttons corresponding to response and clicking the 'Submit' (Figure 4.4). For the 'form filling' condition, participants also had to complete the field for information source. No performance feedback was given.



Ramp Metering Control

Controller ID: No Controller Selected

Your Answer:

○ Lower
○ Increase
● No action required
○ Error

Computer Suggestion:

**Figure 4.4 - Ramp Metering Control**

### 4.1.2.3 *Participants*

23 Undergraduate students (18 male; 5 female) with no prior experience of the task or the user interface design, were recruited to participate in this study. All participants were given two practice trials prior to the experiment so that one could

assume that they were competent in the task demands. Participation was for course credit.

#### *4.1.2.4 Procedure*

This study was approved by the University of Birmingham Ethics Panel (Reference Number ERN_13-0997). Participation was through a purpose-built web interface. All participants attempted the study at the same time in a computer laboratory and completed it over the course of an hour. They were not allowed to speak or interact with each other in any way. Every computer in the laboratory was connected to a server running on the university intranet. Students were given all necessary instructions for completing the experiment in writing, through the web interface with the possibility of asking clarifying questions of the supervising staff.

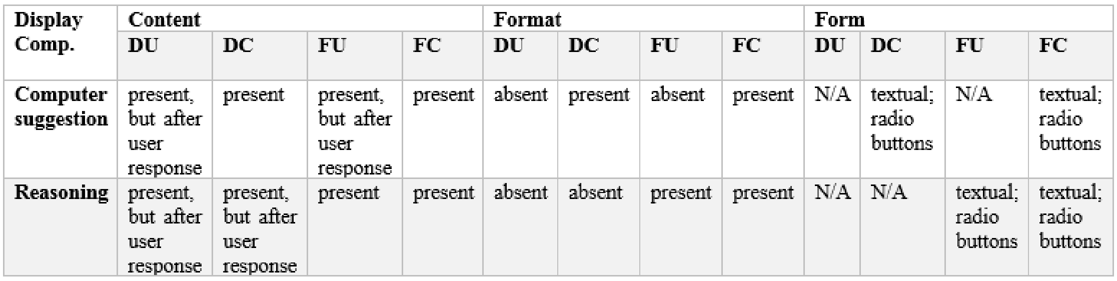
Participants completed trials in both low and high reliability conditions (counter balanced across participants) and completed tasks with all combinations of task and automated support. The total number of trials was 128, 16 in each of the following conditions: HDU, HDC, HFU, HFC, LDU, LDC, LFU, LFC: H or L refers to high or low reliability; D corresponds to decision only, F corresponds to decision plus explanation; U corresponds to cases where the user has to respond first, before the computer recommendation is revealed, while in the C cases, the computer recommendation is shown first. The total number of trials was split into four groups (Table 4.1). In order to control for learning effects, no performance feedback was given.

**Table 4.1 - Trial distribution in each reliability condition**

| **Low reliability** | **no sensor malfunction** | **sensor malfunction** |
|---|---|---|
| **computer correct** | 25% | 25% |
| **computer incorrect** | 25% | 25% |
| **High reliability** | **no sensor malfunction** | **sensor malfunction** |
| **computer correct** | 81.25% | 6.25% |
| **computer incorrect** | 6.25% | 6.25% |

### 4.1.2.5 *Mapping Experimental Design to the Content Format Form Taxonomy*

**Table 4.2 - Experimental Design in terms of Content Format Form**



| Display Comp. | Content | | | | Format | | | | Form | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DU | DC | FU | FC | DU | DC | FU | FC | DU | DC | FU | FC |
| **Computer suggestion** | present, but after user response | present | present, but after user response | present | absent | present | absent | present | N/A | textual; radio buttons | N/A | textual; radio buttons |
| **Reasoning** | present, but after user response | present, but after user response | present | present | absent | absent | present | present | N/A | N/A | textual; radio buttons | textual; radio buttons |

Whether or not the computer gives its response first is an aspect which relates the Format dimension (see CHAPTER 1). This is because it relates to the mode of interaction of the user with the automation, i.e. the factor which triggers the human response. In the case where the user is required to respond first, the triggering factor could be considered the change in the situation, i.e. a change in the information sources. When the computer answers first, an additional factor which could trigger the human response could be the appearance of a computer suggestion. The user might prefer to use the change/appearance on the screen of a computer suggestion as an indication that his input is required over the change in data displayed in the information sources. A change in the current situation might be harder to diagnose as it requires the monitoring of at least two information sources (out of Sensor Data, map and Ramp Metering windows) in parallel. The diagnosticity of the appearance of a computer suggestion as a trigger for action might be much higher in the experimental block where the computer gives its answer first, as it requires the monitoring of just the Ramp Metering Control window.

One may argue that turn has a Content aspect to it, as well. The presence of the computer suggestion gives the user an extra piece of information to consider prior to making a decision, which may result in a higher accuracy of user decision, with a potential higher time cost attached to it. We would argue, however, that turn is rather related to Format for two reasons: i) the computer recommendation is a piece of information that the user has access to in every trial, and ii) turn is related to the mode of interaction, as in the C blocks, the user has to give his response in order to reveal the computer's recommendation.

Reasoning is related to both the Content and the Format dimensions. In terms of Content, the presence of computer reasoning might be interpreted by the user as an additional information source that requires attending to. In terms of Format, this increase in transparency could aid in diagnosing of computer errors by highlighting discrepancies in data presented in the information sources and computer inputs, thus ensuring appropriate reliance. Nevertheless, in both situations the advantage may come with an associated time cost, especially considering that the user is required to give his reasoning, as well. A potential pitfall is that the higher workload that the human operator is faced with in the reasoning condition could cause him copying the computer answer even when it is incorrect, thus leading to complacency and high conformance.

Form is kept constant across all display components.

#### *4.1.2.6 Data Collection and Pre-processing*

The following data were recorded for each user response: participant ID, experimental condition, event number, event time, response time, computer correctness, response correctness and final response. Four out of the 23 participants were excluded from the analysis because they did not attempt all experimental conditions. Before analysis, thresholding was performed on the data for each participant. The cut-off point was set in terms of decision time at *average + 1 x standard deviation*. This resulted in an average of 12.24% and 11.35% trials being filtered out from the low reliability and high reliability conditions, respectively.

The performance of the participants was tested for normality using Shapiro-Wilk. In case of normally distributed data, a repeated measures Analysis of Variance (ANOVA) was performed, otherwise Friedman and subsequent Wilcoxon (Bonferroni adjustment) tests were run for decision time, % correct responses, solution source (i.e., did the participant believe that the solution came from them or from the computer) and match (times when the user's given response matched that of the computer).

## 4.1.3 Results

### *4.1.3.1 Decision Time*

Decision time is defined as the time elapsed from when a new trial is shown to the user up until he makes the final decision, of whether he chooses the computer's suggestion or his own answer as the final solution.

The Shapiro-Wilk test showed that decision time data are not normally distributed. Friedman indicated a significant difference ($\chi^2(7) = 73$, $p < 0.0001$). Subsequent Signed-Wilcoxon tests were run to identify which independent variable caused the this effect. No effect of reliability was found, however there was an effect of task ($Z = -7.56$; $p < 0.001$) (median D = 5.98 s, median F = 9.5 s). This is illustrated by the data in Table 4.3 and Figure 4.5. No other effect was found (Figure 4.6).

**Table 4.3 - Average Decision Times across Conditions**

| | 25% reliability (L) | | 81% reliability (H) | |
|---|---|---|---|---|
| turn \ task | Decision (D) | Decision + Reasoning (F) | Decision (D) | Decision + Reasoning (F) |
| User first (U) | 6.5s | 10.8s | 7.6s | 10.6s |
| Computer first (C) | 6.5s | 10.0s | 6.6s | 10.0s |

Given no effect of reliability, it was decided to split the data into the Low and High reliability conditions, to see if there were differences within these conditions. The Signed-Wilcoxon test showed no significant effect of turn. However, there were significant effects of task in both the low ($Z = -5.35$; $p < 0.001$) (median low decision (LD) = 5.88 s, median low form (LF) = 9.76 s) and high ($Z = -5.37$; $p < 0.001$) (median high decision (HD) = 6.03 s, median high form (HF) = 9.25 s) reliability conditions.

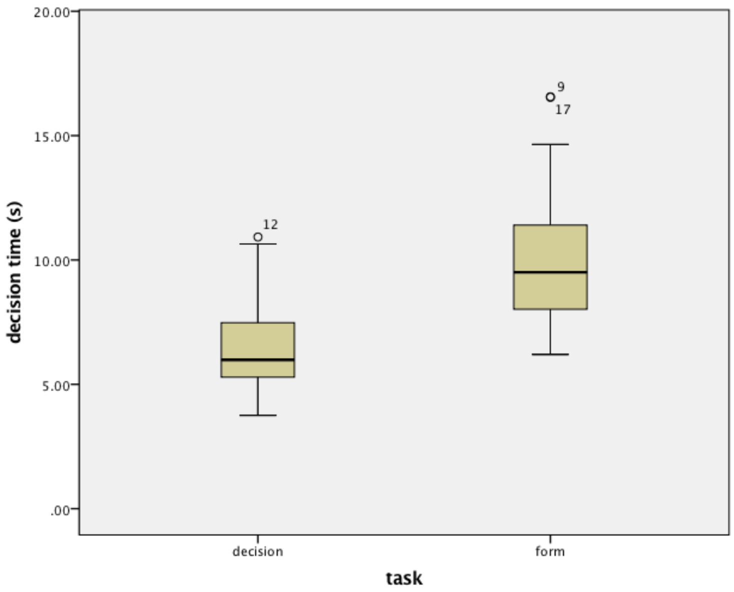


**Figure 4.5 - Decision time in terms of task**

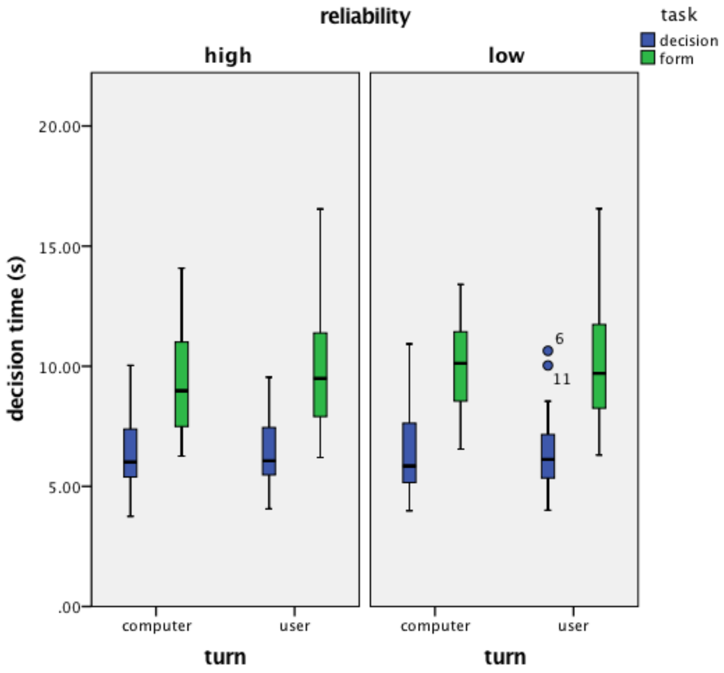


**Figure 4.6 - Decision Time**

### *4.1.3.2 Percentage Correct Responses*

Percentage correct responses is defined as the proportion of responses that were right, in terms of the rules defined in the experiment, out of total responses given by each user.

Normality tests revealed that data were not normally distributed. A run of the Friedman test showed no significant effects. Because there were no significant effects of reliability on correct responses, data within each reliability situation were looked at separately. No statistically significant results were found for the low reliability condition, however, for the high reliability condition the results the of Friedman test were significant ($\chi^2(3) = 9.19$, p = 0.027). The signed Wilcoxon found an effect of task on correctness in the high reliability condition (Z = -2.411; p = 0.016) (median high form (HF) = median high decision (HD) = 100%). A further run of Wilcoxon revealed a difference between the HFC and HDC condition (high reliability-form-computer first and high reliability-decision-computer first) (Z = -2.586; p = 0.01) (median HFC = 93.75%, median HDC = 100%). No other effects were found. See Figure 4.7.

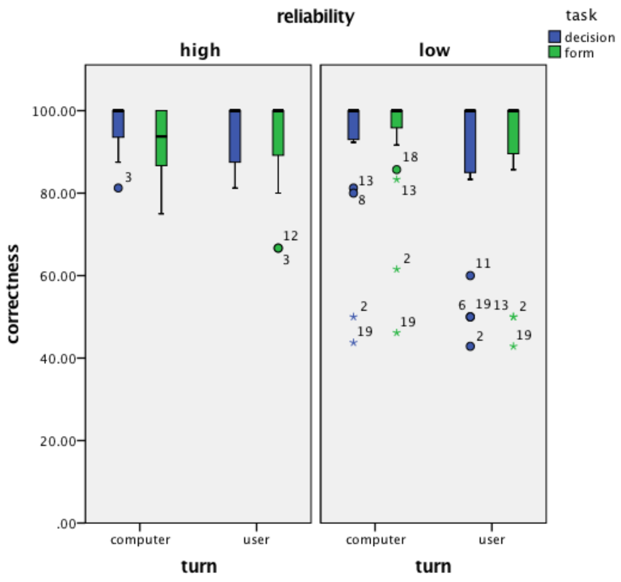


**Figure 4.7 - Percentage Correct Responses**

### *4.1.3.3 Solution Source*

The proportion of trials in which the user selected his response over the computer's as the final answer is named Solution Source. Participants made this decision by either clicking 'Your answer' or 'Computer Suggestion' in the Ramp Metering Control window.

There was a main effect of reliability (Z = -3.15; p = 0.002) (median low = 100%, median high = 100%), illustrated by the signed Wilcoxon test. Participants were more likely to select 'Self' (than Computer) as the source of the solution on the Low reliability condition. The results relating to the solution source present a measure of the users' perceived reliability of the automation. There was neither an effect of turn (user first (U) or computer first (C)), nor of task (decision (D) or form (F)) on solution source. See Figure 4.8.

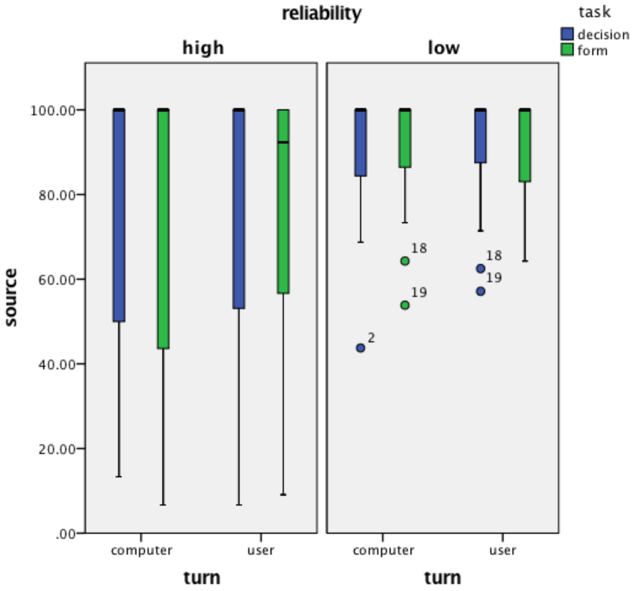


**Figure 4.8 - Solution Source**

### *4.1.3.4 Match*

Match is defined as the proportion of trials in which the answer which the user has given matches (is the same as) the computer's suggestion.

Data for Match were not normally distributed, therefore, a Friedman test was ran, showing the effects on match were found ($\chi^2(7) = 99.3$, $p < 0.001$). A subsequent Wilcoxon was run to check for an effect of reliability. Reliability was found to have an effect on match ($Z = -7.55$; $p < 0.001$) (median low (L) = 25%, median high (H) = 81.25%; 25th percentile low = 25%, 25th percentile high = 76.92%; 75th percentile low (L) = 30.93%, 75th percentile high (H) = 85.11%). No other effects were found. See Figure 4.9.

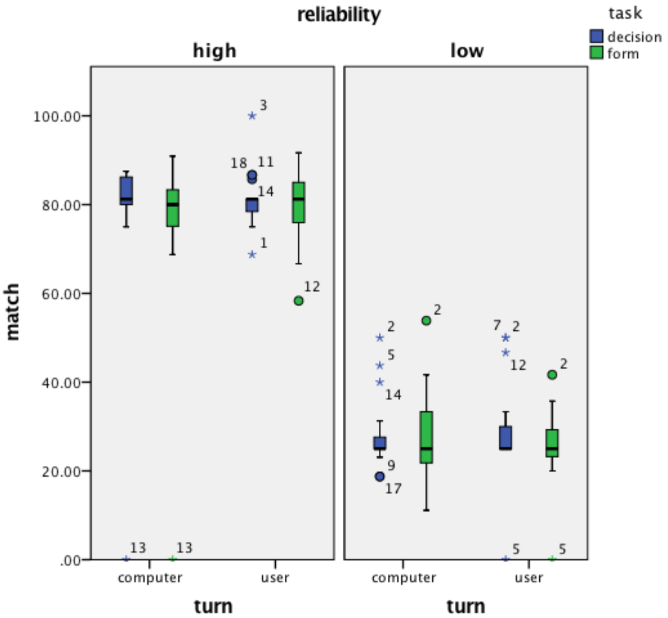


**Figure 4.9 - Decision Match**

Because neither the data for match, nor for solution source were normally distributed, in order to determine the relationship between solution source and match, a Spearman’s rho test was performed. Both variables are measured on an interval scale from 0-100% and a monotonic relationship was found between them (Figure 4.10), therefore, the assumptions for the Spearman’s test were met. There was a weak, negative correlation between solution source and match, which was statistically significant ($r_s = -0.204$, $p = 0.012$).

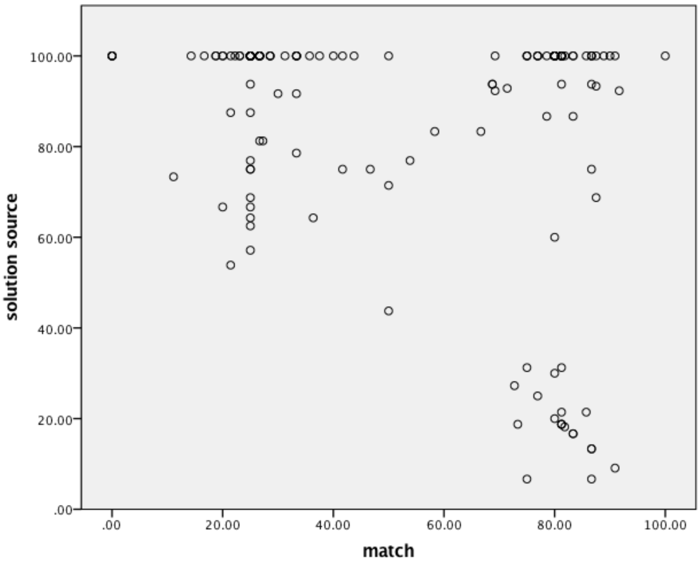


**Figure 4.10 - Monotonic relationship between solution source and match**

### 4.1.4 Discussion

While the experiment shows that completing a form in addition to making a decision incurs a time cost, there are some less obvious findings here. First, when the system has low (25%) reliability, then users are likely to rely on their own interpretation of the system state (and so, regard themselves as the solution source). When the system has higher (81%) reliability, then users will accept advice from the computer (and so, see the computer viable solution source on some as a of the trials). This finding is quite interesting for it indicates that humans are sensitive to automation reliability in spite of the absence of feedback, while previous studies tended to employ response feedback (Dzindolet et al., 2003; Madhavan et al., 2006).

When looking at percentage correct responses, the order in which responses are given has no impact on the users' performance. However, the task (decision, or decision plus form-filling) has an effect on correctness in the high reliability condition, but not in the low reliability condition. In situations where the only the decision is required, the percentage of correct user responses is higher than when he is also required to give his reasoning (i.e. fill in the form). This was a surprising result as one would expect that filling in the form would have the users think twice and re-check whether their decision is correct or not. It seems that in the high reliability condition, form-filling is a source of confusion for the user. It may be that the task of form-filling may have taken a higher priority than that of deciding the course of action, which was the main task. Furthermore, requiring the user to explain his answer in a form using radio buttons, may be a successful means of imposing a particular approach to solving a problem, or it may be a means of externalising procedures. However, doing this may not be the best way to encourage the behaviour of checking given answers. In other words, it may not be the best approach to make the user think twice. This is supported by the results, as match was not affected by task. This result may also mean that requiring the user to give a reason for his answer does not influence his reliance on the computer's answer. From subsequent discussions with some of the users, there has been some indication that form-filling (i.e. giving a reason for the answer) was perceived as a separate task which was attended to separately.

User performance in the low reliability level was sometimes higher than 95% [mean high = 94.04% (st.dev. = 8.36) and mean low = 89.97 (st.dev. = 17.63)] (Figure 4.11).

This could be explained by the fact that users were able to accurately judge the reliability of automation (see match and solution source). However, there seem to be a large number of outliers (10 out of 23) in the low reliability condition. There may have been a subgroup of participants who were not able to accurately judge reliability of the computer and to whom the findings of the Bahrami et al. (2010) study apply. Nevertheless, if humans are able (given the opportunity) to work out how reliable automation is, two heads are always better than one, a finding which is supported by past research (Koriat, 2012).

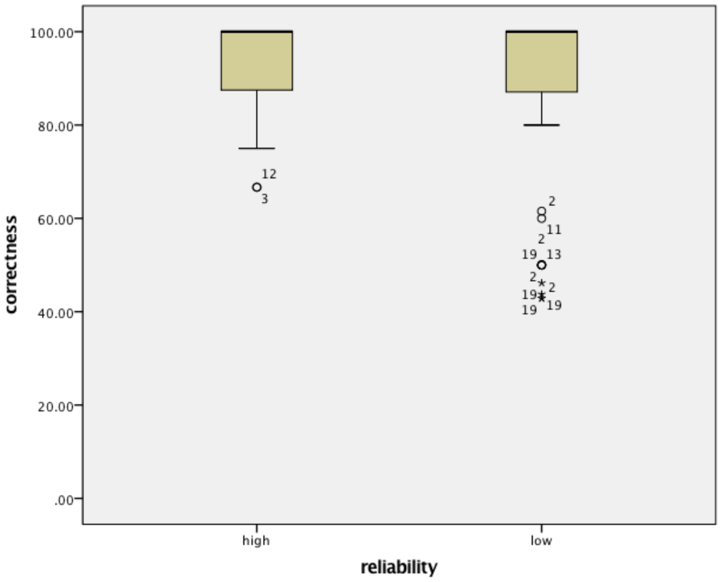


**Figure 4.11 - Percentage Correct Responses in terms of Computer Reliability**

The results which come from the analysis of solution source, suggest that users are able to determine whether the most reliable information source is themselves or the computer. In the low reliability condition, users tend to select themselves as the solution source more often, while in the high reliability condition users prefer to choose computer’s suggestion. For this type of decision task, system reliability has little impact on decision time but does impact on the likelihood that users will accept computer advice. However, this can increase the likelihood of errors persisting within the system. In other words, if users regard the system as having High reliability, they are less likely to intervene when the system has made an error. To illustrate this effect, the 75$^{th}$ percentile of match in the high reliability condition was 85.11%. This shows that some users were likely to exhibit conformance (give the same answer as the computer), provided that the computer is considered to be highly reliable.

There was a significant effect of reliability on match. Moreover, median levels of match for the low and high reliability conditions were 25% and 81.25%, respectively, which are exactly the reliability levels that were set as experimental conditions. If all user were 100% correct, then the match plot would have been a straight line at 81.25% in the high reliability condition and a straight line at 25% in the low reliability condition. The fact that this did not occur can also be seen from the analysis of the percentage correct responses. However, this metric allows us to examine other aspects of the user's behaviour.

Match was not affected by turn. This, perhaps, means that the order in which responses are given does not influence human's reliance on computers. In the context of this experiment, just because the user can see the computer's answer before he gives his own, does not mean that he will copy it. The Form of the control window was meant to stay constant, in terms of layout, regardless of whether the computer or the user went first (Figure 4.4). This was regarded as preferable in order to be able to quantify changes in user behaviour determined by the change in AoF and interaction between the user and the computer. It would be interesting, however, for further research to investigate whether changes in Form in terms of layout (i.e. the position of the response dialog) of the Ramp Metering Control window would produce different effects in human behaviour.

A negative correlation between solution source and match was found: as match is higher, solution source is lower. This means that, when the user's answer is the same as the computer's, the user is more likely to select the computer's suggestion as the final answer. This suggests that participants are more reliant on the computer when they (participant and computer) both arrive at the same answer. Alternatively, it may mean that the user would rather pass accountability for the decision to the computer. Inagaki (2003) believes that authority should always sit with the human, while other researchers like Dekker (2002) and Woods and Cook (2002) suggest that authority, as well as responsibility should be shared between the human and the computer. One can imagine a situation where the computer works on the same task as the human (a form of task sharing, not necessarily the most efficient scenario, but one can presume that humans work on different datasets than computers do) and they arrive to the same answer for a given problem. In case of any issues, the computer could be held accountable for error, or they could share accountability, rather than passing the blame on the human. But, perhaps, this scenario would be an indication of a deeper issue related to the understanding of the subject-matter, requiring resources to be spent on further research rather than on taking disciplinary actions.

# CHAPTER 5    FORM AND FORMAT

In this chapter two User Interfaces are designed to support decision making in a road traffic control task. Both user interfaces are designed to provide the information needed to make critical decisions related to traffic management, in terms of situation awareness and in terms of decision options. Moreover, both user interfaces are also designed to implement principles of ecological interface design. However, the second UI shows a higher degree of integration in the form of task proximity. In addition to comparing the two UI designs, this chapter also considers the impact of the reliability of computer recommendations on decision time and correctness. It is shown that UI2 leads to significantly faster performance on total task time, due to faster performance on the information gathering phase of the task. It is also shown that while performance time with UI1 is affected by computer reliability, this does not affect UI2. On the other hand, decision correctness for UI2 is affected by computer reliability. Impact of UI design on decision making is discussed.

*Part of the section on Proximity Compatibility Principle below has been published in (Morar et al., 2015b). A version of the experiment presented in this chapter has been accepted for publication in IEEE Human-Machine Systems.*

## 5.1 Introduction

This experiment was designed around the SPEEDD Traffic Management use-case. It was set up to investigate differences in user behaviour when using two versions of the final SPEEDD Traffic Management prototype. The design of both user interfaces was informed by CWA and principles of Ecological Interface Design and followed the methodology proposed by Upton and Doherty (2008) (see CHAPTER 2). The two user interfaces are compared in terms of the CFF taxonomy.

### 5.1.1 Proximity Compatibility Principle

The Proximity Compatibility Principle (PCP) is based on the assumption that associated information should be positioned together. This might seem obvious, but it raises two difficult challenges for Human Factors. The first is what one means by ‘associated’ and the second is how this translates into a design recommendation. To elaborate on the first challenge, Wickens and Carswell (1995) suggest that there are two forms of ‘proximity’ to be considered in the design of UIs. The first, ‘display proximity’, suggests that people will see UI components as being associated not simply because they are adjacent, but also because they share common features, such as colour, scale, shape, code. The second form, ‘task proximity’, is defined by the attentional demand involved in obtaining information about a particular system state. There are two main forms of task proximity. Non-integrative task proximity relies on similarity of cues, while integrative task proximity relies on the active combination of information through computation and decision making.

Bennett and Flach (2011) argue that ‘task proximity’ is, essentially, a form of 'match mental model' test. Consequently, there is little to be gained from introducing the concept of 'task proximity' as the suggestion is that users match UI contents with their mental model. However, it is possible that this critique misreads the concept, as ‘task proximity’. Rather than solely being a matter of matching UI content to mental model, task proximity is more closely aligned to the concept of Distributed Cognition than this critique allows. In particular the representation of a task can be considered as the problem space in which the operator’s decisions are framed (Zhang and Norman, 1994). For example, consider the idea of a polygon display (Figure 5.1) in which a collection of parameters which the operator needs to monitor and manage are presented to define the 'envelope' in which system state is performing. Rather than seeking to maintain control of each parameter separately, the operator will (more likely) be trimming the process in order to keep the envelope within limits and, as this envelope becomes distorted, the operator will focus attention on specific parameters. From this, one could suggest that this integrative UI (in which all parameters are available at glance) provides good support of task proximity for normal operations, but that, as the system tips into an unstable mode, it might become less appropriate for managing specific parameters.

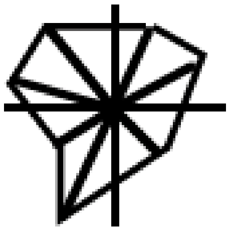
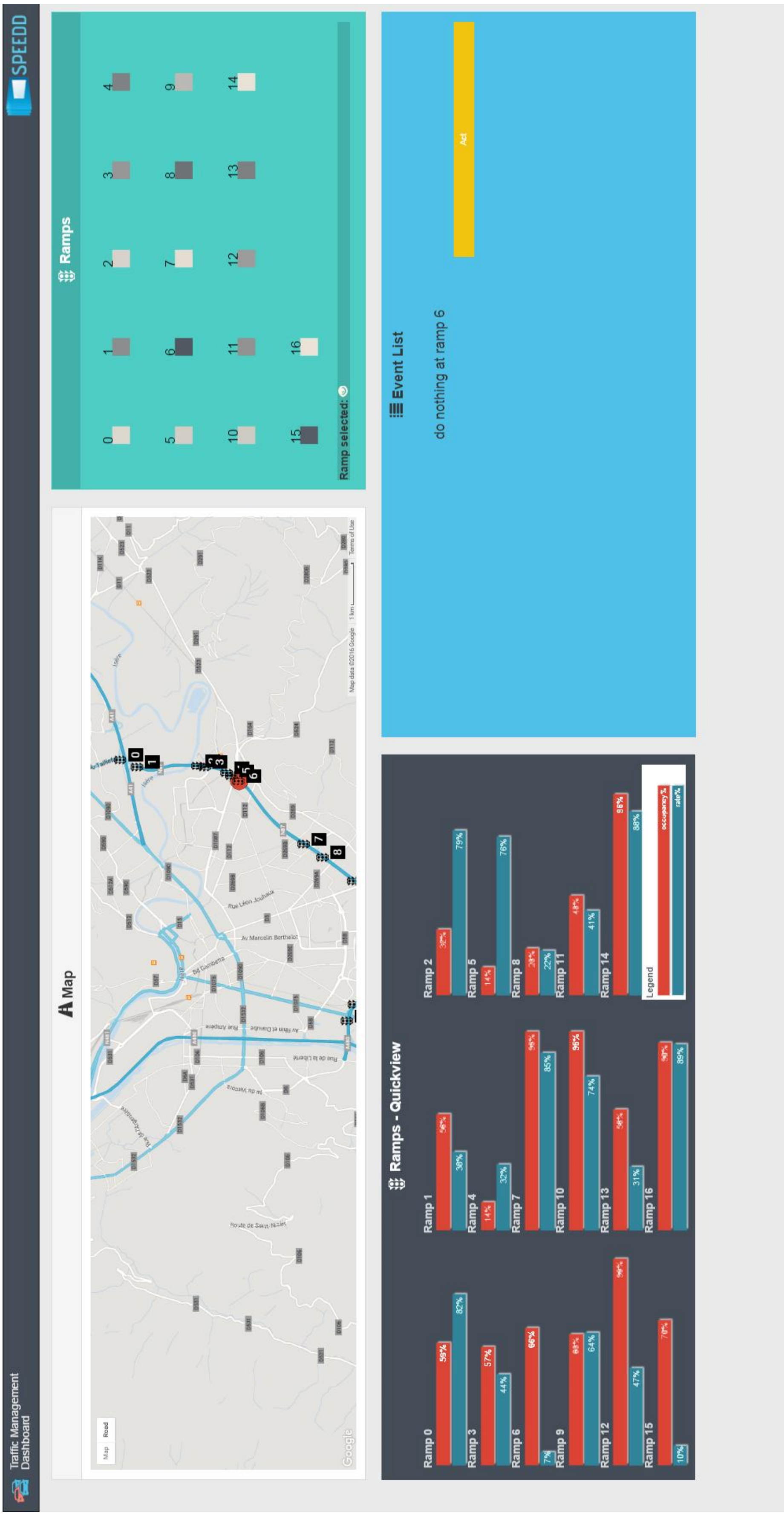


**Figure 5.1 - Polygon display (adapted from Figure 6 (Zhang, 1996))**

The basic conclusion of PCP is that when a task demands attention to be divided between several sources of information, then an integrative UI produces superior performance, but when the task demands attention be focused on single sources of information then non-integrative UI produces superior performance (Carswell and Wickens, 1987; Carswell, 1992; Wickens and Carswell, 1995). For each UI component, the reliability of the displayed information coupled with the relevance of this information to decision making (i.e., its diagnosticity) would define the 'worth' of the component. This supports Woods (1988) proposal that designs should aim to support information extraction by the operator (in terms of allowing the operator to respond to emergent properties which they can interpret on the basis of their experience and knowledge) rather than simply for information availability which requires the operator to search and combine specific pieces of information. This observation has two implications. The first is to consider when UIs should morph from integrative to non-integrative UIs. The second is how operators might perceive 'integration' in displayed information.

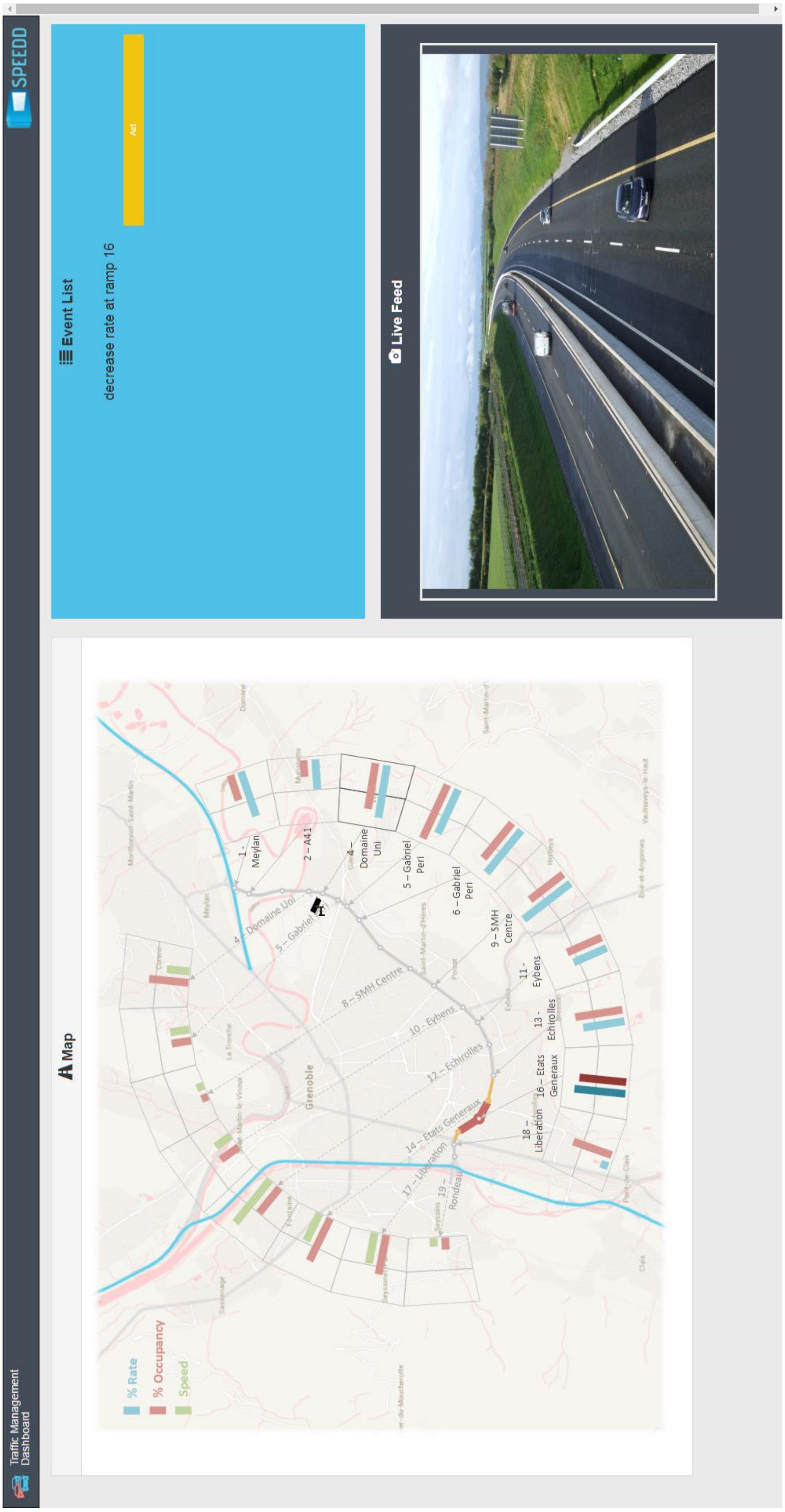
As UI2 (Figure 5.3) is designed to support integrative processing, it should lead to faster information gathering and decision times. However, as both user interfaces are designed to support the means-ends analysis for EID, there should be no differences in accuracy. In terms of the effect of automation reliability, we might expect low reliability to lead to increase in decision time (because of the increased uncertainty that this induces). We would also expect decision accuracy and decision match (i.e., whether or not the user agrees with the automation's recommendation) to vary with automation reliability. EID relates to both Content and Form. In terms of the information requirements, it relates to Content, however the notion of direct perception fits in more with the dimension of Form, along with PCP. See Figure 1.18 for more information (CHAPTER 1).

The hypothesis is that user decision time will be higher when using UI1 (Figure 5.2) (lower degree of integration) and lower when using UI2 (Figure 5.3) (higher degree of integration). Decision correctness is hypothesised to remain constant, based on the fact that the information available for decision making is constant for both UIs and that participants were faced with all three reliability levels when using each UI.



do nothing at ramp 6

Figure 5.2 - UI 1 modified for the experimental task



decrease rate at ramp 16

**Figure 5.3 - UI 2 modified for the experimental task**

## 5.2 Method

An experiment was devised in order to test how performance and overall user behaviour differs while using the two different user interfaces and also in response to varying degrees of computer reliability (Table 5.1).

**Table 5.1 - Independent Variables**

| Factors | Levels |
|---|---|
| **User Interface** | 1 – EID<br>2 – EID + PCP |
| **Automation reliability** | Low – 20%<br>Medium – 50%<br>High – 80% |

### 5.2.1 Task

The experimental task was developed around a realistic Traffic Management scenario. The participants had to respond to two types of alerts (or events) presented by the automatic system: congestion and overflow. For simplicity, ramp metering rate was equated with the frequency that cars are able to pass at a traffic light so that a high rate means that cars can pass quicker than on a low metering rate. A low metering rate is defined as a value below 50%, while a value above this mark is considered to be a high rate. In UI 1, the current ramp metering rates are shown by the blue bars in the "Quickview" window (Figure 5.6) and in UI 2, by the blue bar in the map window (Figure 5.7).

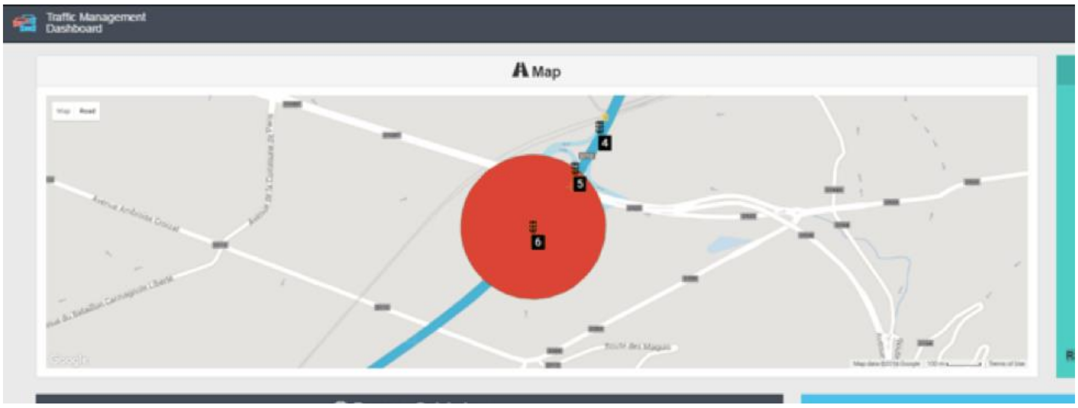


**Figure 5.4 - Congestion view in UI 1**

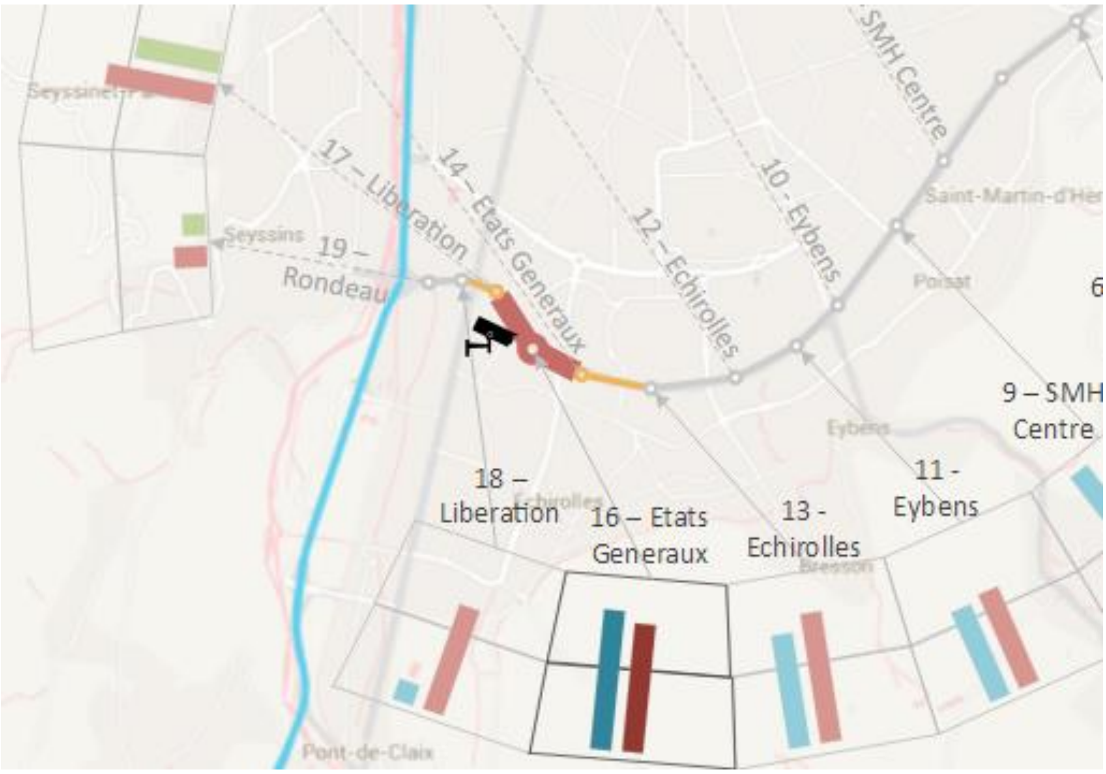


**Figure 5.5 - Congestion view in UI 2**

The congestion event relates to the traffic on main artery. The appearance of this event signals a build-up of traffic in the vicinity of a ramp. In this case, the operator needs to limit the number of cars that can get onto the main road. Therefore, a congestion event requires that the rate of the inbound ramp in closest vicinity to the alert is low. In the first user interface, congestion is shown by an red circle on the map (Figure 5.4), while in the second interface it is shown by an increased width and a red colouring of the portion of the road in question (Figure 5.5).

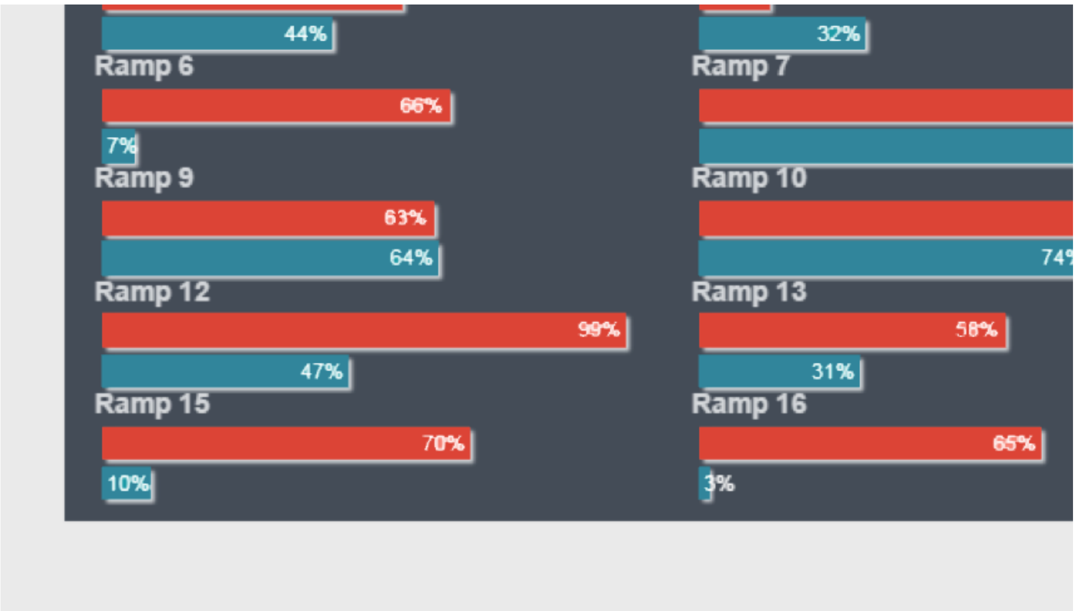


**Figure 5.6 - Overflow view in UI 1**

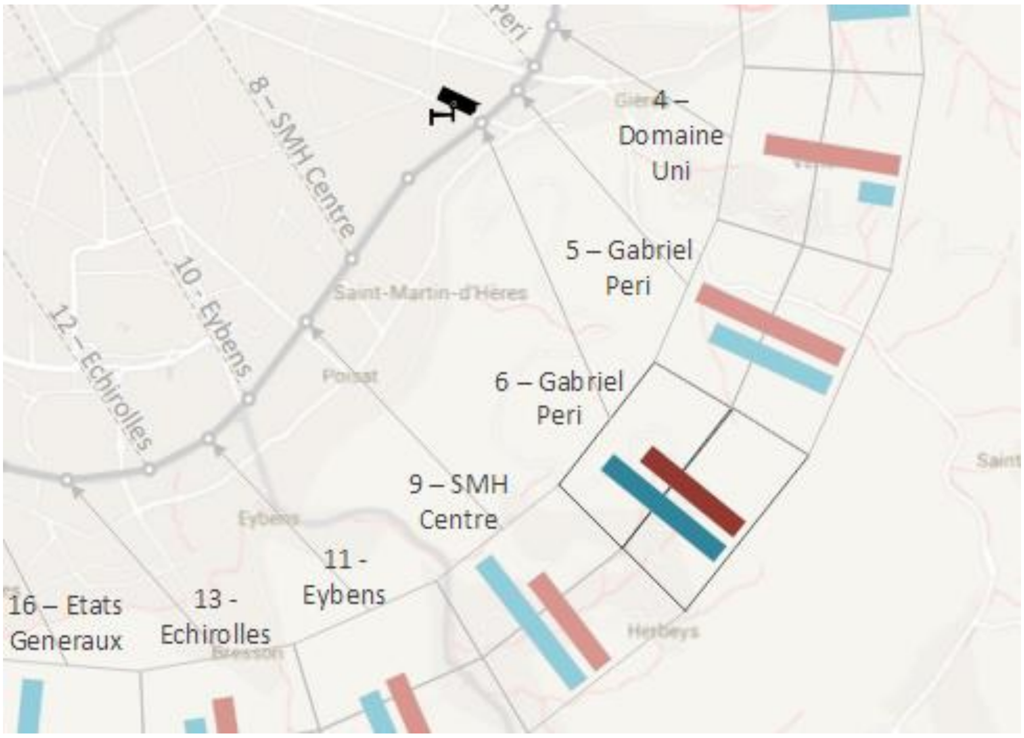


**Figure 5.7 - Overflow view in UI 2**

Overflow is defined as the build-up of traffic on one of the inbound ramps leading to the main road. In the case of an overflow alert, the operator is required to increase the amount of cars that are able to join the main artery, provided that there is no congestion at that location. Therefore, the overflow event requires that the metering rate at the ramp in question is high. Overflow is signalled by a value of over 50% of the density bar. In the first UI the density bar is in the “Ramps – Quickview” window (red bar, Figure 5.6), while in the second UI, density is represented by a red bar on the map (Figure 5.7).

Each new event was triggered by a computer generated message appearing in the event list window. This message was a recommendation of whether to increase, decrease or leave the metering rate at a particular ramp unchanged. It simulated the output of an automated system which gives operators suggestions on the best course of action given a detected event.

### 5.2.2 Procedure

Participants were given a briefing on the experimental task followed by instructions on how to use the interfaces. Participants then began a practice session in order to familiarise themselves with the user interfaces. The practice session consisted of 10 trials, 5 with each user interface. The practice trials were in the same format of those presented in the main experiment but generated randomly for each participant. The computer reliability level was set to 50%. During this session, participants were encouraged to ask any clarifying questions regarding both the user interfaces and the experimental task.

Two independent variables were defined: 1) the user interface that was used to complete the task, and 2) the reliability of the automated system that presented the participants with the suggestion of what action to be performed. The two user interfaces are shown in Figure 5.2 and Figure 5.3. To better control the experiment, users were required to respond to one event at a time. This leads to only one computer suggestion being shown in the event list, for both UIs, and one CCTV view in the case of UI 2, compared to multiple views in the initial interface.

Automation reliability was set at three levels: low (20%), medium (50%) and high (80%). The reliability levels related to the proportion of computer suggestions that were correct in a given block (condition). Therefore, in the high reliability condition, 80% of the suggested actions were correct solutions to the events that were presented in that condition.

The main experiment consisted of 60 trials and was split into six blocks, 10 trials per block. Each block represented one of the possible combinations of the two user interfaces and the three computer reliability levels. Participants were given a 10-second break between each block.

The start of each new trial was signalled by a computer suggestion appearing in the Event List window. The message consisted of a recommended action and the number of the ramp controller in question. In order to validate the computer suggestion, participants had to identify whether the event was of a congestion or an overflow type. The users then had to decide whether to increase, decrease or leave the ramp metering rates unchanged, depending on the current rate levels as seen in Table 5.2. Participants were allowed to use this table for reference throughout the experiment. This bypassed the need of memorising the rules and was also in accord with the information that Traffic Operators gave us, more specifically that the procedures for taking action are fixed and there is very little, if any, variability when making a control action (CHAPTER 2). When the user was ready to give a response, he would click on the “Act” button present in the event list window. This revealed a list of the possible actions (in the form of a radio buttons list) to take in regards to the metering rate (i.e. increase, decrease, nothing). The user would then select their answer and press the “Submit” button below the list. This signified the end of the trial and the beginning of a new one. Participants were instructed to complete the trials as quickly and as correctly as possible.

**Table 5.2 - Correct Responses in Terms of Event and Rate Level**

| | **Congestion** | **Overflow** |
|---|---|---|
| **Low Rate** | do nothing | increase rate |
| **High Rate** | decrease rate | do nothing |

### 5.2.3 Mapping Visual Variables to the Content Format Form Taxonomy

**Table 5.3 - Mapping Visual Variables to Content Format Form**

| Visual Variable | Content | | Format | | From | |
|---|---|---|---|---|---|---|
| | UI1 | UI2 | UI1 | UI2 | UI1 | UI2 |
| **Congestion Display** | present | present | | | graphical as a circle on the map at the location in question;<br>colour red | graphical as a highlighted node and/or segment on the road (map);<br>colour red |
| **Ramp Rates** | present | present | | | textual + graphical;<br>blue bars in the ramps window below the map;<br>separate window from map | textual + graphical;<br>blue bars integrated in the map;<br>close to the ramp in question;<br>integrated in main map;<br>higher task proximity |
| **Ramp Occupancies** | present | present | | | textual + graphical; | textual + graphical; |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | red bars in the ramps window below the map;<br>separate window from map; | red bars integrated in the map;<br>close to the ramp in question;<br>integrated in main map;<br>higher task proximity; |
| **CCTV** | absent from main display/ available on request | present | needs to be explicitly invoked | camera icon moves to attended location; | image | image;<br>higher task proximity |
| **Computer Suggestion** | present | present | | | textual | textual |

The main differences between the two User Interfaces relate to the application of PCP through the increase of task proximity and, thus to the dimension of Form (see Table 5.3). An aspect which relates to Format is the presence of the CCTV on the main display in the second UI. The reason why this change is not regarded as pertaining to Content is that, CCTV feed is also available in the first UI, however the user has to explicitly bring it up (i.e. the UI requires different interaction).

Using the same colour to indicate congestion and Ramp occupancies (red) for both UIs, a high display proximity is achieved (Form). UI2 shows a higher task proximity because, when the user attends to an event, the camera icon moves to the location (node and ramp) in question and Ramp rates and occupancies are highlighted by increasing the opacity of the border around them (Form). Moreover, the CCTV feed also changes when the user is attending to an event (Format), UI2 requiring user less interaction for accessing this information as compared to UI1, where the user has to bring up the CCTV by clicking on a node on the road. Task Proximity is further increased in the second UI by integrating the information presented in the Ramps-Quickview window (UI1) into the main map (UI2) (Form).

The dimension of Content is kept constant across the two UIs by making the same information available to the users. Furthermore, the Event List window did not incur any modifications.

The application of the Proximity Compatibility Principle and the reduction of information accessing cost for the CCTV feed in the second UI should lead to faster decision times and, potentially, higher decision accuracies than when using the first UI.

### 5.2.4 Participants

24 people took part in the experiment [13: male; 11: female; age range: 22-29]. None of the participants had any prior experience of working in Traffic Management.

It was considered that there was no need to include domain experts as participants in the experiment because the task of ramp metering control had not yet been adopted in the control centre the project partnered with at the time of writing. Therefore, the DIR-CE traffic managers did not have any expertise in the task of controlling metering rates and they would have had to undergo training. Considering that the availability of traffic experts is extremely low, training non-experts was deemed a good alternative.

The experiment met University of Birmingham ethics approval (Reference Number ERN_13-0997). All data were anonymised and participants provided informed consent.

### 5.2.5 Data Collection and Analysis

#### *5.2.5.1 Dependent Variables*

Five dependent variables were defined in terms of the two user interface versions and the three reliability levels.

The dependent variables were:

- information gathering time: the interval between the start of a trial and the time the user pressed the "Act" button;
- time to submit decision: the interval between the time that the answer options were revealed to final decision;
- total task time: the median time to make a decision, in other words, the average time to complete a trial.
- decision accuracy: the percentage of correct decisions out of the number of trials;
- decision match: when a user's decision was the same as the computer;
- subjective workload: measured using the NASA TLX.
- subjective usability: System Usability Scale

For each trial the following data were gathered: trial start time, trial end time, act button press time, submit button press time, trial number, user decision, computer suggestion, block number, UI version and the following derived metrics: act interval (time elapsed from trial start until the user presses the act button), submit interval (time elapsed from when the user presses the act button until he presses the submit button), total trial duration (i.e. act interval + submit interval), user-computer decision match and user decision correctness.

Data for each participant was stored in a separate Comma Separated Variables (csv) file on a secure University server. Pre-processing was carried out on each participant data in order to remove outlier trials (trials which took very long to respond to) where participants may have been engaged in other tasks or where clarifying questions have been asked. A thresholding of mean + 1 s.d. was used on the total trial duration. This resulted in the exclusion of 13.68% of the total number of trials. A Shapiro-Wilk test was then run on the remaining data in order to check for normality. Where conditions for normality were met (i.e., $p>0.05$), an ANOVA test was performed, followed by a pairwise analysis (Bonferroni adjustment), where appropriate. For data which were not normally distributed (ie. $p<0.05$), a Friedman test was ran, followed by a Signed Wilcoxon test for pairwise comparisons (Bonferroni adjustment), where appropriate. All statistical tests were performed using IBM SPSS v24.

The study was split into 6 experimental blocks, each consisting of 10 trials, amounting to a total of 60 trials. In each experimental block one of the possible combinations of the two user interfaces (no PCP, PCP) and the three computer reliability levels (low, medium, high) were employed. The order in which blocks were presented to participants was as follows: first half of the participants (in the order of arrival) were presented with user interface 1 followed by user interface 2, while the second half of participants first completed trials using user interface 2 followed by user interface 1. Within each user interface, the order in which computer reliability changed was random (using a pseudo-random number generator) for each participant. Moreover, the order in which participants were shown the 10 trials within each block was also random (using a pseudo-random number generator). In order to control for learning effects, no performance feedback was given to participants.

The user interfaces for both the practice session and the main experiment, were displayed on a 22” monitor (1920x1080 resolution) and the interaction was achieved using a standard mouse.

## 5.3 Results

### 5.3.1 Total Task Time

Total task time was defined as the mean time needed to make a decision, in other words, the average time to complete a trial. Data were not normally distributed for the two UIs. Therefore, a Signed Wilcoxon test was performed. This showed that users performed faster when using UI2 (median UI2 = 10.05s; median UI1 = 14.30s) (Z = -7.076, p < 0.001) (Figure 5.8).

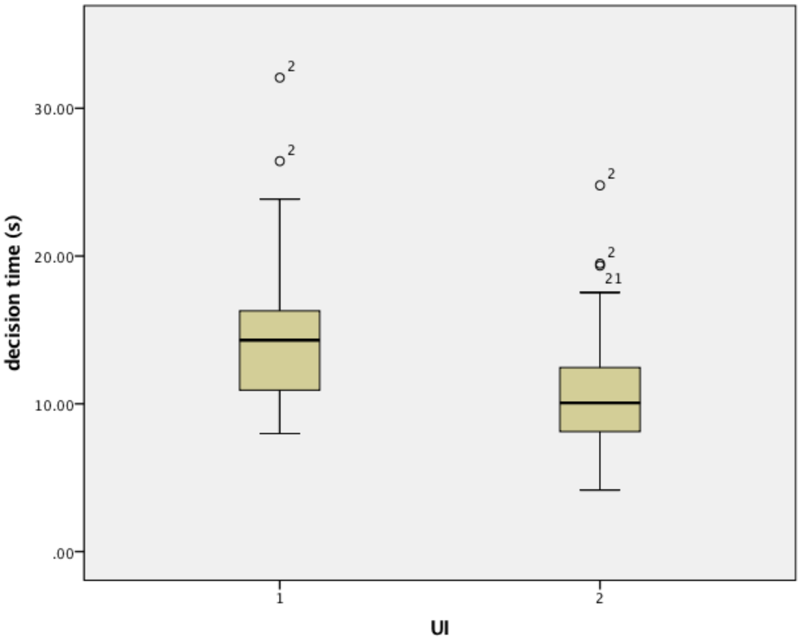


**Figure 5.8 - Decision time for the two UIs**

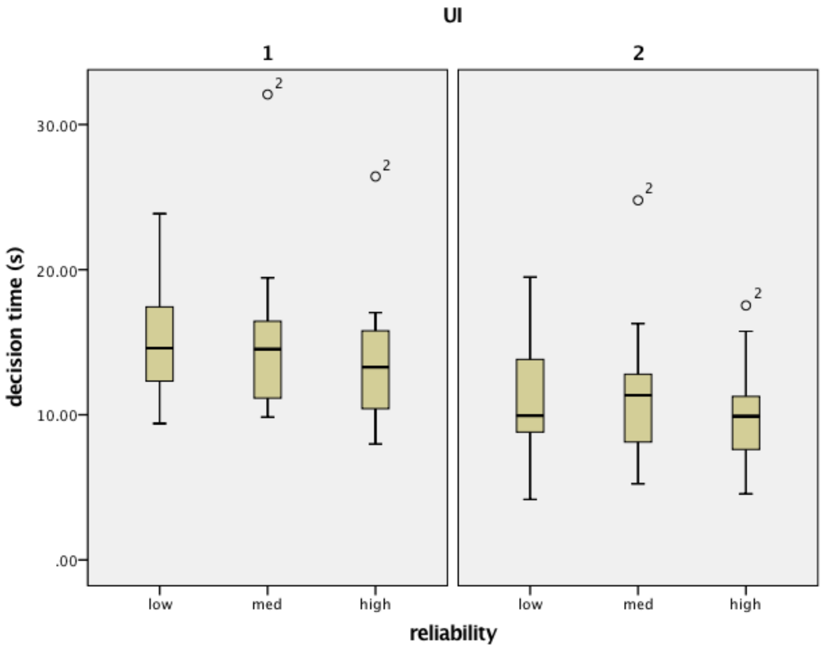


**Figure 5.9 - Decision time for the different UIs and reliability levels**

Within UI1, Shapiro-Wilk showed data to be non-normal. A Friedman test revealed differences between the different reliability levels for UI1 ($\chi^2(2)$ = 10.583, p = 0.005). Post hoc analysis with Wilcoxon signed-rank tests showed that users were faster to make a decision in the high reliability condition when compared to the medium (Z = -2.51, p=0.036) condition (median low = 14.58s, median high = 13.27s).

For UI2, data were not normally distributed. No significant effects were found by running the Friedman test. However, the signed Wilcoxon test revealed a significant difference between the high and the medium reliability levels when using UI2 (Z = -2.51, p=0.036) (median med = 11.34s, median high = 9.89s). These results can be seen in Figure 5.9.

### 5.3.2 Information Gathering Time

Information gathering time was defined as the interval between the start of a trial and the time the user pressed the “Act” button which revealed the answer options. Data for the two user interfaces were not normally distributed. A Wilcoxon signed-rank test revealed that users are quicker to click the act button when using the second user interface (median UI2 = 2.63s; median UI1 = 7.12s) (Z = -6.813, p < 0.0001). This is shown in Figure 5.10.

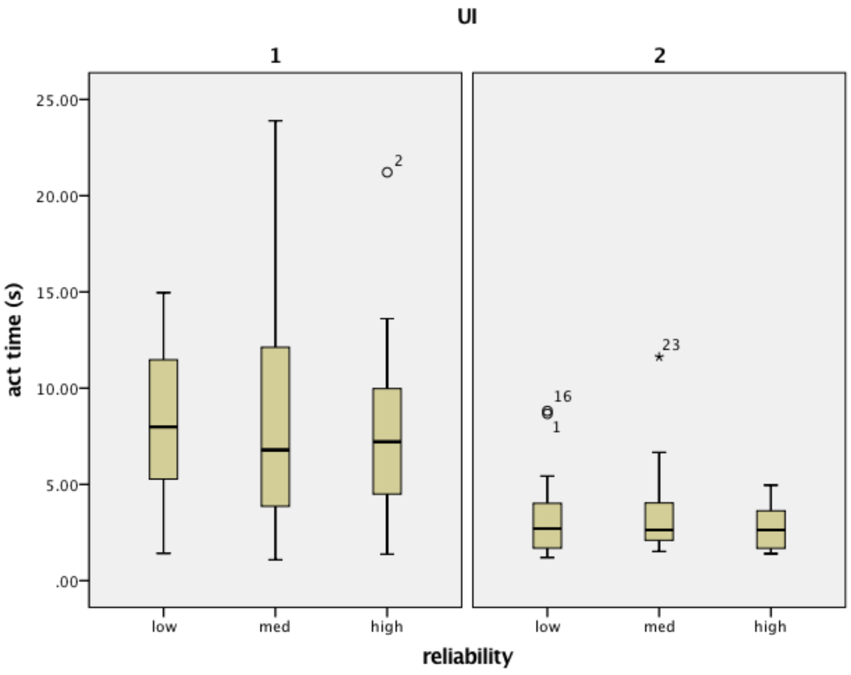


**Figure 5.10 - Act time for the different UIs and reliability levels**

Taking each user interface individually, act times in terms of reliability for UI1 were normally distributed. ANOVA did not show any significant effects of reliability and the pairwise comparisons did not reveal any differences either. Data were not normally distributed for UI2. A Friedman test was run showing no significant differences.

### 5.3.3 Time to Submit Decision

Time to submit decision was defined as the interval between the time that the answer options were revealed to the user up to the time the final decision was submitted. First, differences between the UIs were investigated. Data were non-normal, therefore a Wilcoxon signed-rank test performed. Even though not immediately apparent from plotting the data (see Figure 5.11), this revealed a slight advantage for using UI1 (Z = -2.424, p=0.015) (median UI2 = 7.04s; median UI1 = 5.89s).

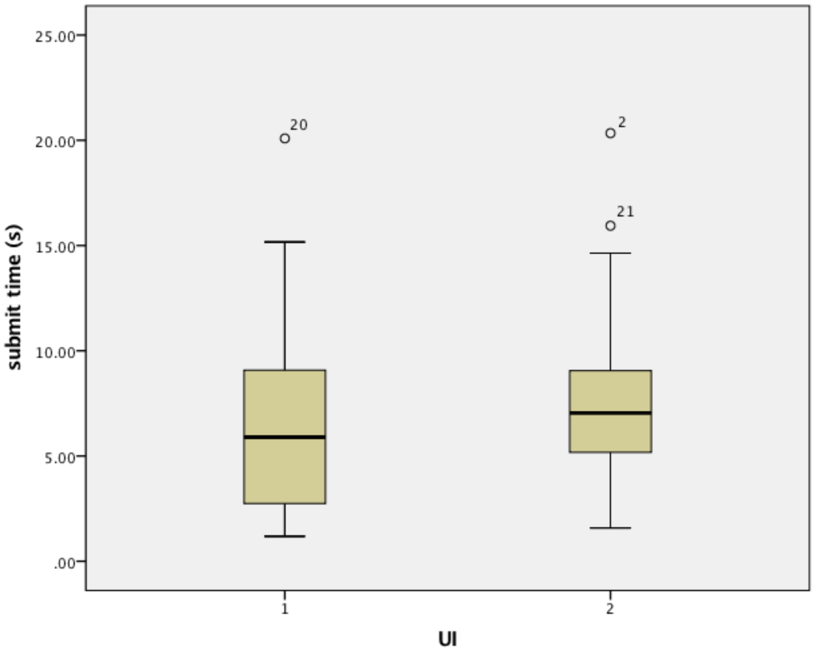


**Figure 5.11 - Submit time for two user interfaces**

To investigate the effects of reliability within each of the user interfaces, we begin by testing for normality using the Shapiro-Wilk test. Submit time data were not normal for UI1 and normal for UI2. A Friedman test for UI1 data showed not significant differences between the three defined reliability levels, hence no further tests were performed. Similarly, no differences were found within UI2 when performing an ANOVA. See Figure 5.12.

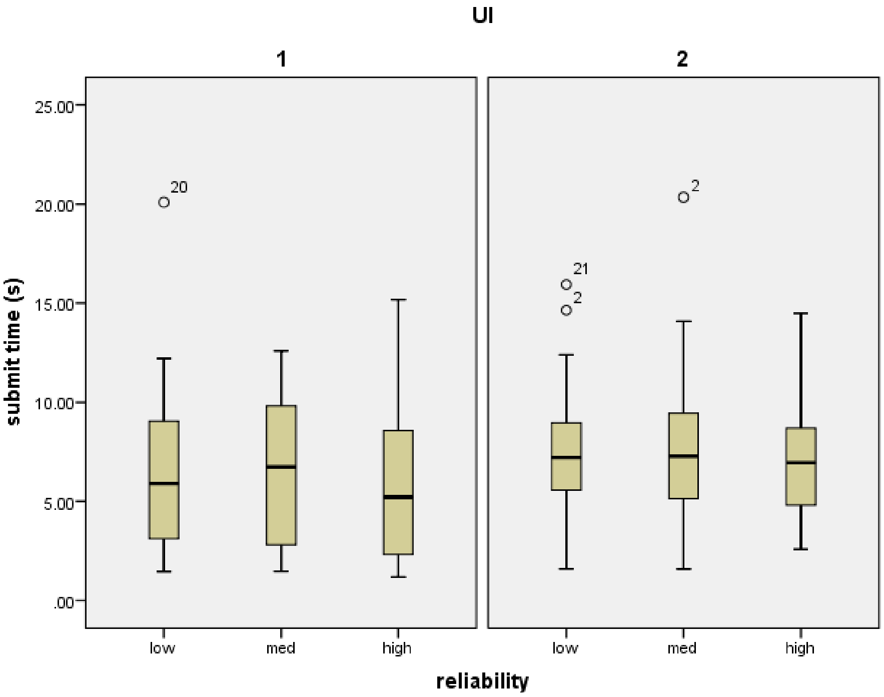


**Figure 5.12 - Submit time for the different UIs and reliability levels**

### 5.3.4 Decision Correctness

Decision correctness refers to the percentage of correct decisions out of the total number of trials engaged in. All decision correctness data were non-normal. No significant differences were found in terms of decision correctness between the two user interfaces when running the Wilcoxon signed-rank test. However, median correctness for UI2 was slightly lower than UI1 (median UI1 = 95%, median UI2 = 90%), although differences were not significant.

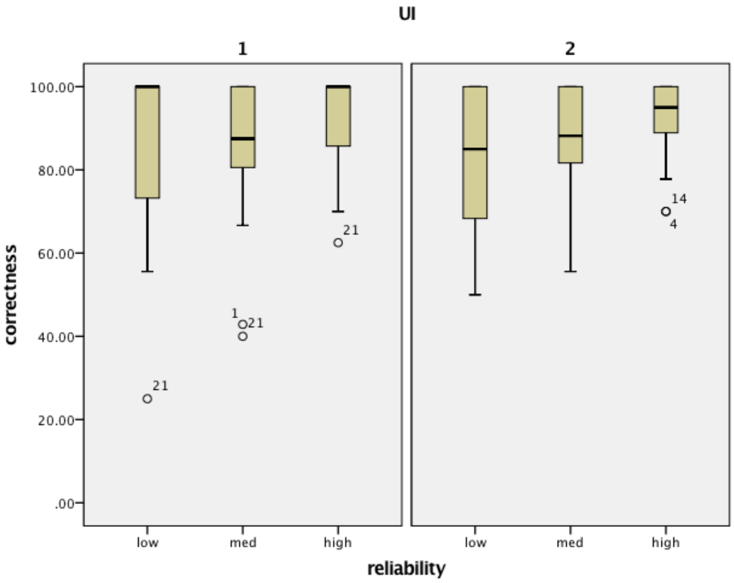


**Figure 5.13 - Decision correctness for each UI in terms of computer reliability**

When looking for an effect of reliability on decision correctness, it was found that there were no significant differences between the three reliability conditions for UI1. However, Friedman showed an effect for UI2 ($\chi^2(2) = 6.29$, p = 0.043). A subsequent run of the Wilcoxon signed-rank test identified that users were more correct when in the high reliability condition (Z = -2.411, p=0.048) as compared to the low condition (median low = 85.0%, median high = 95.0%). These results are shown in Figure 5.13.

A further test was carried out in this situation, looking to determine if any differences could be spotted between the three reliability levels (low, medium and high) when looking at the two user interfaces together (Figure 5.14). Since all data were non-normal, a Friedman test was performed. The result showed that a difference was present ($\chi^2(2) = 8.132$, p = 0.017). The Wilcoxon signed-rank test revealed participants were more correct in the high reliability condition as compared to both the medium (Z = -2.749, p=0.018) and low (Z = -2.924, p=0.009) situations (median low = 100%, median medium = 87.5%, median high = 100%). There was no significant difference between the low and the medium conditions.

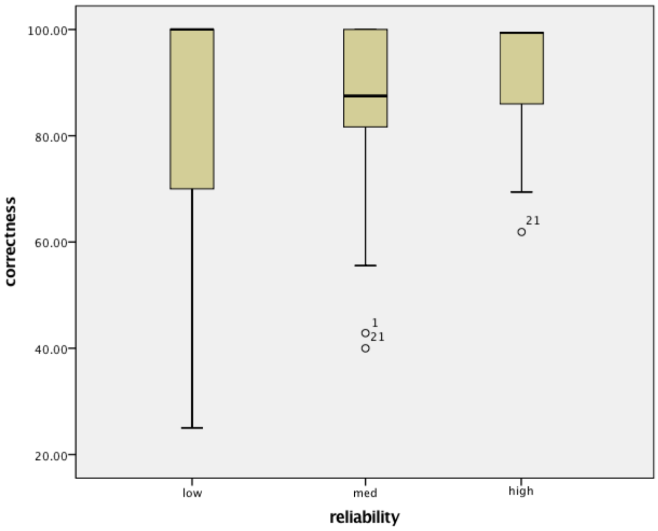


**Figure 5.14 - Decision Correctness for the two UIs together**

### 5.3.5 Decision Match

When a user's decision was the same as the computer suggestion for a particular trial, we say that a decision match occurred. Data for the two UIs were not normally distributed. The performed Wilcoxon signed-rank test did not show any difference between user interfaces in terms of decision match.

When looking for effects of reliability on decision match within each UI, it was revealed that data were non-normal for UI2 and the low reliability condition of UI1, and normal for the medium and high reliability conditions of UI1.

A Friedman test for UI1 revealed some differences in decision match between the reliability levels ($\chi^2(2) = 42.25$, $p < 0.001$). The Wilcoxon signed-rank test showed that there was a lower decision match in the low reliability condition than in the medium ($Z = -4.144$, $p < 0.001$) and the high reliability condition ($Z = -4.258$, $p < 0.001$) (median low = 20%, median medium = 50%, median high = 80%). A paired samples T-test was also performed between the medium and high conditions, since their data were normally distributed. The results showed that there was a higher decision match in the high reliability condition than in the medium reliability condition ($t = -9.410$, $p < 0.001$; mean medium = 49.32%, stdev = 17.37; mean high = 79.28%, stdev = 11.57). Figure 5.15 illustrates this effect.

In terms of UI2, the Friedman test found statistically significant differences between the three reliability levels ($\chi^2(2) = 47.06$, $p < 0.001$). A further run of the Wilcoxon test found differences between all possible pairs of the three reliability levels. Match was higher in the high reliability condition than both the low (Z = -4.289, p < 0.001) and medium (Z = -4.293, p < 0.001) conditions. Furthermore, match levels were higher in the medium than in the low reliability condition (Z = -4.109, p < 0.001). Median levels for match were 21.11%, 50% and 80%, for the low, medium and high reliability condition, respectively (see Figure 5.15).

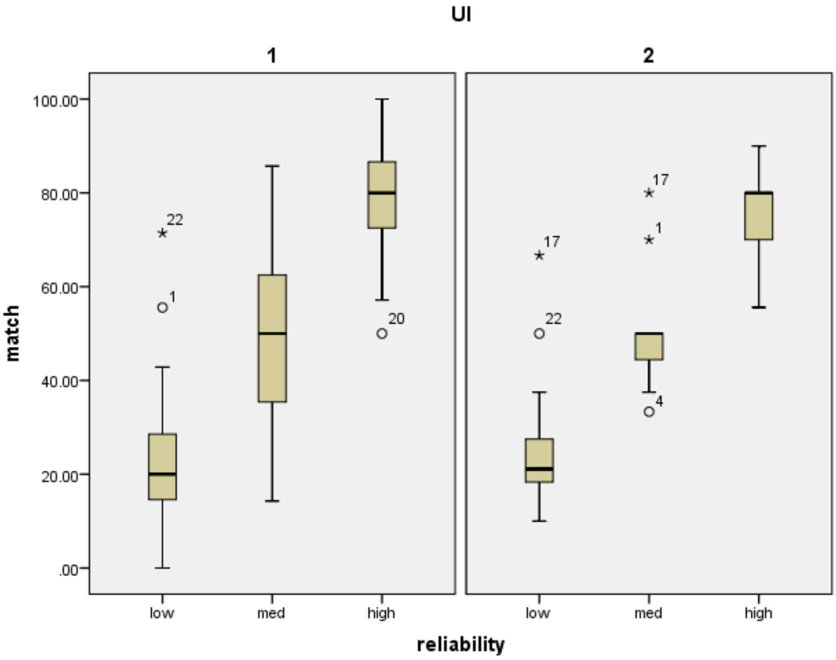


**Figure 5.15 - Decision match for the different UIs and reliability levels**

### 5.3.6 Workload

The NASA TLX results normally distributed for both UIs. All the assumptions were satisfied for performing a paired-sample t-test. No significant difference was found between the two UIs in terms of subjective workload (t(23) = 2.045, p = 0.053). The mean score for the first interface was 59.08 (stdev = 18.58), while for the second, 52.45 (stdev = 21.77). See Figure 5.16.

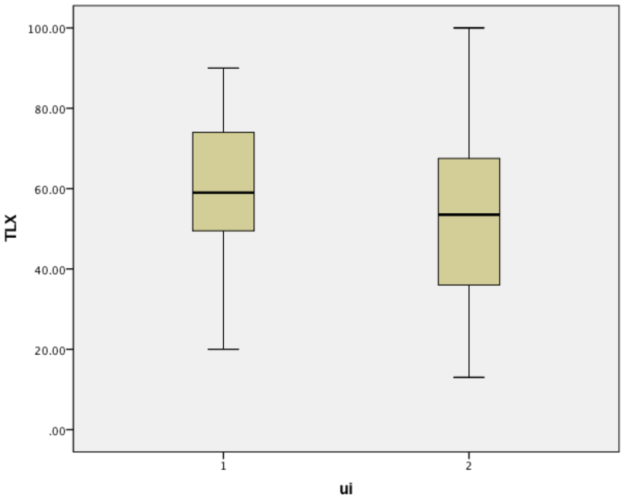


**Figure 5.16 - NASA TLX scores**

### 5.3.7 Usability

The results of the SUS questionnaire were normally distributed for UI1 and non-normal for UI2, therefore a signed Wilcoxon test was performed. Results showed that there was a preference for UI2 that was statistically significant (Z = -2.859, p = 0.004) (median UI1 = 50.0, median UI2 = 77.5). User interface 2 scored above the 68 margin (mean UI1 = 49.37, mean UI2 = 69.37), indicating above average usability (Figure 5.17).

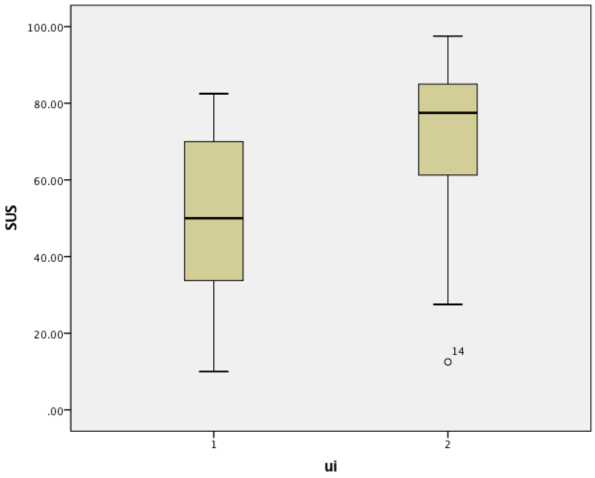


**Figure 5.17 - SUS scores**

Figure 5.18 shows the SUS score of the TM UI versions 1, 2 and 3, as rated by SMEs. An increasing trend can be spotted in the SUS score such that, with every iteration, the UI achieves a higher score, even though not above the 68-point threshold.

UI 1, in experiment, is a slightly simplified version of TM_V3 (where the Driver behaviour window is excluded because it is not used in the experiment). Students who took part in the experiment rated this specific version more harshly than domain experts, giving it an average score of 49. However, they have rated the final version of the TM UI (TM_V5.0 - UI2, in experiment) with an average score of 69, which is considered as above average usability.

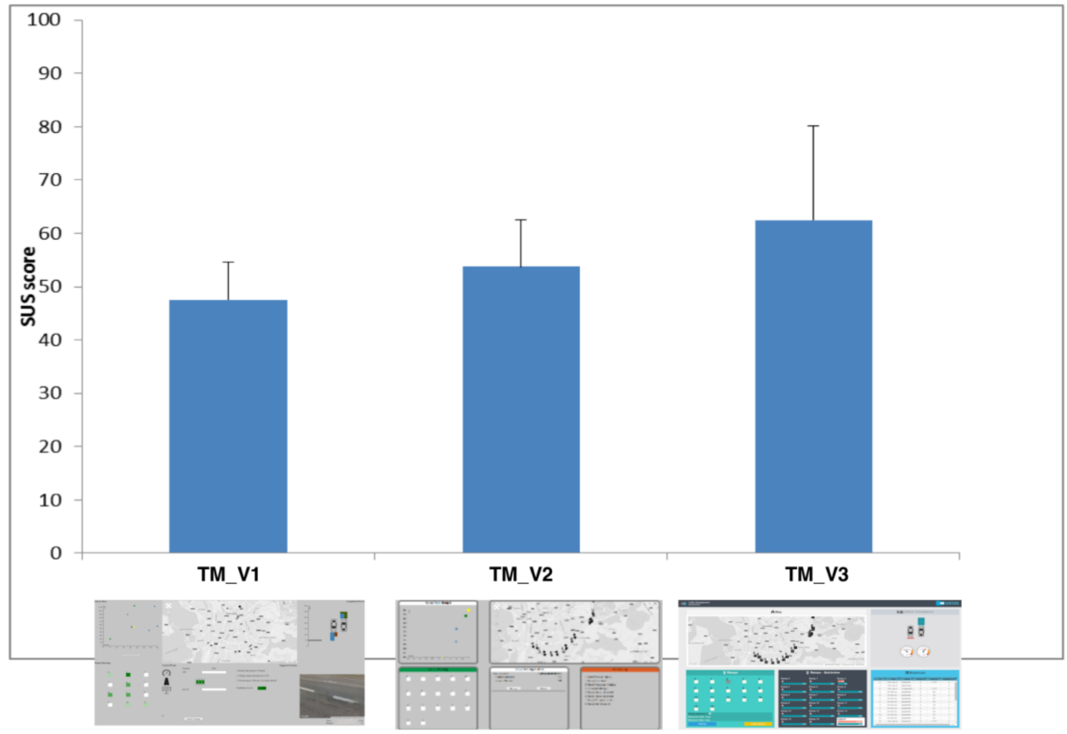


**Figure 5.18 - SME Usability ratings for the TM UIs**

## 5.4 Discussion

The analysis of the mean decision time exposed a large difference between the two user interface versions. However, in order to explore this effect further, we look at the two components of total decision time (i.e. time to act and time to submit). An interesting effect can be spotted: time to submit does not vary by a large amount, as it can be seen in Figure 5.11, whereas a large effect of UI version was identified for time to act (see Figure 5.10), with a difference between means of approximately 5 seconds. This suggests that the two intervals (act and submit) relate to two distinct stages in operator decision-making. The first one, the act interval, being the information gathering stage, while the submit interval, the final checking and response submission stage. Assuming that this is what is actually happening, then UI 2 speeds up the process of gathering information.

This large improvement in decision time that the second interface has brought comes with no reduction in decision performance. Figure 5.19 shows % correct responses for each UI with reference to the average computer reliability. However, despite the large reduction in the total time to complete a trial, subjective workload scores stay relatively constant (see Figure 5.16).

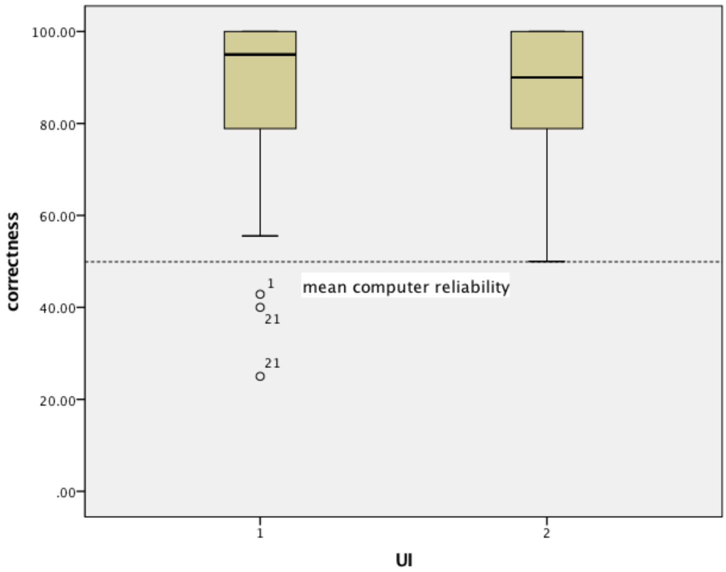


**Figure 5.19 - Decision correctness with reference to average computer reliability across experimental conditions**

Although there was no difference in decision correctness in terms of the user interfaces, an effect of reliability was identified. User performance improves as automation reliability increases, as can be seen Figure 5.13. However, it seems that users could more accurately gauge computer reliability in the low and high conditions when using UI1 than when using UI2. Even though these effects were not statistically significant, the results may suggest that too much integration of information could lead to complacency and conformance and, thus, to a reduced ability to spot automation errors. Alternatively, it may be that the user sees the extra time cost incurred by checking the automation response when using the second display as outweighing the overall benefits of slightly more correct decisions.

The significant effect of reliability on decision time could point to the fact that users are able to distinguish between the different reliability levels, resulting in a more cautious approach to decision-making in the low and medium reliability conditions, as compared to the high reliability condition. In terms of quantifying this sensitivity to the computer reliability, we look at % decision match (Figure 5.20). Participants achieve mean match levels of 24.02% (std. error = 2.47), 48.67% (std. error = 2.32) and 78.25% (std. error = 1.85) for the low, medium and high reliability condition, respectively. This illustrates that participants are able to accurately determine whether they should follow the computer recommendation, considering that the computer's reliability level was set at 20, 50 and 80% for the low, medium and high condition, respectively.

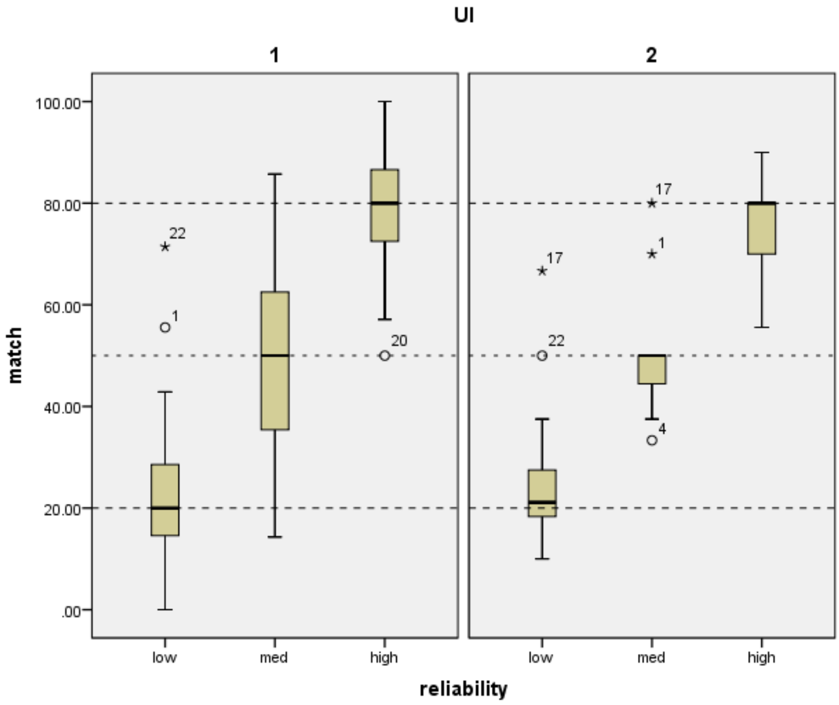


**Figure 5.20 - Decision match for the different UIs and reliability levels with reference to computer reliability in the respective blocks**

A potential criticism to the match metric being an accurate indication of the users' sensitivity to the computer reliability level is that when users are 100% correct, then match levels are 20%, 50% and 80% for each reliability condition respectively. And this is true, provided that users are 100% correct. However, this is not the case, participants achieving mean correctness scores of 84%, 86% and 92%, in the low, medium and high condition, respectively. Therefore, in the low reliability condition, for example, the minimum match score in this case would be around 4%. However, the actual score is very close to the computer's set reliability level, i.e. 24% vs 20%. Another argument is that match and correctness are not linked. More specifically, the user’s response is not given as an acceptance or rejection of the computer’s recommendation, but as a decision of whether to increase, decrease or leave the ramp metering rates unchanged.

# CHAPTER 6 CONCLUSIONS AND FURTHER RESEARCH

## 6.1 Research Questions

1. What effects does automation reliability have on human decision making?
2. How can we design user interfaces to help users cope with these effects?

The experiments presented in the thesis reveal interesting findings in two domains: 1) User Interface Design and 2) Joint Decision Making. These two areas are traditionally studied separately, however this work presents the benefits of bringing the two domains together. The experiments were designed around a simulated Traffic Management task and user interfaces employed were developed as part of the SPEEDD project. All user interfaces were developed according to EID principles and following the design methodology proposed by (Upton and Doherty, 2008). The Content/Format/Form (CFF) taxonomy was further used in order to aid in the discussion of how results could inform future display designs.

## 6.2 Experiment 1 – A Baseline for Joint Human-Automation Decision Making and Implications for UI Design

In 2010, Bahrami et al. presented a study which showed that two heads are better than one, provided that dyad members (in a perceptual decision-making task) have similar sensitivities and had the ability to freely communicate. The work presented in this PhD is based around Human-Automation systems, in which humans are working and cooperating with computers/automation. We investigated whether Bahrami's findings can be extended to dyads in which one of the members is a computer.

The first study, presented in CHAPTER 3, there was no communication between the human and the computer, apart from the computer displaying its suggestion to the human. Moreover, overall automation reliability was very low (25%). This made the experiment more similar to a signal-detection task, rather than a study of human-automation collaboration. Three students exhibited performances similar to those of expert traffic operator, achieving correctness scores greater than 95%. However, 13 out of the 16 participants were unable to spot system errors in the form of incongruence of displays. This effect could be explained by the fact that ‘low-performing’ student were unable to judge the ‘worth’ of the displays for congruence, focusing of validating the automated suggestion. Another explanation for this effect could be given by a phenomenon in radiology research called ‘satisfaction of search’. It was observed that some medical practitioners terminate their visual search at the first sign of pathology (Berbaum et al., 1994, 1990; Samuel et al., 1995). In this case, the most ‘salient’ type of pathology was automation correctness in terms of input data. However, this approach did not take into account the possibility of malfunctioning sensors leading to corrupt data, which the incongruence case simulated. In terms of display Form, this finding could indicate the need for an information source to indicate when incongruence occurs. Alternatively, this finding suggests the need for a change in display Form that would highlight to the user the ‘worth’ of checking for incongruence. Perhaps, this could translate into an increased salience for the ROIs in question (Form), or the ‘fusion’ of the information sources that could disagree in an integrative display (Format).

## 6.3 Experiment 2 – Format

The first study identified that there was a need to include more levels of automation reliability if there would be any discussion to be made in terms of human-automation systems. The second study (CHAPTER 4), investigated the issue varying levels of automation reliability and the issue of communication (i.e. automation transparency). Employing different reliability levels (25% and 81%) proved to be a good way to simulate ‘sensitivity’ for the case of the computer member in the dyad, so as to approach the experiment design of Bahrami et al. (2010). It was assumed that human sensitivity was constant, i.e. ability to perform the task was constant across the experiment. Care was taken in order to counter learning effects and no performance feedback was given. Bahrami et al. (2010) approach this issue in the same manner. Apart from reliability, two other independent variables were introduced: turn (the order in which responses are given, i.e. computer first, user first) and task (whether or not justifications for responses are given). There were no effects of reliability, turn, or task on decision accuracy, however, in the high reliability condition, when the computer answered first, there was an effect of task.

Because in experiment 1 users’ attention was mainly on validating the computer suggestion, one could presume that by hiding the computer suggestion, users will be more accurate at spotting errors (in terms of incongruence). However, this is not what the second experiment showed. Turn did not seem to have an impact on decision correctness. Perhaps, the user was not negatively affected by the presence of the computer decision, as the Bahrami et al. (2010) study indicated. This could be explained by the users' ability to adequately judge automation reliability, suggested by the results of match and solution source. In the low (25%) reliability, condition, users are more likely to rely on their own response, while in the high (81%) reliability condition, users tend to accept the automation's recommendation. This is an interesting finding because it suggests that humans are sensitive to automation reliability even in the absence of performance feedback, while previous studies tended to employ feedback (Dzindolet et al., 2003; Madhavan et al., 2006).

It was hypothesised that the inclusion of reasoning with the decision would have the effect of increasing decision accuracy by: i) allowing the agents to communicate, justifying their decisions, and ii) slowing the users down and making them 'think twice'. While, the added task of filling in a form did have a negative impact on decision time, it did not lead to a better performance. Form-filling (or, reporting) is still a very big part of what traffic operators do and, while this is a good means of keeping track of what happened, or of externalising institutional procedures, it might not help them do a better job. In fact, this action of form-filling might be perceived as an additional, possibly irrelevant task to the job of 'managing' traffic. This leads to the idea that form-filling could be automated, thus saving a considerable amount of time (by nearly 40%, in this study).

The ability of the dyad members to communicate, justifying their decisions, did not influence decision accuracy in the low reliability condition and, in this sense, results are coherent with previous research which stated that communication is not necessary for two heads to be better than one (Bang et al., 2014; Koriat, 2012). However, in the high reliability condition, task did influence on user performance, but not as one might expect from the Bahrami et al. (2010) study. Form-filling reduced decision accuracy, rather than increase it. This could be explained by the fact that users perceived the action of form-filling as an extra task which was not necessarily related to the main task of managing traffic. This idea was also supported by subsequent discussions with some participants. It may be that these two tasks were perceived as having an equal priority and the first task suffering a reduced attendance due to the addition of the second.

In terms of UI design, the finding that form-filling slows down decision making and that decision accuracy suffers in the high reliability condition suggests that a change in Format should be made. Removing the form-filling task could potentially benefit the operators. However, as there is no decrease in decision accuracy in the low reliability condition (when form-filling was employed), form-filling could be used in order to inform and train automation (on-line learning) in non-time-critical situations (for example, in the case of scheduled road works instead of traffic accidents).

A further interesting finding of this experiment is related to the notions of Authority and Responsibility. In situations were the user gave the same answer as the computer (i.e. decision match occurred), users tended to select the computer as the final solution source. This could indicate that, in these situations users preferred that the computer was held accountable for the decision. Dekker (2002) and Woods and Cook (2002) suggested responsibility should be shared between them and the automation. This finding points towards a change in display Format: it may be more desirable to give operators the opportunity to over-rule computer decisions, thus giving them the final authority and, therefore, responsibility over the outcome of their decision, but only in cases where mismatch occurs. In cases where user and computer decisions match, it might be more desirable consider automation as the final authority and, thus, holding the computer responsible in the event of improper operation.

## 6.4 Experiment 3 – Form and Format

Two user interfaces for a traffic management application have been tested. However, UI2 showed a higher integration of information sources (in terms of task proximity). User behaviour in terms of decision correctness and decision time was measured with each user interface and with varying levels of computer reliability, in a simulated traffic monitoring task.

In terms of decision correctness, users were positively affected by the increase in computer reliability, a result which is consistent with the two experiments previously presented. However, decision correctness did not differ between the two interfaces.

This was an expected behaviour, as the UIs did not differ in terms of Information Content.

In terms of Joint Decision Making, it seems that participants were able to adequately judge the reliability of automation even in the absence of feedback and in lieu of communication between them and automation. This is an effect that was seen in the previous experiment (CHAPTER 4) as well. This can be inferred from the fact that, match levels are approximately equal to automation reliability in each condition. In contrast to what researchers such as Dzindolet (Dzindolet et al., 2003) suggested, effects of poor automation on user reliance do not persist over experimental blocks, but users re-evaluate their position regarding automation reliability in a continuous fashion.

Lu Wang et al. (2009) have found that displaying automation reliability to users has a positive effect on their reliance on automation. However, this leaves out uncertain situations, where the computer cannot accurately judge its correctness. For example, automation could compute a result based on corrupt data. The computer can have a high confidence in its answer, but it can be completely wrong in terms of the real situation, as it may not possess all the data required to make a decision. An example of this type of automation error is shown in experiments 1 and 2, where automation computes a correct answer based on the data in the graph, but the consideration of contextual information illustrates inconsistencies in the data (displayed as information source incongruence). Moreover, findings from the last experiment (4) suggest that strategic conformance can appear when automation confidence is shown to the user. Therefore, knowing that users are able to judge automation reliability, seeing strategic conformance occurring and understanding that, in uncertain situations, automation is not the best judge of its reliability level, leads to the conclusion that displaying automation confidence along with its decision is not necessarily the best design choice, in terms of Format. A better approach might be to make the user aware of the data used to make a particular decision, or, more abstracted, the reasons for making the decision.

When looking at the two user interfaces individually, we see no significant differences between the three reliability levels when using UI1. However, when using UI2, users were more correct when in the high reliability condition as compared to the low and medium conditions. It seems that the overall effect of reliability on decision performance is due to the results produced when UI2 was used. This leads to the conclusion that the higher degree of integration in UI2 somehow results in the user trusting the computer more than when using UI1, thus showing a higher conformance. Perhaps, the advantages of responding quicker to an event outweighed the advantages of a more scrutinous attitude towards the computer recommendation, when using the second interface. Moreover, it may be that the placement of information sources all across the first UI (requiring integrative processing) encouraged the search for information more than having all the information in one place. This could have lead to the slightly higher decision correctness with UI1 (even though not significant).

Finally, while there is a definite advantage in terms of decision time of using a more integrated display, the results of this study hint at the fact that there may be a loss in decision accuracy. Although, there was a time advantage when using UI2, this was not reflected in the reporting of subjective workload. The subjective usability metric (SUS), however, showed a clear preference for UI2.

In terms of display design, this study points towards an advantage of using a display which supports integrative processing (UI2 shows a higher degree of integration). However, even though it was not preferred by the users, results indicated that using UI1 might increase decision accuracy in case of lower automation reliability. Perhaps, the solution regarding the final Form of the display sits somewhere in the middle. It might be more appropriate to show the users an non-integrated display (UI1) when automation has a low level of reliability and an integrated display (UI2), otherwise.

## 6.5 Summarising Results in Terms of Content, Format and Form

| **EXP** | **Research Question** | **Display aspect** | | | **Effect on User Behaviour** | **Implication for Further Design** |
|---|---|---|---|---|---|---|
| | | **Content** | **Format** | **Form** | | |
| **1** | 1 | | | data which was required to check for informatio | Users were unable to judge the 'worth' of checking for | - increase salience of information sources |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  |  | n source congruence was spread across the screen (required integrative processing) | information congruence leading to:<br>- low performance | - integrate information sources with similar % viewing time in a single ROI<br>- increase legibility of ROIs with higher maximum dwell times |
| **2** | 1, 2 |  | form-filling as a means to give reasoning and for the purpose of documentation |  | Form-filling was perceived as separate from the main task leading to:<br>- decreased performance<br>- increased decision time | - remove form-filling<br>- require only the computer to justify recommendation |
| **2** | 1, 2 |  | ability to choose final solution source |  | Participants relied on the computer when their answers matched, hinting towards shared responsibility | - opportunity for the user to over-rule computer decision in case of mismatch, otherwise sharing responsibility |
| **3** | 1, 2 |  |  | UI supporting | - using a display | - show users an non-integrated |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  |  | non-integrative vs integrative processing | showing more integration lead to faster responses<br>- results also hinted at the possibility of lower decision accuracy when using the display showing higher integration | display (UI1) when automation has a low reliability level and an integrated display (UI2) otherwise |
| **2, 3** | 1, 2 | automation confidence was not displayed |  |  | - 2 and 3 showed that users can adequately judge automation reliability in absence of feedback | - might be better to not display confidence but the reason for the automated decision |

### 6.5.1 Socio-Technical Constraints on UI Design

The process of design of the SPEEDD UIs has raised the idea that the influence of socio-technological constraints on interface design can be categorised and tracked in terms of the CFF taxonomy. This lead to the compilation of the table in APPENDIX I . This table is summarised in Table 6.1. A diagram (see Figure 6.1) has been developed as a result of the knowledge gathered from undergoing the design process for both SPEEDD use-cases and the several forms of evaluation ran as part of this PhD. This is an updated version of the diagram presented in Figure 2.6 (CHAPTER 2).

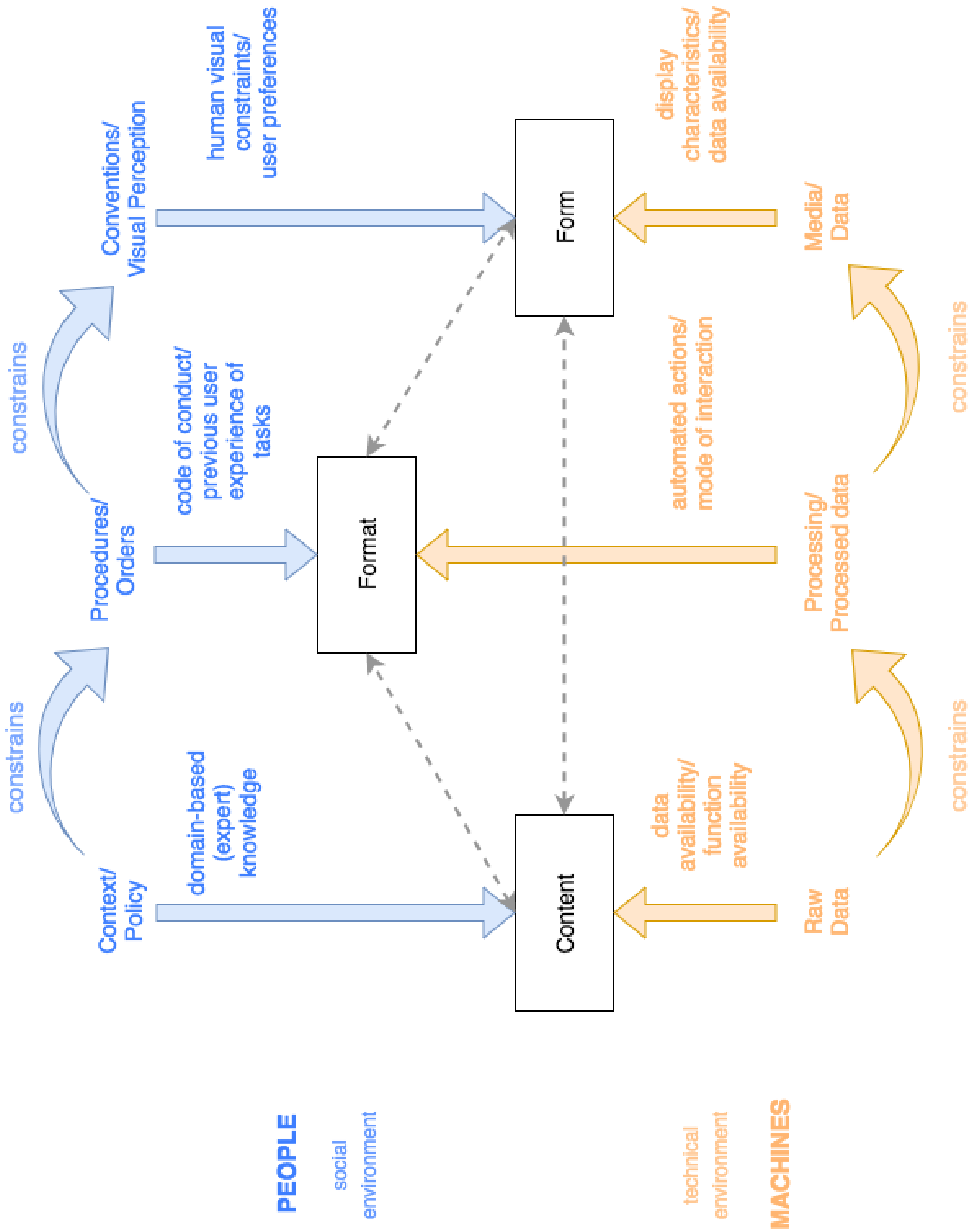

PEOPLE
social environment
technical environment
MACHINES
Context/ Policy
constrains
Procedures/ Orders
constrains
Conventions/ Visual Perception
domain-based (expert) knowledge
code of conduct/ previous user experience of tasks
human visual constraints/ user preferences
Content
Format
Form
data availability/ function availability
automated actions/ mode of interaction
display characteristics/ data availability
Raw Data
constrains
Processing/ Processed data
constrains
Media/ Data

**Figure 6.1 - Socio-technological Constraints on User Interface Design**

Our experience has shown that Content, Format and Form are interrelated. Changing the Form of a UI component might lead to Format being affected, whereas changing the Content might lead to a change in Format and Form as well. Examples of such events happening can be seen in APPENDIX I in places where we see marks in more than one UI dimension. Let's look at change number 7. The addition of information regarding driver behaviour in both directions of the Grenoble ring road (change in Content), lead to the information source looking differently (change in Form).

Moreover, this interrelation of Content, Format and Form also became apparent as a result of the experiments presented in this thesis. Changing the Form of the main map in experiment 3 (CHAPTER 5) to a road schematic, also changed the way in which users interacted with it and the layout of the UI (Format).

**Table 6.1 - Instances of Social- and Technical-driven Changes in UI Design**

| **Environment** | **Social** | | | **Technical** | | |
|---|---|---|---|---|---|---|
| **Dimension** | **Content** | **Format** | **Form** | **Content** | **Format** | **Form** |
| **Reason** | domain-based knowledge | a) company politics regarding operations/ code of conduct<br>b) previous user experience of performing the task; | a) human visual constraints<br>b) user preference | a) data availability<br>b) system function availability | a) automated functions<br>b) specific mode of interaction with background automation | a) data availability |
| **Instances in APPENDIX I (reason)** | 4, 7, 9, 10, 18, 20, 27, 31, 33, 26 | 9 (a), 17 (b), 22 (b), | 7 (b), 31 (b), 32 (a), 33 (b) | 1 (a), 2 (b), 4 (ab), 5 (a), 12 (ab), 21 (a), 27 (a), | 11 (ab), 13 (b), 22 (b) | 15, 16, 21, 23 |

#### *6.5.1.1 Design Considerations*

The table in APPENDIX I lists D (design consideration) as a driver for UI changes in several instances. However, in Table 6.1 and Figure 6.1, Design is not an environment that can pose constraints on the UI. This is because Design is the space that is subject to constraints from the social and technical environments. D refers to changes in UI design that have been initiated by the designer, independently from the social and technical constraints and based on past research and previous design experience.

#### *6.5.1.2 Social Constraints on UI Design*

APPENDIX I lists all the changes that the UIs for the TM use-case underwent along with the drives behind them. Looking at these changes in terms of Content, Format and Form, we are able to identify what imposed these constraints on each UI dimension. Taking the perspective of the social environment, Content seems to have been driven exclusively by domain-based knowledge. APPENDIX I confirms this fact. This is an expected result considering that the information requirements were the output of CWA along with expert interviews. Let us take design change number 7, for example (see APPENDIX I ). The driver behaviour window was modified (in version 1.5 from version 1.0) to include both north- and south-bound traffic on the ring road. This requirement came up in the interviews with the DIR-CE Grenoble traffic managers, when discussing UI version 1.0.

In terms of Format, two main social drives appear to have constrained the design of the SPEEDD UIs. They are code of conduct and previous user experience of performing a particular task. The former can be exemplified by change number 9, while the latter by change 17. Change number 9 was marked by the addition of the Activity window in the second version of the TM UI. This change came as a result of discussions with traffic managers in Grenoble, who stated that activity logging is one of their primary responsibilities. The map in the third version of the TM UI was modified so that it does not pan to the location of a detected or predicted congestion. This is due to the fact that operators were used to a static map and they found the automatic zooming in and panning to a location somewhat distracting.

All changes in the dimension of Form (looking at the social environment) are found to be driven by one of two factors: human visual constraints and user preference. For example, the circular design was replaced with a radial design (change 32) because operators found the ramps hard to read due to the circular placement. User preference, however, was a driver for change 33, where the Grenoble map was added in background of the schematic road (TM UI). This was done because the operators were used to see all road intersections felt like there was a loss of context when moving from the interactive map visualisation to the schematic representation of the ring road.

#### *6.5.1.3 Technical Constraints on UI Design*

Looking at APPENDIX I  and the diagram showing how the socio-technical environment constrains User Interface design (see Figure 6.1) we can see how the underlying architecture made its mark on the UI. First, let's take dimension of Content. The designer can only show on screen data that is available somewhere in the system or information derived from a number data points available in the system. The key word to note here is 'availability'. The UI should not show information that is not available in the overall system, otherwise it would be meaningless - or, potentially more serious - misleading to an operator trying to control that system, or to an analyst investigating a case. Likewise, the UI must not display control actions that are not supported in the underlying architecture.

Table 6.1 illustrates examples where data availability and function availability within the runtime architecture constrained the SPEEDD UIs throughout the design process. Let's take change 1, where the Road User Goals window which was present on the Initial Layout of the TM UI, was not implemented in the first version of the UI because the data were not available in the technical environment. Change 2, however illustrates a case where the Open tasks and scheduled events window was not integrated in the first prototype TM UI because there was no database to store these data incorporated in the runtime architecture.

The dimension of Format is informed and constrained by the way in which processed data are handled in the system, by function availability and by the 'agreed' source of the course of action. The first issue relates to whether the processed data are displayed to the user or it is used internally as an input to a separate automated module. The second issue is concerned simply with whether the function in question is implemented at the technical side. The last issue refers the agent which decides the action to be taken in a particular situation; be it the computer, or the operator/analyst. This can vary from system to system, from situation to situation and it can even be adaptive, in that it can change within a system and within a particular type of situations. For example (shown in APPENDIX I , see 11) is the simplification of the Control Panel window in version 2.0 of the TM UI, due to the fact that fine-tuning of the ramp rates was assigned to automation, leaving the human operator with the task of monitoring them and setting bounds for the ramp rate values.

It is less obvious, however, how the dimension of Form is constrained by the technological environment. Indeed, the underlying system does not have a lot to say regarding the Form aspect of UI design, but the hardware on which the system runs does. Here, we speak of physical displays (projector, big screen, desk monitor, multiple or single screens, etc.), input devices (mouse, keyboard, joystick, custom keyboards, touchscreens, touchpads, microphones, movement sensors (such as Kinect) etc.). For example, if the operator will use a touchscreen to interact with the UI, buttons have to be bigger than in the case when one uses a conventional mouse + keyboard setup. Moreover, in the first situation, the designer is to avoid textual inputs at all costs. Another important aspect is button placement. This is dictated also by the positioning of the screen, so that the designer might want to place buttons close to hand (in the case where interaction speed is of prime importance), or conversely in a hard to reach position (in cases where human interaction is considered undesirable). In the context of the SPEEDD project, traffic operators use single desk-mounted screen and, as input modalities, they use a mouse and a keyboard, thus the dimension of Form is far less constrained by media in the case of alternative display and input devices. These interaction media were known prior to the design of the SPEEDD UIs. Therefore, constraints generated by them are not part of the list of changes, since the UIs have been designed for these media.

However, there are other technical factors that constrain UI Form, i.e. data availability. This is exemplified by change number 21 (see APPENDIX I ). In version 3.0 of the TM UI, the activity log went through a complete redesign from previous versions due to the technical requirement of displaying more data available within the architecture for the purpose of monitoring the correct operation of the automation.

## 6.6 Limitations and Further Research

The User Interface Design example (section 1.7) illustrated how Content/Format/Form (CFF) can be used to make more informed design choices and more clearly track them, while the experiments presented in this PhD started to show how CFF can be used in the experimental evaluation of User Interface designs. The work presented in this thesis is limited to the investigation of a relatively small number of changes in the dimensions of Format and Form. Varying Content has not been investigated due to the large body of literature that shows how information requirements for a certain work-domain can be extracted. The studies presented in this thesis show some interesting findings but are nonetheless limited in some respects.

Experiment 1 (CHAPTER 3) looked from the perspective of PCP at how users understand automation errors, however the overall automation reliability was very low (25%). This issue was further addressed in experiments 2 (CHAPTER 4) and 3 (CHAPTER 5). Due to the unavailability of the Eye-Tracking device in further studies, the results related to CFF from the metrics derived in experiment 1 require further validation in other studies.

In the second experiment (CHAPTER 4), Form in terms of layout of the Ramp Metering Control window was kept the same (i.e. the user answer was always shown above the computer answer, regardless of whether the computer or the user went first) (Figure 4.4). An interesting question for further research to investigate would be whether the layout (the position of the response dialog) of the Ramp Metering Control window would produce different effects in human behaviour. In this experiment we have also looked at how would showing the computer reasoning affect user behaviour in terms of decision time and decision correctness. However, a pitfall of the approach was that the user was also required to give his reasoning, so that any changes in user behaviour cannot be attributed solely to the computer showing its reasoning, but also to the fact that the user had to fill out his ‘form’. A further experiment could be designed to test how would the computer giving his reasoning, without also requiring the user to do the same, affect user behaviour.

The discussion of experiment 3 (CHAPTER 5) could have been enriched by the consideration of Form. This would have required that an additional difference (in the dimension of Form) would be introduced, possibly by having one UI show a lower display proximity. This discussion would have allowed for comparison with the final experiment, where it was shown that changes in user behaviour due to differences in Form can overshadow those due to differences in Format. However, the addition of this further independent variable would have made it harder to pin-point the causes for the change in human behaviour.

The socio-technical constraints diagram (Figure 6.1) is a considerable step forward for methodologies of User Interface design and evaluation, however, it is the result of the study of two use-cases and four experiments and further work is required to validate it.

# REFERENCES

Ahn, J., Taieb-Maimon, M., Sopan, A., Plaisant, C., Shneiderman, B., 2011. Temporal visualization of social network dynamics: prototypes for Nation of Neighbors, in: Social Computing, Behavioral-Cultural Modeling and Prediction. Springer, pp. 309–316.

Ahrens, J., Hendrickson, B., Long, G., Miller, S., Ross, R., Williams, D., 2011. Data-Intensive Science in the US DOE: Case Studies and Future Challenges. Comput. Sci. Eng. 13, 14–24. https://doi.org/10.1109/MCSE.2011.77

Baber, C., Starke, S., Chen, X., Morar, N., Howes, A., Cooke, N., Bak, P., 2014. Design of User Interface for SPEEDD Prototype.

Bahner, J.E., Hüper, A.-D., Manzey, D., 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. Int. J. Hum.-Comput. Stud. 66, 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., Frith, C.D., 2010. Optimally Interacting Minds. Science 329, 1081–1085. https://doi.org/10.1126/science.1185718

Bainbridge, L., 1983. Ironies of automation. Automatica 19, 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P.E., Lau, J.Y.F., Roepstorff, A., Rees, G., Frith, C.D., Bahrami, B., 2014. Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. Conscious. Cogn. 26, 13–23. https://doi.org/10.1016/j.concog.2014.02.002

Batty, M., 2013. Big data, smart cities and city planning. Dialogues Hum. Geogr. 3, 274–279. https://doi.org/10.1177/2043820613513390

Bennett, K.B., Flach, J.M., 2011. Display and Interface Design: Subtle Science, Exact Art. CRC Press.

Bennett, K.B., Nagy, A.L., Flach, J.M., 2012. Visual Displays, in: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics. John Wiley & Sons, Inc., pp. 1177–1208. https://doi.org/10.1002/9781118131350.ch42

Berbaum, K.S., El-Khoury, G.Y., Franken, E.A., Kuehn, D.M., Meis, D.M., Dorfman, D.D., Warnock, N.G., Thompson, B.H., Kao, S.C.S., Kathol, M.H., 1994. Missed fractures resulting from satisfaction of search effect. Emerg. Radiol. 1, 242–249. https://doi.org/10.1007/BF02614935

Berbaum, K.S., Franken, E.A.J., Dorfman, D.D., Rooholamini, S.A., Kathol, M.H., Barloon, T.J., Behlke, F.M., Sato, Y., Lu, C.H., El-Khoury, G.Y., Flickinger, F.W., Montgomery, W.J., 1990. Satisfaction of Search in Diagnostic Radiology. Invest. Radiol. 25, 133.

Bertin, J., 1983. Semiology of graphics: diagrams, networks, maps / Jacques Bertin ; translated by William J. Berg. The University of Wisconsin Press, Madison, Wis. ; London.

Borst, C., Bijsterbosch, V.A., Paassen, M.M. van, Mulder, M., 2017. Ecological interface design: supporting fault diagnosis of automated advice in a supervisory air traffic control task. Cogn. Technol. Work 19, 545–560. https://doi.org/10.1007/s10111-017-0438-y

Brooke, J., 1996. SUS-A quick and dirty usability scale, in: Usability Evaluation in Industry. Taylor and Francis., London, pp. 189–194.

Burns, C.M., Ho, G., Arrabito, G.R., 2011. Mapping Ecologically to Modalities. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 55, 335–339. https://doi.org/10.1177/1071181311551069

Byrne, E.A., Parasuraman, R., 1996. Psychophysiology and adaptive automation. Biol. Psychol., Psychophysiology of Workload 42, 249–268. https://doi.org/10.1016/0301-0511(95)05161-9

C. Melody Carswell, 1992. Choosing Specifiers: An Evaluation of the Basic Tasks Model of Graphical Perception. Hum. Factors 34, 535–554. https://doi.org/10.1177/001872089203400503

Carswell, C.M., Wickens, C.D., 1987. Information integration and the object display An interaction of task demands and display superiority. Ergonomics 30, 511–527. https://doi.org/10.1080/00140138708969741

Cook, K., Thomas, J., 2005. Illuminating the Path: The Research and Development Agenda for Visual Analytics.

Cossalter, M., Mengshoel, O.J., Selker, T., 2011. Visualizing and Understanding Large-Scale Bayesian Networks., in: Scalable Integration of Analytics and Visualization.

Cummings, M.L., Mastracchio, C., Thornburg, K.M., Mkrtchyan, A., 2013. Boredom and distraction in multiple unmanned vehicle supervisory control. Interact. Comput. 25, 34–47.

de Visser, E.J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., Parasuraman, R., 2012. The World is not Enough: Trust in Cognitive Agents. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 56, 263–267. https://doi.org/10.1177/1071181312561062

Debernard, S., Cathelain, S., Crévits, I., Poulain, T., 2002. AMANDA Project: Delegation of tasks in the air-traffic control domain. Presented at the COOP 2002: Cooperative Systems Design-A Challenge of the Mobility Age.

Dekker, S.W.A., 2002. Reconstructing human contributions to accidents: the new view on error and performance. J. Safety Res. 33, 371–385. https://doi.org/10.1016/S0022-4375(02)00032-4

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P., 2003. The role of trust in automation reliance. Int. J. Hum.-Comput. Stud., Trust and Technology 58, 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Ellis, G., Dix, A., 2006. An Explorative Analysis of User Evaluation Studies in Information Visualisation, in: Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization, BELIV ’06. ACM, New York, NY, USA, pp. 1–7. https://doi.org/10.1145/1168149.1168152

Few, S., 2013. Information Dashboard Design: Displaying Data for At-a-Glance Monitoring, Second Edition, Second edition edition. ed. Analytics Press, Burlingame, Calif.

Flach, J.M., Tanabe, F., Monta, K., Vicente, K.J., Rasmussen, J., 1998. An Ecological Approach to Interface Design. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 42, 295–299. https://doi.org/10.1177/154193129804200324

Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., Beller, J., 2011. Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. Cogn. Technol. Work 14, 3–18. https://doi.org/10.1007/s10111-011-0191-6

Folds, D., Brooks, J., Stocks, D., Fain, W., Courtney, T., Blankenship, S., 1993. Functional Definition of an Ideal Traffic Management System. Atlanta Ga. Ga. Tech Res. Inst.

Garin, F., Baber, C., Starke, S., Morar, N., Howes, A., Kofman, A., 2015. Evaluation of SPEEDD prototype 1 for Road Traffic Management (No. D8.3).

Germanwings crash: Co-pilot Lubitz “practised rapid descent” [WWW Document], 2015. . BBC News. URL http://www.bbc.co.uk/news/world-europe-32604552 (accessed 4.21.16).

Gibson, J.J., 2014. The Ecological Approach to Visual Perception: Classic Edition. Psychology Press.

Greef, T. de, Lafeber, H., Oostendorp, H. van, Lindenberg, J., 2009. Eye Movement as Indicators of Mental Workload to Trigger Adaptive Automation, in: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (Eds.), Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 219–228.

Greengard, S., 2009. Making Automation Work. Commun ACM 52, 18–19. https://doi.org/10.1145/1610252.1610261

Griethe, H., Schumann, H., 2006. Visualizing uncertainty for improved decision making. SimVis.

Halford, G.S., Baker, R., McCredden, J.E., Bain, J.D., 2005. How Many Variables Can Humans Process? Psychol. Sci. 16, 70–76. https://doi.org/10.1111/j.0956-7976.2005.00782.x

Ham, D.-H., Yoon, W.C., 2001. The effects of presenting functionally abstracted information in fault diagnosis tasks. Reliab. Eng. Syst. Saf. 73, 103–119. https://doi.org/10.1016/S0951-8320(01)00053-9

Hancock, P.A., Scallen, S.F., 1996. The future of function allocation. Ergon. Des. Q. Hum. Factors Appl. 4, 24–29.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, in: Hancock, P.A., Meshkati, N. (Eds.), Advances in Psychology, Human Mental Workload. North-Holland, pp. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Helander, M.G., Landauer, T.K., Prabhu, P.V., 1997. Handbook of Human-Computer Interaction. Elsevier.

Hoc, J.-M., 2001. Towards a cognitive approach to human–machine cooperation in dynamic situations. Int. J. Hum.-Comput. Stud. 54, 509–540. https://doi.org/10.1006/ijhc.2000.0454

Hoc, J.-M., 2000. From human – machine interaction to human – machine cooperation. Ergonomics 43, 833–843. https://doi.org/10.1080/001401300409044

Hoc, J.-M., Debernard, S., 2002. Respective Demands of Task and Function Allocation on Human-Machine Cooperation Design: a Psychological Approach (AAAI Technical Report No. WS-02-03). AAAI.

Hoc, J.-M., Lemoine, M.-P., 1998. Cognitive Evaluation of Human-Human and Human-Machine Cooperation Modes in Air Traffic Control. Int. J. Aviat. Psychol. 8, 1–32. https://doi.org/10.1207/s15327108ijap0801_1

Hoff, K.A., Bashir, M., 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Hum. Factors 57, 407–434. https://doi.org/10.1177/0018720814547570

Hollnagel, E., 2001. Extended cognition and the future of ergonomics. Theor. Issues Ergon. Sci. 2, 309–315. https://doi.org/10.1080/14639220110104934

Hollnagel, E., 1987. Information and reasoning in intelligent decision support systems. Int. J. Man-Mach. Stud. 27, 665–678. https://doi.org/10.1016/S0020-7373(87)80023-8

Human engineering for an effective air-navigation and traffic-control system, 1951. . National Research Council, Div. of, Oxford, England.

Inagaki, T., 2003. Automation and the cost of authority. Int. J. Ind. Ergon., Selected papers from the second cyberspace conference on ergonomi cs, CybErg 1999 31, 169–174. https://doi.org/10.1016/S0169-8141(02)00193-2

Jamieson, G.A., Vicente, K.J., 2001. Ecological interface design for petrochemical applications: supporting operator adaptation, continuous learning, and distributed, collaborative work. Comput. Chem. Eng. 25, 1055–1074. https://doi.org/10.1016/S0098-1354(01)00678-0

Johnson, A.W., Oman, C.M., Sheridan, T.B., Duda, K.R., 2014. Dynamic task allocation in operational systems: Issues, gaps, and recommendations, in: 2014 IEEE Aerospace Conference. Presented at the 2014 IEEE Aerospace Conference, pp. 1–15. https://doi.org/10.1109/AERO.2014.6836205

Kammerer, Y., Gerjets, P., 2010. How the Interface Design Influences Users' Spontaneous Trustworthiness Evaluations of Web Search Results: Comparing a List and a Grid Interface, in: Proceedings of the 2010 Symposium on Eye-Tracking Research &#38; Applications, ETRA '10. ACM, New York, NY, USA, pp. 299–306. https://doi.org/10.1145/1743666.1743736

Kaniarasu, P., Steinfeld, A., Desai, M., Yanco, H., 2012. Potential measures for detecting trust changes, in: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp. 241–242.

Kibangou, A., Morbidi, F., Schmitt, E., Hempel, A., Baber, U.C., Cooke, N., 2015. User requirements and scenario definition.

Kim, J., Moon, J.Y., 1998. Designing towards emotional usability in customer interfaces—trustworthiness of cyber-banking system interfaces. Interact. Comput., HCI and Information Retrieval 10, 1–29. https://doi.org/10.1016/S0953-5438(97)00037-4

Koriat, A., 2012. When Are Two Heads Better than One and Why? Science 336, 360–362. https://doi.org/10.1126/science.1216549

Lee, J., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J.D., 2008. Review of a Pivotal Human Factors Article: “Humans and Automation: Use, Misuse, Disuse, Abuse.” Hum. Factors J. Hum. Factors Ergon. Soc. 50, 404–410. https://doi.org/10.1518/001872008X288547

Lee, J.D., Moray, N., 1994. Trust, self-confidence, and operators’ adaptation to automation. Int. J. Hum.-Comput. Stud. 40, 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J.D., See, K.A., 2004. Trust in Automation: Designing for Appropriate Reliance. Hum. Factors J. Hum. Factors Ergon. Soc. 46, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Licklider, J.C.R., 1960. Man-Computer Symbiosis. IRE Trans. Hum. Factors Electron. HFE-1, 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Lu Wang, Greg A. Jamieson, Justin G. Hollands, 2009. Trust and Reliance on an Automated Combat Identification System. Hum. Factors 51, 281–291. https://doi.org/10.1177/0018720809338842

Lyons, J.B., Stokes, C.K., 2011. Human–Human Reliance in the Context of Automation. Hum. Factors J. Hum. Factors Ergon. Soc. 0018720811427034. https://doi.org/10.1177/0018720811427034

Madhavan, P., Wiegmann, D.A., 2007. Similarities and differences between human–human and human–automation trust: an integrative review. Theor. Issues Ergon. Sci. 8, 277–301. https://doi.org/10.1080/14639220500337708

Madhavan, P., Wiegmann, D.A., Lacson, F.C., 2006. Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. Hum. Factors J. Hum. Factors Ergon. Soc. 48, 241–256. https://doi.org/10.1518/001872006777724408

Masalonis, A.J., Duley, J.A., Parasuraman, R., 1999. Effects of Manual and Autopilot Control on Mental Workload and Vigilance During Simulated General... Transp. Hum. Factors 1, 187.

Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An Integrative Model of Organizational Trust. Acad. Manage. Rev. 20, 709–734. https://doi.org/10.5465/AMR.1995.9508080335

McIlroy, R.C., Stanton, N.A., 2015. Ecological Interface Design Two Decades On: Whatever Happened to the SRK Taxonomy? IEEE Trans. Hum.-Mach. Syst. 45, 145–163. https://doi.org/10.1109/THMS.2014.2369372

Merritt, S.M., Heimbaugh, H., LaChapell, J., Lee, D., 2013. I Trust It, but I Don’t Know Why: Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. Hum. Factors J. Hum. Factors Ergon. Soc. 55, 520–534. https://doi.org/10.1177/0018720812465081

Meyer, J., Feinshreiber, L., Parmet, Y., 2003. Levels of automation in a simulated failure detection task, in: IEEE International Conference on Systems, Man and Cybernetics, 2003. Presented at the IEEE International Conference on

Systems, Man and Cybernetics, 2003, pp. 2101–2106 vol.3. https://doi.org/10.1109/ICSMC.2003.1244194

Morar, N., Baber, C., Duncan, A., Bak, P., 2015a. What You See Is What You Do: applying Ecological Interface Design to Visual Analytics. Workshop Proc. EDBTICDT 2015 Jt. Conf. March 27 2015 Bruss. Belg. 1330, 125–131.

Morar, N., Baber, C., Starke, S., Fournier, F., 2015b. Missing Key Information: How Automation Failure Can Be Misinterpreted, in: International Annual Meeting of the Human Factors and Ergonomics Society.

Mueller, K., Garg, S., Nam, J.E., Berg, T., McDonnell, K.T., 2011. Can Computers Master the Art of Communication?: A Focus on Visual Analytics. IEEE Comput. Graph. Appl. 31, 14–21. https://doi.org/10.1109/MCG.2011.39

Muir, B.M., 1987. Trust between humans and machines, and the design of decision aids. Int. J. Man-Mach. Stud. 27, 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

Natan Morar, Chris Baber, 2017. Joint Human-Automation Decision Making in Road Traffic Management. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 61, 385–389. https://doi.org/10.1177/1541931213601578

Onnasch, L., Wickens, C.D., Li, H., Manzey, D., 2014. Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. Hum. Factors J. Hum. Factors Ergon. Soc. 56, 476–488. https://doi.org/10.1177/0018720813501549

Parasuraman, R., Cosenzo, K.A., De Visser, E., 2009. Adaptive Automation for Human Supervision of Multiple Uninhabited Vehicles: Effects on Change Detection, Situation Awareness, and Mental Workload. Mil. Psychol. April 2009 21, 270–297. https://doi.org/10.1080/08995600902768800

Parasuraman, R., Manzey, D.H., 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. Hum. Factors J. Hum. Factors Ergon. Soc. 52, 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., Mouloua, M., Molloy, R., 1996. Effects of Adaptive Task Allocation on Monitoring of Automated Systems. Hum. Factors J. Hum. Factors Ergon. Soc. 38, 665–679. https://doi.org/10.1518/001872096778827279

Parasuraman, R., Riley, V., 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. Hum. Factors J. Hum. Factors Ergon. Soc. 39, 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Wickens, C.D., 2008. Humans: Still Vital After All These Years of Automation. Hum. Factors 50, 511–520. https://doi.org/10.1518/001872008X312198

Philip Chen, C.L., Zhang, C.-Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf. Sci. 275, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

Rani, P., Sarkar, N., Adams, J., 2007. Anxiety-based affective communication for implicit human–machine interaction. Adv. Eng. Inform. 21, 323–334. https://doi.org/10.1016/j.aei.2006.11.009

Rasmussen, J., 1983. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Trans. Syst. Man Cybern. SMC-13, 257–266. https://doi.org/10.1109/TSMC.1983.6313160

Rasmussen, J., Vicente, K.J., 1989. Coping with human errors through system design: implications for ecological interface design. Int. J. Man-Mach. Stud. 31, 517–534. https://doi.org/10.1016/0020-7373(89)90014-X

Ricci, F., Rokach, L., Shapira, B., 2011. Introduction to Recommender Systems Handbook, in: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.), Recommender Systems Handbook. Springer US, pp. 1–35. https://doi.org/10.1007/978-0-387-85820-3_1

Rovira, E., Cross, A., Leitch, E., Bonaceto, C., 2014. Displaying Contextual Information Reduces the Costs of Imperfect Decision Automation in Rapid Retasking of ISR Assets. Hum. Factors J. Hum. Factors Ergon. Soc. 0018720813519675. https://doi.org/10.1177/0018720813519675

Samuel, S., Kundel, H.L., Nodine, C.F., Toto, L.C., 1995. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. Radiology 194, 895–902. https://doi.org/10.1148/radiology.194.3.7862998

Schmidt, K., Rasmussen, J., Brehmer, B., Leplat, J., 1991. Cooperative work: A conceptual framework. Distrib. Decis.-Mak. Cogn. Models Coop. Work 75–110.

Sheridan, T.B., 2002. Humans and Automation: System Design and Research Issues. John Wiley &amp; Sons, Inc., New York, NY, USA.

Sheridan, T.B., 2000. Function allocation: algorithm, alchemy or apostasy? Int. J. Hum.-Comput. Stud. 52, 203–216. https://doi.org/10.1006/ijhc.1999.0285

Simons, D.J., Levin, D.T., 1997. Change blindness. Trends Cogn. Sci. 1, 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2

Sims, J., Vashishtha, D., Rani, P., Brackin, R., Sarkar, N., 2002. Stress detection for implicit human-robot co-operation, in: Automation Congress, 2002 Proceedings of the 5th Biannual World. Presented at the Automation Congress, 2002 Proceedings of the 5th Biannual World, pp. 567–572. https://doi.org/10.1109/WAC.2002.1049497

Sinha, R., Swearingen, K., 2002. The Role of Transparency in Recommender Systems, in: CHI ’02 Extended Abstracts on Human Factors in Computing Systems, CHI EA ’02. ACM, New York, NY, USA, pp. 830–831. https://doi.org/10.1145/506443.506619

Stanton, N.A., 2006. Hierarchical task analysis: Developments, applications, and extensions. Appl. Ergon., Special Issue: Fundamental Reviews 37, 55–79. https://doi.org/10.1016/j.apergo.2005.06.003

Starke, S.D., Baber, C., Cooke, N.J., Howes, A., 2017. Workflows and individual differences during visually guided routine tasks in a road traffic management control room. Appl. Ergon. 61, 79–89. https://doi.org/10.1016/j.apergo.2017.01.006

Stevens, S.S., 1946. On the Theory of Scales of Measurement,. Science 103, 677–680.

Tintarev, N., Masthoff, J., 2012. Evaluating the effectiveness of explanations for recommender systems. User Model. User-Adapt. Interact. 22, 399–439. https://doi.org/10.1007/s11257-011-9117-5

Tintarev, N., Masthoff, J., 2007. A Survey of Explanations in Recommender Systems, in: 2007 IEEE 23rd International Conference on Data Engineering Workshop. Presented at the 2007 IEEE 23rd International Conference on

Data Engineering Workshop, pp. 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

Townsend, A.M., 2013. Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia. W. W. Norton & Company.

Upton, C., Doherty, G., 2008. Extending Ecological Interface Design principles: A manufacturing case study. Int. J. Hum.-Comput. Stud. 66, 271–286. https://doi.org/10.1016/j.ijhcs.2007.10.007

Vanderhaegen, F., Crevits, I., Debernard, S., Millot, P., 1994. Human-machine cooperation: Toward an activity regulation assistance for different air traffic control levels. Int. J. Human–Computer Interact. 6, 65–104. https://doi.org/10.1080/10447319409526084

Vicente, K.J., 1999. Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. CRC Press.

Vicente, K.J., Christoffersen, K., Pereklita, A., 1995. Supporting operator problem solving through ecological interface design. IEEE Trans. Syst. Man Cybern. 25, 529–545. https://doi.org/10.1109/21.370186

Vicente, K.J., Rasmussen, J., 1992. Ecological interface design: theoretical foundations. IEEE Trans. Syst. Man Cybern. 22, 589–606. https://doi.org/10.1109/21.156574

Vicente, K.J., Rasmussen, J., 1990. The Ecology of Human-Machine Systems II: Mediating “Direct Perception” in Complex Work Domains. Ecol. Psychol. 2, 207–249. https://doi.org/10.1207/s15326969eco0203_2

Wanner, F., Jentner, W., Schreck, T., Stoffel, A., Sharalieva, L., Keim, D.A., 2015. Integrated visual analysis of patterns in time series and text data - Workflow and application to financial data analysis. Inf. Vis. 1473871615576925. https://doi.org/10.1177/1473871615576925

Wickens, C.D., Carswell, C.M., 1995. The Proximity Compatibility Principle: Its Psychological Foundation and Relevance to Display Design. Hum. Factors 37, 473.

Wickens, C.D., Clegg, B.A., Vieane, A.Z., Sebok, A.L., 2015. Complacency and Automation Bias in the Use of Imperfect Automation. Hum. Factors J. Hum. Factors Ergon. Soc. 0018720815581940. https://doi.org/10.1177/0018720815581940

Wickens, C.D., Dixon, S.R., 2007. The benefits of imperfect diagnostic automation: a synthesis of the literature. Theor. Issues Ergon. Sci. 8, 201–212. https://doi.org/10.1080/14639220500370105

Willsher, K., 2015. Germanwings crash: co-pilot practised descent move on previous flight. The Guardian.

Woods, D., Dekker, S., 2000. Anticipating the effects of technological change: A new era of dynamics for human factors. Theor. Issues Ergon. Sci. 1, 272–282. https://doi.org/10.1080/14639220110037452

Woods, D.D., 1985. Cognitive Technologies: The Design of Joint Human-Machine Cognitive Systems. AI Mag. 6, 86. https://doi.org/10.1609/aimag.v6i4.511

Woods, D.D., Cook, R.I., 2002. Nine Steps to Move Forward from Error. Cogn. Technol. Work 4, 137–144. https://doi.org/10.1007/s101110200012

Zhang, J., 1996. A representational analysis of relational information displays. Int. J. Hum.-Comput. Stud. 45, 59–74. https://doi.org/10.1006/ijhc.1996.0042

Zhang, J., Norman, D.A., 1994. Representations in Distributed Cognitive Tasks. Cogn. Sci. 18, 87–122. https://doi.org/10.1207/s15516709cog1801_3

# APPENDIX I

## 6.7 Changes in UI Design and drives for them related to CFF

| Change No. | Use-Case – UI version | Design Change | Driven by – reason<br>(Technical domain – T, sociological domain – S, Design consideration – D) | Content | Format | Form |
|---|---|---|---|---|---|---|
| **1** | TM – 1.0 | removed Road User Goals window | T – data not available | X | | |
| **2** | TM – 1.0 | Removed open tasks, scheduled events, etc. | T – no ability to store user events | X | | |
| **3** | TM – 1.0 | Colour difference between predicted and detected congestion | D | | | X |
| **4** | TM – 1.0 | CCTV feed not integrated in architecture, mock CCTV used (Google Maps StreetView) | T – CCTV not part of the architecture<br>S – CCTV was required by operators | X | | |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **5** | TM – 1.0 | Driver behaviour window shows only average speed of traffic and average distance between drivers | T – only data available in the architecture bus regarding driver behaviour | X |  |  |
| **6** | TM – 1.5 | Increased ease of selection of individual ramps | D |  | X |  |
| **7** | TM – 1.5 | Added north- and south-bound traffic information for Driver behaviour window | S – Grenoble Ring Road had traffic going in both directions | X |  | X |
| **8** | TM – 1.5 | Suggested Actions window was added | D – window in which computer could display control recommendations |  | X |  |
| **9** | TM – 2.0 | Activity window added | S – logging is a primary activity that they perform and the presence of a log is mandatory | X | X |  |
| **10** | TM – 2.0 | Driver behaviour window removed | S – operators’ work does not involve the direct control of road users’ behaviour, this being achieved by long-term governmental campaigns | X |  |  |

| | | | | | | |
|---|---|---|---|---|---|---|
| **11** | TM – 2.0 | Simplified control panel window where the user can set bounds for the ramp metering control unit and not absolute values | T – fine adjustments to ramp rates is automated | | X | |
| **12** | TM – 2.0 | Removed lane closures and variable message signs | T – not dealt with in SPEEDD architecture | X | | |
| **13** | TM – 2.0 | Suggested Actions window has been removed | T - outputs of the automated system are concerned with ramp metering levels to be applied at each particular ramp and not as suggestions of what actions the user should perform | | X | |
| **14** | TM – 2.0 | enlargement of the map | D | | | X |
| **15** | TM – 2.0 | integration of the CCTV window into the map window | D – data availability | | | X |
| **16** | TM – 2.0 | Sensor names and locations linked to map | D – data availability | | | X |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **17** | TM – 3.0 | Map no longer pans to congestion (detected or predicted) location | S – operators found that to be distractive |  | X |  |
| **18** | TM – 3.0 | Ramps-Quickview window added | S – operators highlighted the importance of showing ramp queue lengths | X |  |  |
| **19** | TM – 3.0 | the means of displaying ramp rates – changed from purely textual to textual and coloured bars | D |  |  | X |
| **20** | TM – 3.0 | Sensor Data window removed | S - hard to read and a continuous view of the historical data were deemed unnecessary | X |  |  |
| **21** | TM – 3.0 | Activity log window redesigned – name changed to Event List and made tabular | T – display more data for each automatically detected event | X |  | X |
| **22** | TM – 3.0 | ‘trimming’ of ramp metering rate bounds moved from the main UI to a pop-up dialog and Control Panel window removed | S, T – operators would not be expected to constantly correct and contribute to the computer’s actions |  | X | X |

| | | | | | | |
|---|---|---|---|---|---|---|
| **23** | TM – 4.0 | integration of multiple information sources into one view – map, the Ramps and Ramps-Quickview windows have been replaced by the circular display | D – data availability | | | X |
| **24** | TM – 4.0 | Map changed to a schematic road representation, split into segments at the locations of inbound and outbound ramps | D | | | X |
| **25** | TM – 4.0 | Ramps are represented by nodes (circles). Each node is linked by a thin arrow to a set of bars on the outer circle, the direction of the arrow indicating whether the node represents an inbound or outbound ramp | D | | | X |
| **26** | TM – 4.0 | Bespoke CCTV window reintroduced | S - operators make extensive use of the CCTV panels for most of their tasks | X | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **27** | TM – 4.0 | Indication of traffic speed at node location added | S – operators pointed out that an indication of average traffic speed would complement the overview of the road status<br>T – speed data were available in the data bus | X | | |
| **28** | TM – 4.0 | Congestion shown as an increase in size of the node and change in colour to red, instead of a circle | D – circle as signifying congestion, no longer salient feature in new map display | | | X |
| **29** | TM – 4.0 | Road occupancy shown by the colour of the road segment, red signifying high density, yellow – medium, while grey showing normal to low levels of density | D – increased diagnosticity of congestion | | | X |
| **30** | TM – 4.0 | Ramp rates and occupancy bars are linked to a physical location on the map | D – allow for global patterns to be spotted | | | X |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **31** | TM – 4.0 | Text representing actual values of ramp rates, occupancies and speed were removed | S – operators have pointed out that they rarely need to know precise values and they are more interested in ramp states | X |  | X |
| **32** | TM – 5.0 | Circular design changed to radial | S – operators found the circular placement of ramps hard to read |  |  | X |
| **33** | TM – 5.0 | Map added in background of the schematic road | S, D - replacing the initial map with a schematic of the road results in some loss of spatial context | X |  | X |
| **34** | TM – 5.0 | Display of Congestion event changed - Increasing segment thickness in addition to colouring it red | D – increase in diagnosticity of congestion |  |  | X |
| **35** | TM – 5.0 | Live Feed window changed by adding cycling views from other parts of the road network | D – increase in Situation Awareness | X |  |  |