



Scalable Data Analytics,
Scalable Algorithms, Software Frameworks
and Visualization ICT-2013 4.2.a

Project FP6-619435/SPEEDD

Deliverable D5.2

Distribution Public



<http://speedd-project.eu>

D5.2 – Second version of visual analytics suite for proactive decision support

Chris Baber, Sandra Starke, Xiuli Chen, Natan Morar, Andrew
Howes, and Neil Cooke

(University of Birmingham)

Status: FINAL

December 2015

Project

Project Ref. no	FP7-619435
Project acronym	SPEEDD
Project full title	Scalable Proactive Event-Driven Decision Making
Project site	http://speedd-project.eu/
Project start	February 2014
Project duration	3 years
EC Project Officer	Alina Lupu

Deliverable

Deliverable type	Report
Distribution level	Public
Deliverable Number	D5.2
Deliverable Title	Second version of visual analytics suite for proactive decision support
Contractual date of delivery	M22 (December 2015)
Actual date of delivery	December 2015
Relevant Task(s)	WP5/Tasks 5.1, 5.2, 5.3
Partner Responsible	UoB
Other contributors	
Number of pages	72
Author(s)	C. Baber, S. Starke, X. Chen, A. Howes, N. Morar, N. Cooke
Internal Reviewers	
Status & version	Final
Keywords	Cognitive Work Analysis, Visual Analytics, User Interface Design, Decision Modelling, Eye Tracking

Executive Summary

This report details activity undertaken during year 2 of the SPEEDD project for Work Package 5. The primary aim of this work package is to develop and evaluate novel User Interface (UI) designs to support human decision making with big data and proactive, event-driven decision systems.

In year 1, our attention laid in understanding the operator decision making in the use case domains of the SPEEDD project, building and testing a model of decision making and developing experimental designs and measures to evaluate information search and decision making. This was reported in D5.1.

In year 2, the information requirements of operators in the different domains were used to specify information representation. In this report, we describe the ongoing development of a methodology to support an auditable approach to making visual design decisions. Once a set of design options can be identified, the next challenge is to evaluate these against the demands of the decision tasks that they are intended to support. To this end, the decision model provides a means to conduct such evaluation. The benefit of the model is that reveals the underlying strategy that one would expect a rational decision maker to apply when using a given UI design to support a given type of decision. This means that the output is not simply a matter of evaluating a design solely against aesthetic or usability criteria, but requires a deeper understanding of how the UI design interacts with decision strategy.

The output of the model is compared with human participants performing the same credit card fraud analysis task. There is good agreement between the results of the model and the human performance. An implication of this finding is that, with much practice, people are able to recognize the validity of specific information sources for the decisions they are making. Following the experiment, participants were asked to explain their strategy and many of them suggested that they developed ‘stories’ to explain relations between the data. This raises the intriguing possibility that the ‘stories’ are, in one sense, the heuristic method that people apply to assigning validity to information cues and, in another, the emerging awareness of the cue validity values that they are using. This is a novel proposal and one that we seek to explore further in future work.

When people work with automated decision support, there is an ongoing question of how they respond to automation failure. We report two experiments in which participants are required to identify (in a road traffic monitoring task) whether the automation suggestion is correct or not. In one experiment, we demonstrate that expert performers are able to attend appropriately to all relevant cues, while non-experts consistently miss of the cues. Thus, cue validity arises from the interaction between representation of the information and the user’s knowledge of the domain. Interestingly, a subsequent experiment showed that simply drawing participants’ attention to the need to consider cues relevant to two decision tasks greatly reduced this effect. This suggests that the design of the visualization should not be solely concerned with presenting data but also with prompting which action and decision the user needs to make.

Contents

Executive Summary	3
1. Introduction.....	8
History of the document.....	8
1.1 Purpose and Scope of the Document	8
1.2 Structure.....	9
2 Approach to User Interface Design	11
2.1 Mapping Information Content to Cognitive Work Analysis.....	11
2.2 Load-Balance Diagrams in Ecological Interface Designs.....	12
2.3 How do Ecological Interface Designs affect operator performance?	14
2.4 Defining a Space of Representations	18
2.5 Conclusions.....	23
3 Modelling Decision Making Using Different User Interface Designs.....	24
3.1 Introduction.....	24
3.2 Aims of Modelling	26
3.3 Defining the Context of Credit Card Fraud Analysis.....	26
3.4 Using Visualization.....	27
3.4.1 Visual Perception	27
3.5 Modelling Theory	28
3.5.1 Problem formulation	28
3.5.2 Observation Model.....	31
3.5.3 Peripheral vision	31
3.5.4 Time Cost.....	32
3.5.5 Learning	32
3.6 Results.....	33
3.6.1 Information used and accuracy	33
3.6.3 Decision time distribution.....	34
3.7 Discussion	35
4 Experiments on Decision Making using Different User Interface Designs: Credit Card Fraud Use Case	36
4.1 Introduction.....	36
4.2 Describing Human Decision Making.....	36
4.3 Experiment 5.2.1	37

4.3.1 Experimental Design.....	37
4.3.1.1 Scenario	38
4.3.2 Method	39
4.3.4 Results.....	42
4.4 Discussion.....	47
5 Experiments on Decision Making using Different User Interface Designs: Traffic Management .	50
5.1 Introduction.....	50
5.2 Study 5.2.2	51
5.2.1 Experimental Design.....	51
5.2.2 Method	51
5.2.3 Results.....	53
5.2.4 Conclusions to Experiment 5.2	55
5.3 Study 5.2.3	56
5.3.1 Experimental Design.....	56
5.3.3 Method	57
5.3.4 Results.....	57
5.3.5 Conclusions.....	60
5.4 Discussion.....	61
6 Defining Performance Baselines.....	62
6.1 Introduction.....	62
6.2 Can time be a useful measure of baseline performance?	62
6.2.1 A Critical Path Model of Experiment 5.2.3	62
6.2.2 What's wrong with Critical Path Models?.....	64
7 Discussion.....	66
7.1 Introduction.....	66
8. References.....	68

List of Tables

Table 1: Summary of results for Jungk et al. (1999) study (* $p < 0.01$; ** $p < 0.05$).....	14
Table 2: Mapping Bertin's (1983) reading and encoding to the Information Content in figure 2. 21	
Table 3: Categorising alternative UI designs (for Fraud management).....	25
Table 4: Nine information sources.....	27
Table 5: Nine information sources used in the experiment.....	40
Table 6: Comparison of Average Times across different forms of Match under all conditions....	60
Table 7: Comparison of Predicted and Observed times for Experiment 5.3.....	64

List of Figures

Figure 1: Strategy for WP 5.....	8
Figure 2: Deriving Information Content from Cognitive Work Analysis views for the Traffic Management UI	11
Figure 3: Load Balance Diagrams in Ecological Interface Designs	12
Figure 4: Example of a Fundamental Diagram for describing traffic behavior (de Wit et al., 2012).13	
Figure 5: Representing traffic flow and vehicle density on initial prototypes of the SPEEDD UI for Traffic Management	13
Figure 6: UI designs evaluated by Jungk et al. (1999).....	15
Figure 7: Representing the same problem in different ways.....	16
Figure 8: Space of Representations for Tower of Hanoi puzzle (Zhang and Norman, 1994).....	17
Figure 9: Effect of changes in dimension of problem solving performance	18
Figure 10: Visual Design Process [Upton and Doherty, 2008].....	19
Figure 11: Bertin's variables for an image.....	20
Figure 12: Initial UI Concepts for WP8.....	21
Figure 13: Concept UI for WP8.....	22
Figure 14: Heat map of suspicious trading.	24
Figure 15: Four interface variants for credit fraud detection.	25
Figure 16: The color acuity function used in the optimal control model.	32
Figure 17: Information Sources used by the model	33
Figure 18: Accuracy achieved by the model.....	34
Figure 19: Model decision time	34
Figure 20: User interfaces for the four experimental conditions	38
Figure 21: General layout of the abstracted user interface in the four experimental conditions. ..	39
Figure 22: Boxplots summarising six outcome measures for each condition.....	42
Figure 23: Viewing networks sorted by the experimental condition.	44
Figure 24: Dwell times for the nine information sources for all experimental conditions.	45
Figure 25: Frequency of cue use ordered by cue validity.	49
Figure 26: User Interface for the Traffic Management Experiment	52
Figure 27: Boxplots for decision times for the different response categories.....	53
Figure 28: Mean correct responses in terms of scenario for each group.....	54
Figure 29: Boxplots for decision times different response categories	55
Figure 30: Percentage view time per region of interest (ROI).....	55
Figure 31: User Interface for Decision Only condition	57
Figure 32: User Interface for Complete Form and Make Decision condition	57
Figure 33: Mean Decision Time across all conditions.....	58
Figure 34: Mean Correct Response for conditions	59
Figure 35: Comparison of Solution Source.....	59
Figure 36: Extract from CPM for Experiment 5.2.3	63

1. Introduction

History of the document

Version	Date	Author	Change Description
0.1	02/11/2015	Chris Baber (UoB)	Set up the document and initial content
0.2	05/11/2015	Sandra Starke (UoB)	Chapter 4 added
0.3	06/11/2015	Xiuli Chen (UoB)	Chapter 3 added
0.4	07/11/2015	Chris Baber (UoB)	Chapters 1, 2 and 6 added
0.5	08/11/2015	Natan Morar (UoB)	Chapter 5 added
0.6	13/11/2015	Chris Baber (UoB)	Draft Report submitted for internal review
1.0	07/12/2015	Chris Baber (UoB)	Final report submitted following corrections resulting from internal review

1.1 Purpose and Scope of the Document

This is the second report on activity under Work Package 5 (Real-time Visual Analytics for Proactive Decision Support). As stated in the Description of Work, “The primary objective of this work package is to explore the impact of real-time proactive decision computation on human decision-making in Big Data applications.” This requires two strands of research: the first focuses on Ergonomics / Human Factors, particularly in terms of understanding the nature of decision making in the SPEEDD use case domains; the second focuses on the design and evaluation of Visual Analytics to support such decision making. In order to meet the primary objective, the Work Package involves three tasks:

- T5.1: Modelling Decision-Making as a Socio-Technical Activity
- T5.2: Defining Objective Metrics for Evaluating Decision-Making
- T5.3: Real-Time Visualization for Human Decision-Making

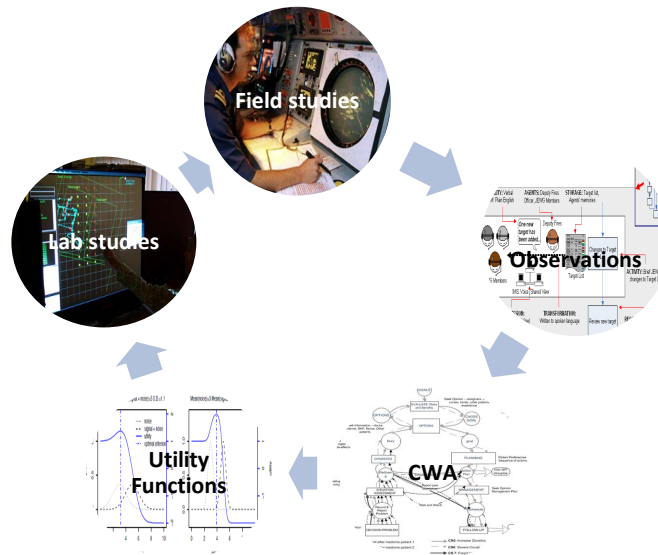


Figure 1: Strategy for WP 5

As figure 1 highlights, in Work Package 5, we take an Ergonomics / Human Factors approach in which work activity of practitioners is observed and described (using Cognitive Work Analysis) which leads to a description of decision strategy (in the form of utility functions) and requirements for user interface designs, which are evaluated through laboratory studies and review by Subject Matter Experts.

T5.1 involved CWA of operator activity in the use case domains. This was completed in year 1 (although the analysis will be updated with each new set of observations in the field). A challenge for SPEEDD is to analysis of these observational data (and data from laboratory experiments) to define baselines against which SPEEDD prototypes can be evaluated, in terms of changes to operator decision making.

T5.2 involved the definition and initial testing of a novel model of human decision making. This was completed in year 1 and reported at the CHI conference (Chen et al., 2015). The challenge for year 2 was to demonstrate the applicability of the model to SPEEDD use cases.

T5.3 involves the design of User Interfaces for SPEEDD. In year 1, the aim was to map operator activity to information requirements and develop concept displays that presented this information. In year 2, the aim is to map information requirements to representation in order to develop a design space in which to develop novel visualisations.

1.2 Structure

The report begins with a discussion of how SPEEDD WP5 defines user interface design. In Chapter 2: Approach to User Interface Design, the central question is how do we go from the ‘content’ identified by the Cognitive Work analysis (CWA) to the ‘presentation’ on the UI? We have demonstrated (in D5.1) that CWA can identify the information requirements of operators for the tasks and goals under consideration. This provides an appreciation of the information content that should appear on displays. However, there is little guidance in the Human Factors or Human-Computer Interaction literature as to what visual form is most appropriate for presenting this content. We are developing a design process which seeks to bridge this gap from defined content to visual design.

While it might be possible to generate visual design from the approach outlined in chapter 2, we recognise that there will be a set of alternative options for most visual designs. Consequently, SPEEDD is seeking to develop a means by which the resulting design space can be parameterised and modelled. Chapter 3: Approaches to Modelling Use of User Interface Designs is primarily concerned with the question of how do we predict how users might respond to competing visual designs? In order to address this question, we further develop the optimal decision model (presented in D5.1) to compare User Interfaces with different visual designs and with different interaction requirements.

The results of the model developed in chapter 3 suggest that strategies might vary in response to UI design. We wished to explore how people respond to these designs, not only in terms of whether their decision making could exhibit similar characteristics to those assumed in the model but also in terms of their visual search behaviour (i.e., how they looked for information in the different UIs).

In addition to considering the strategies that are adopted in search for information, SPEEDD is also interested in how people make decisions with automated support. Chapter 5: Experiments on User

Performance:. Traffic Management presents two experiments in which participants make decisions on the ramp-metering task that forms the basis of SPEEDD prototype one. We are particularly interested in whether participants can notice and respond to automation failure and how such failure might affect their performance. We also consider how the need to produce a report (of road management activity) could have an impact on operator performance.

The question of how automation impacts on performance is considered in Chapter 6: Defining Performance Baselines. This issue was considered in D5.1 and is further explored in this chapter. We demonstrate that temporal models based on task analysis (using Critical Path Analysis) can provide some reasonable predictions, but argue that such modelling focuses on superficial aspects of performance and neglects the strategies that decision makers apply.

The report concludes with Chapter 6: Discussion, which reviews the main findings from the experiments and outlines the planned work for the next 12 months.

2 Approach to User Interface Design

2.1 Mapping Information Content to Cognitive Work Analysis

In D5.1 (The Design of the User Interface for the SPEEDD Prototype) the process followed for User Interface design in SPEEDD was outlined. This process begins with an appreciation of the activity performed by Subject Matter Experts in the domain of interest and results in a definition of information requirements for this activity. Detailed descriptions of the activity have been provided in D7.1 (User Requirements and Scenario Definitions [Fraud] - update), D7.2 (Initial Evaluation Report), D8.1 (User Requirements and Scenario Definition [Traffic Management]) D8.3 (Evaluation of SPEEDD prototype 1 for Road Traffic Management). The design process outlined in D5.1 is sufficient to define the User Interface in terms of what might be displayed to the user, but does not consider the manner in which the information should be represented. The intention was to work within the tradition of Cognitive Work Analysis, which moves from observation of Subject Matter Experts performing their work *in situ* to specifying information needs and then designing 'ecological User Interfaces'. This process was described in detail in D5.1 and outlined in figure 2 for the traffic management use case.

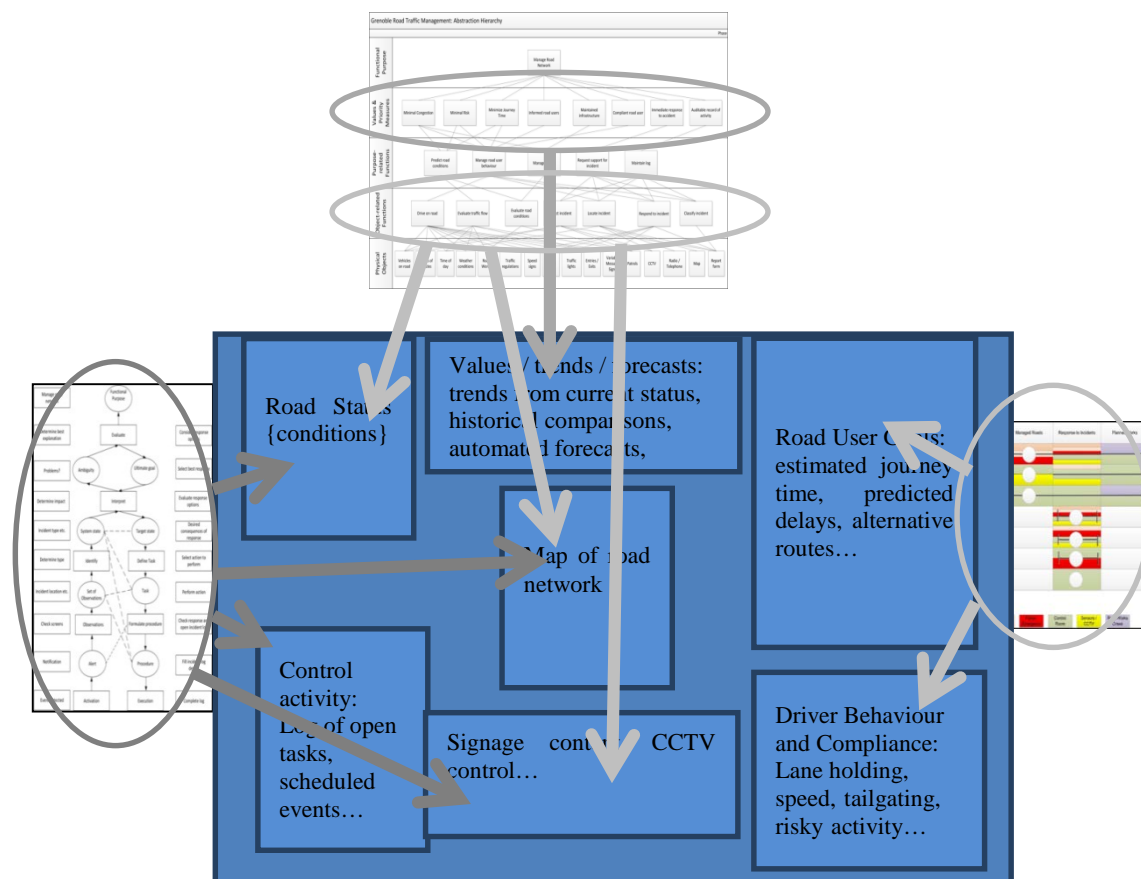


Figure 2: Deriving Information Content from Cognitive Work Analysis views for the Traffic Management UI

CWA (like other Human Factors or Human-Computer Interaction methods) is able to make recommendations for the information content of a UI but does not translate these into a form that a graphic designer is able to use (Bennett and Flach, 2011; Bisantz et al., 2003; Lintern, 2005, 2012). In other words, the approach identifies what information *should* be displayed to the user but does not specify *how* this should be displayed. Of course, the manner in which information could be displayed should be left to the discretion of the designer (as far as possible) and this means that there will be many options which could be used for presentation. These options constitute the ‘design space’ from which it ought to be possible to select some presentations as being more appropriate than others. This requires a clear definition of what the term ‘appropriate’ means, e.g., in terms of the operator, the decision and the domain. We would like the design process to define the space of possible representations that could be used for the User Interface (UI), and to have some means of defining appropriateness beyond solely asking for user opinion and preference. However, we do not wish this process to be overly prescriptive as this would not only remove the opportunity for the UI designer to produce novel designs but could lead to all UI designs becoming homogenous.

2.2 Load-Balance Diagrams in Ecological Interface Designs

The key challenge that the UI designs in SPEEDD seeks to address is how to represent the parameters that an operator needs to know in order to ensure stability in the system, e.g., in terms of maintaining traffic flow or in terms of ensuring acceptable use of credit cards. In D5.1 it was noted that CWA considers the issue of balance across parameters in terms of the Values and Parameters (in the Abstraction Hierarchy) that constrain the operator decision space. In Ecological Interface Design (EID), a common approach to UI design is to represent competing demands in the form of a polygon display. However, this leads to the complaint EID displays often take a similar form to Process Control displays used in power stations or factories. For example, figure 3, show design prototypes for a system to haemodialysis decision in an operating theatre (on the left), for a cement mill simulation (in the centre), and for a military planning tool (on the right). While these are very different domains (and one would assume that this would mean that the information requirements would be very different), the use of a polygon to display the relationship between two parameters in the system means that they all look remarkably similar. Similarity could be interpreted as a mark of consistency across domains but we feel it reflects a lack of imagination on the part of these designers.

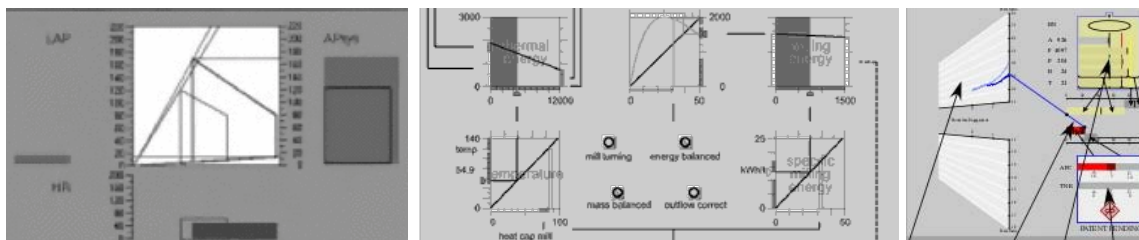


Figure 3: Load Balance Diagrams in Ecological Interface Designs

One reason why the load-balance design is so popular in the ecological interface design tradition is that it reflects the assumption in Cognitive Work Analysis (from which such designs are derived) that the primary goal of the operator is to monitor interacting parameters that describe system behaviour. For traffic management use-case in SPEEDD, such parameters could be presented in the Fundamental Diagram that relates vehicle density to traffic flow (figure 4).

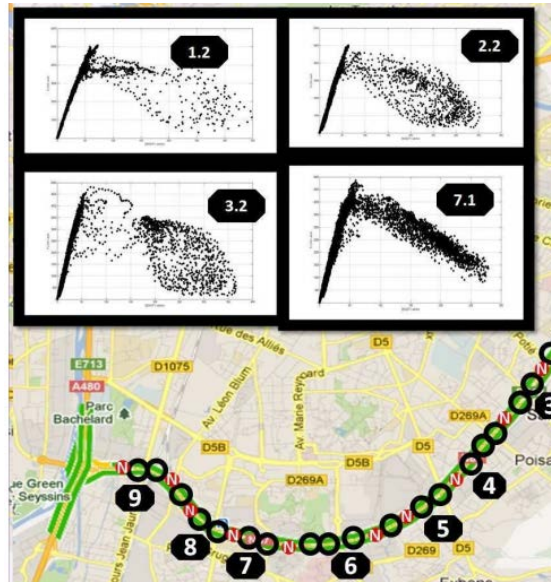


Figure 4: Example of a Fundamental Diagram for describing traffic behavior (Canudas de Wit et al., 2012). Traffic density is mapped against vehicle density for specific locations over time.

In the initial prototype UI for Traffic Management, we developed a display that presented rate x density in a manner analogous to the Fundamental Diagram (figure 5).

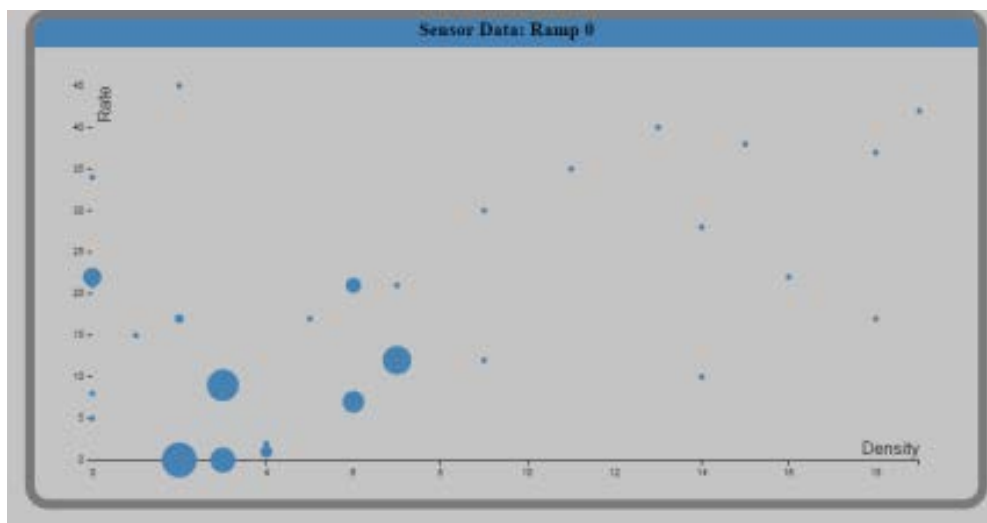


Figure 5: Representing traffic flow and vehicle density on initial prototypes of the SPEEDD UI for Traffic Management

Presenting information that supports tasks associated with balancing two parameters is assumed to be a key aspect of the metacognitive activity of operators. In other words, in addition to monitoring the status of a system in terms of specific parameters, it is vital that the operators are able to consider how these parameters interact in order to understand the underlying dynamics of the system. However, it is not obvious why the polygon, load-balance used in EID would be the *only* means of presenting such information.

2.3 How do Ecological Interface Designs affect operator performance?

Jungk et al. (1999) compared three forms of UI for an anaesthetist application. The designs are shown in figure 6. Each display was used to complete 38 different tasks in which the participant had to decide if the reading was acceptable or not. Table 1 shows a summary of the findings from this study.

Table 1: Summary of results for Jungk et al. (1999) study (* $p < 0.01$; ** $p < 0.05$)

Measure	User Interface		
	Trend-line	Profilogram	EID
Trial Time (s)*	133 (± 80)	221 (± 125)	237 (± 127)
Rate of misdiagnosis errors**	37%	19%	13%
Visual Fixations*	205 (± 120)	368 (± 212)	268 (± 180)
Manual action*	15 (± 10)	26 (± 15)	29 (± 22)

It is apparent that the UI designs in this study resulted in measurable differences in performance. In general the trend-line (with which the participants were likely to have been most familiar) resulted in fewer actions and visual fixations than the other displays, which translates into faster time but results into many more errors. In comparison the EID and profilogram took more time but resulted in fewer errors. The authors suggested that the EID could be seen as preferable (with fewer errors and fewer visual fixations than the profilogram), but our aim in presenting this study is to highlight two points. The first is that, as one might expect, performance varies across display designs. The implication is that some designs can be demonstrably superior to others, depending on the task, users and context. The second is that the range of measures reveals different aspects of performance. For the Jungk et al. (1999) study, one can see a speed-accuracy ‘trade-off’ (in which faster performance times seemed to result in more error); although it is not obvious that this reflects a deliberate ‘strategy’ on the part of the users, i.e., given the nature of the study it was unlikely that participants would be seeking to get more diagnoses wrong, so much as a consequence of the strategy that arose from the use of particular UI designs. In other words, the ‘cost’ of searching for information could interact with either the ‘cost’ of processing that information or the ‘value’ of that information in making a decision. This suggests that evaluation of UI designs should consider not only basic performance aspects, such as time and error, but also the different strategies that operators employ when using these designs. For the SPEEDD project, this strategy is reflected by the modelling approach we employ (see chapter 3) and by the experiments that have been conducted relating to user performance (see chapters 4 and 5).



Figure 6: UI designs evaluated by Jungk et al. (1999)

While Ecological Interface Designs are intended to support direct perception of information, the underlying theory as to why this should occur has been less well developed. One version of Distributed Cognition views objects-in-the-world as the means to represent information, which, in turn, cue or encourage specific cognitive activity (Zhang and Norman, 1994, 1995). Thus, a problem can be represented in several ways. The representation could call to mind the procedures that are required or could prompt specific actions. For example, Zhang and Norman (1994) presented the same ‘Tower of Hanoi’¹ problem using different representations (figure 7). Results from a set of experiments indicated that manner in which the problem is solved was affected by the representation.

¹ The Tower of Hanoi (or Tower of London) problem requires the movement of discs from one peg to another. The aim is to move the complete stack, by moving one disc at a time subject to the rule that larger discs are not meant to sit on smaller discs (or vice versa, depending on the version of the task).

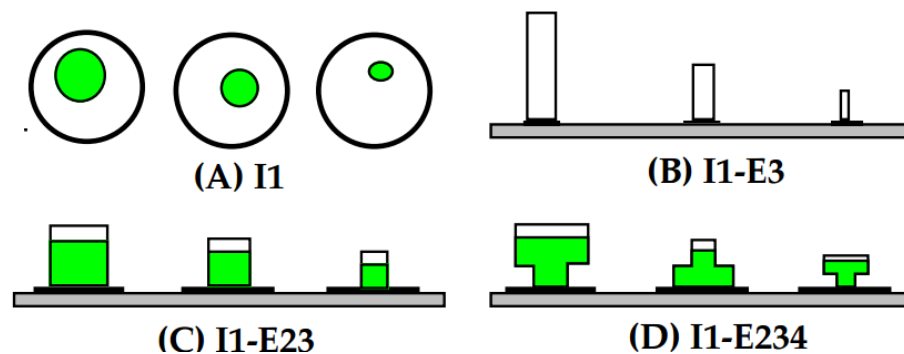


Figure 7: Representing the same problem in different ways

Zhang (1997) accounts for these differences through a process of alignment of the internal representations are held by the person in the form of schema with external representations through processes of perception. Thus, an appropriate external representation can support perception of salient features and cue specific actions. He called such representations Relational Information Displays (RIDs) and suggested that these ‘represent the relations between information dimensions’ (Zhang, 1997, p.59). This suggests that specific representations ought to call to mind expected relations in a set of data. For Zhang and Norman (1994), the space of representations for the Tower of Hanoi puzzle could be defined as shown in figure 8. Presenting the puzzle using manipulations of the dimensions resulted in differences in performance, as shown in figure 9. The implication is that changes in the representation lead to changes in performance. For our design process, we would like to be able to predict this relationship in such a way as to describe the design space of plausible alternative representations, based on an understanding of the ‘puzzle’ that people need to solve and the actions that they need to perform.

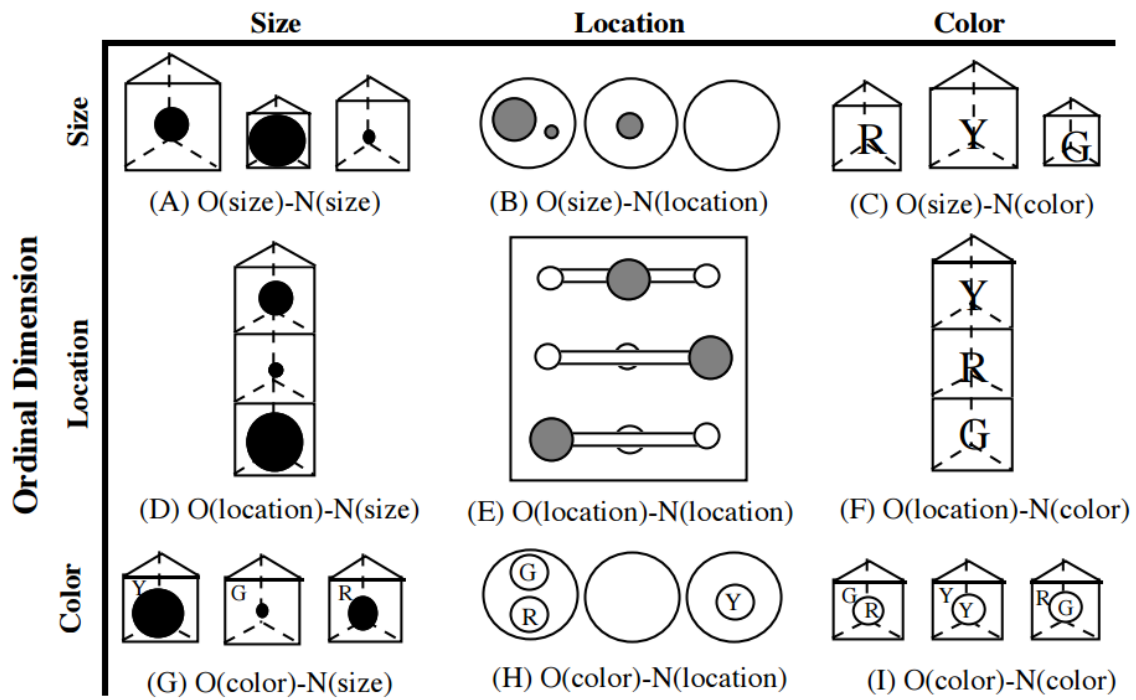


Figure 8: Space of Representations for Tower of Hanoi puzzle (Zhang and Norman, 1994)

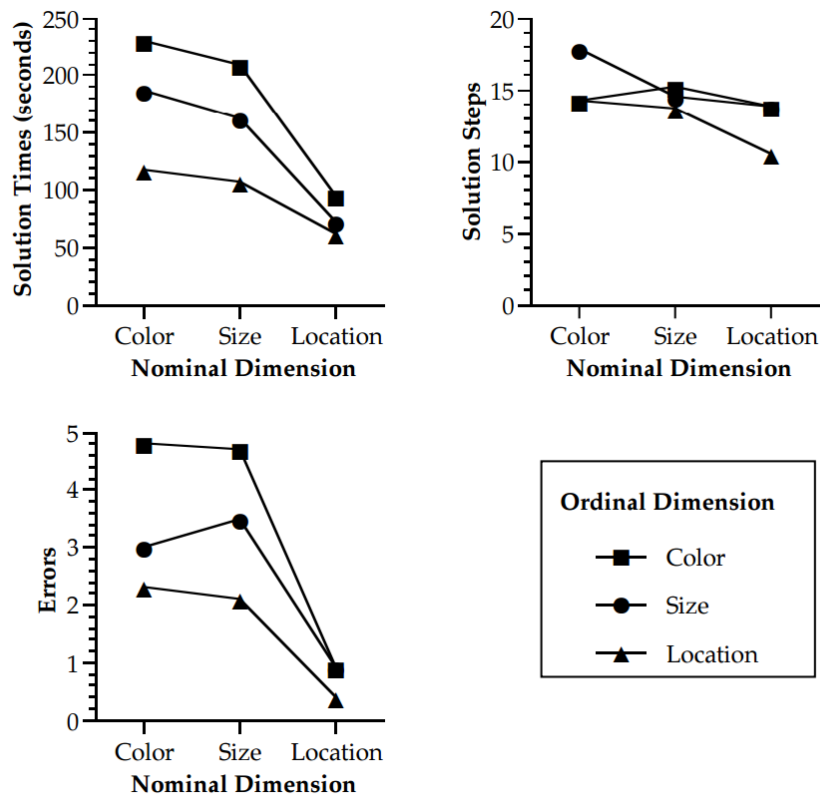


Figure 9: Effect of changes in dimension of problem solving performance

2.4 Defining a Space of Representations

D5.1 demonstrated how CWA can produce a space of information content, and the polygon displays used in EID (figure 3) illustrate one approach to presenting the parameters which operators are seeking to balance. As the previous section notes, however, different UI designs can have marked and predictable effects not only on users performance but also on their understanding of the problem that they are attempting to solve, and on the strategy that they apply to decision making. Our challenge in year 2 is to develop an approach to UI design that reconciles the challenges of EID with the recognition that UI designs influence the strategy that people might apply in decision making.

As a starting point, figure 2 was derived from a process that mapped the Cognitive Work Analysis views to the Information Content that we believed was important to the described activity. We need to map this to a space of possible representations. One approach would be to follow the work on Upton and Doherty (2008) who develop UI designs following the process shown in figure 10.

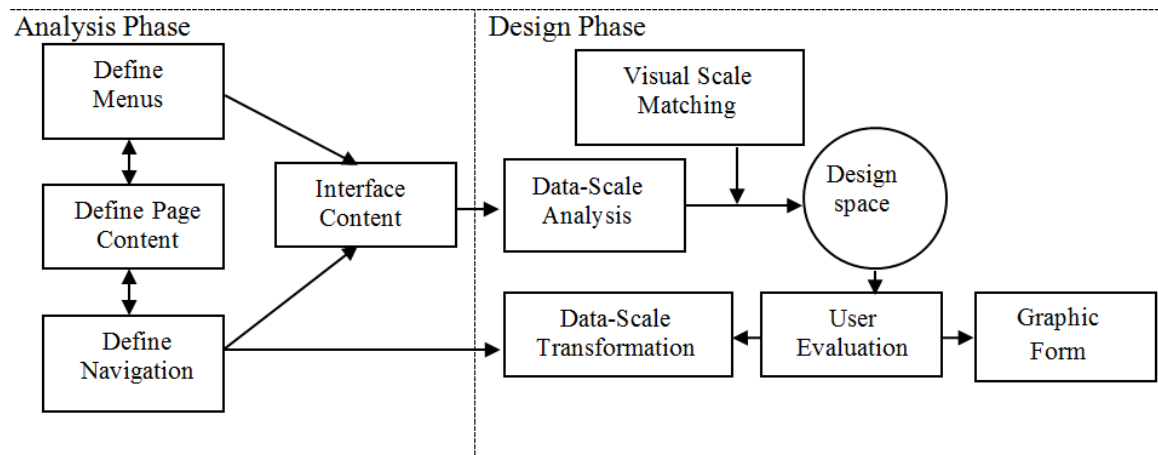


Figure 10: Visual Design Process [Upton and Doherty, 2008]

Key to the process shown in figure 7 are the notions of Data-Scale analysis / Visual Scale matching which were proposed by Bertin (1983). These provide a means of defining the range of forms which could be appropriate for the information that needs to be presented to the operator.

Bertin (1983) begins with a high-level set of activity that a person might do in response to a visual display. He terms these ‘Forms of Perception’ and defines these as:

- Associative – which defines the ease with which changes in several items can be recognized in a group;
- Selective – which defines the ease with which changes in one item can be recognised;
- Quantitative – which defines the ease with which information on the display can be read and values judged.
- Order – which defines the ease with which items can be compared and sequenced.
- Length – which defines the number of intervals a scale might contain, e.g., how many numbers can be displayed between 0 and 10, or how many shades of grey can be distinguished on a display?

Relating these forms of perception to the Skills-Rules-Knowledge framework (Rasmussen, 1983) that we discussed in D5.1, one can see that Associative and Selective Forms of Perception are likely to be skill-based (i.e., automatic) behaviours, while Order and Length are likely to be rule-based behaviours, and Quantitative is a combination of rule-based and knowledge-based behaviours. Consequently, the effort (cost) required to extract information using the Associative and Selective forms of perception can be assumed to be much lower than the Quantitative form.

These notions of *how* the content of an visual display can be ‘read’ are echoed in the Human Factors literature, e.g., Sanders and McCormick (1992) talk of Quantitative reading (for precise data reading); Qualitative reading (for reading trends etc.); Check readings (for deciding if parameters are within limits); Situation awareness (for deciding how the system is performing). There has been much research which compares different forms of representation for these different forms of reading. For example, for car speedometers, knowing the vehicle’s precise speed involves Quantitative reading and performance (and

preference) is superior with the digital, alphanumeric display. However, if the task was to interpret the speed relative to a speed limit, this would be a form of Qualitative reading and curvilinear displays results in superior performance. This indicates that there can be predictable relationships between the nature of the task that the user is performing and the manner in which information is presented.

Bertin (1983) considered presentation to consist of a set of primitive shapes, as shown in figure 8. Each primitive supports a particular type of encoding, e.g., in terms of position, proximity, placement, size, label etc., and each type of encoding relates to the type of reading activity that the user will perform. The implication is that there ought to be a set of encodings which are more suitable than others for each type of reading.

LES VARIABLES DE L'IMAGE									
POINTS			LIGNES			ZONES			
XY 2 DIMENSIONS DU PLAN	x	x	x	/	~	/	15 1 9 14 1 1 16 21 2 2 14 15 1	2 1 18 2 1 21 15 1 1 2 9	OQ ≠
Z TAILLE	■	■	■	/	~	/	■	■	OQ ≠
VALEUR	■	■	■	/	~	/	■	■	O ≠
LES VARIABLES DE SÉPARATION DES IMAGES									
GRAIN	■	■	■	/	~	/	■	■	O ≠
COULEUR	■	■	■	/	~	/	■	■	≠
ORIENTATION	■	■	■	/	~	/	■	■	≠
FORME	■	■	■	/	~	/	■	■	≠

Figure 11: Bertin's variables for an image

The notion that one can categorise forms of perception in response to information content (admittedly in acontextual and abstract terms) and that these can be related to forms of perception is illustrated by table 2. These relations could be used to either critique a design concept or propose a new one.

Table 2: Mapping Bertin's (1983) reading and encoding to the Information Content in figure 2.

Information Content	Form of Perception	Encoding
Road status	Associative	Zones / Couleur
Trends / forecasts	Associative	Ligne
Control activity	Quantitative	Alphanumeric
Controlled objects	Selective	Points (status); Video (content)
Map of road network	Order	Zones
Road user goals	Order	Valeur
Driver behaviour	Associative	Valeur / Couleur

The initial UI prototypes (for the use-case Traffic Management: ramp metering) were developed in D8.1 (User Requirements and Scenario Definition [Traffic Management]) D8.3 (Evaluation of SPEEDD prototype 1 for Road Traffic Management) and shown in figure 12. These present the information content that might be appropriate to the Operator's task in separate windows on the screen. Each display has a map of the road network and some indication of road status (typically as a version of a Fundamental Diagram to show traffic activity and as a grid of coloured squares to show ramp status). The Controlled Objects is presented in the form of a window showing video (usually this is to be called up by the operator by clicking a specific point on the map). In the two displays on the left of figure 12, road user goals and driver behaviour are illustrated using colours and icons. However, the operators (during the initial evaluation reported in D8.3) felt that this information was not part of their primary responsibilities. Control activity was represented in form of short reports (typically in the bottom right of the displays).

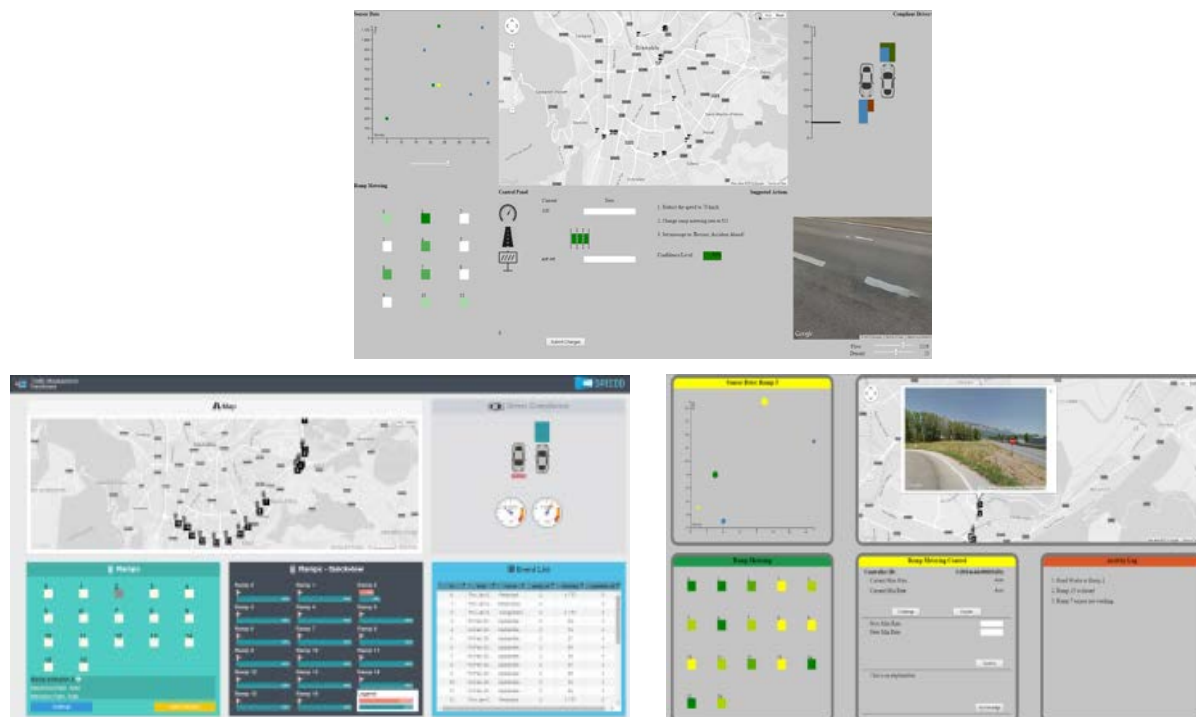


Figure 12: Examples of Initial UI Concepts for WP8

What the UI concepts in figure 12 do not have is a direct relationship between the ramp status displays and the map. This requires the operator to translate from the map to the ramp status (i.e., by reading the number of the ramp on the map and then looking for this in the grid). Furthermore, comparison between the status of ramps is performed in the grid rather than on the map. This could mean that identifying patterns and trends could be problematic. Reducing the need for such translation and increasing the opportunity to identify trends in the data is a central requirement for an ecological UI (because this would support the aim of enhancing ‘direct perception’ of system status to the operator). Redesigning the UI led to figure 13. In figure 13, the map of the road network is central to the display. The road is divided into zones. This is intended to support the Order form of perception of the displayed information, i.e., to enable comparison of multiple zones. Colour-coding of the status of the zones is intended to support both associative and order forms of perception. Data that relate to each ramp are shown in bar-charts surrounding the display. These are intended to support associative form of perception. The use of three concentric circles is intended to support the Order form of perception; both in terms of the comparison of the current readings (inner circle) with ‘historical average’ (middle circle) and predicted (outer circle); this design is also intended to enhance Situation Awareness (see D5.1). Linking of bar-chart to specific ramp is currently performed using lines linking map to charts, but this is being reviewed and evaluated.

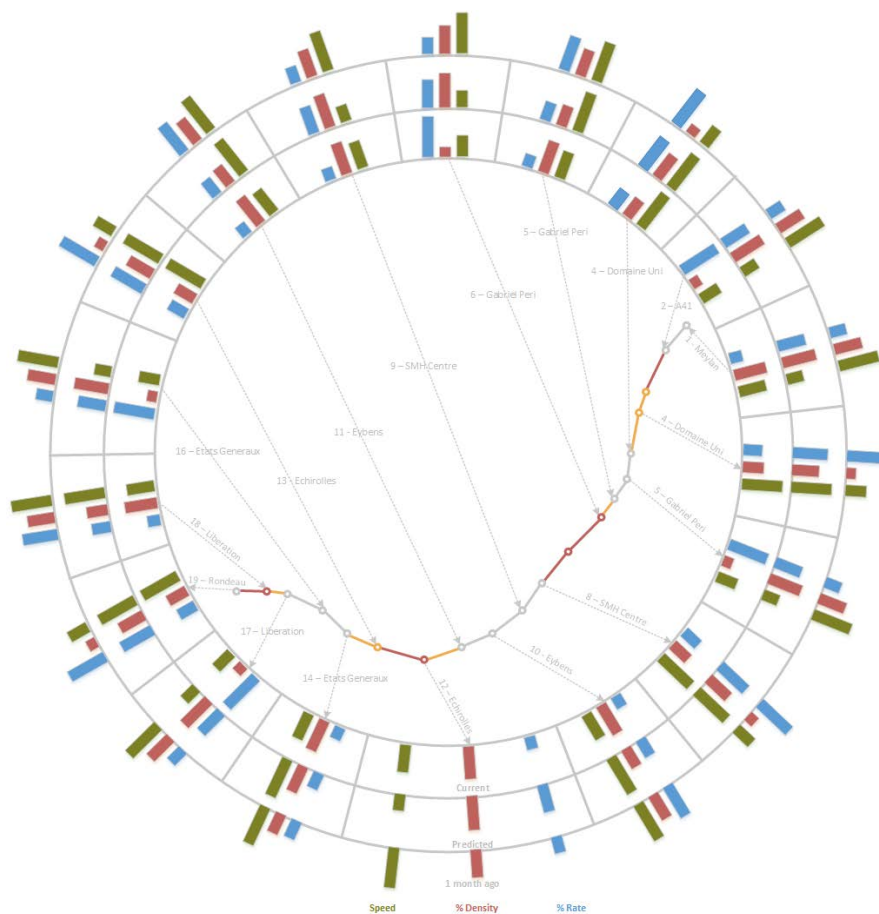


Figure 13: Concept UI for WP8

2.5 Conclusions

In this chapter, our aim has been to develop an approach to UI design that allows mapping of work activity to information content (using CWA) and from information content to presentation (using Bertin's (1983) notion of forms of perception). This provides a means through which design decisions can be considered and questioned. Thus, we could ask what form of perception a particular form of information presentation might support and how this relates to the decision the operator needs to make. This views the evaluation of a given design (at least during the formative assessment as the design develops) as a way of questioning the assumptions that underpin the design. While we have found this a useful means of thinking about design, and while we believe that the categories and types that are employed support a shift from an entirely subjective process to designing UI, we recognize that there remains much ambiguity in the approach. Not least of which relates to the consideration of which form of perception are most appropriate for which type of task, and which form of presentation is most appropriate for which form of perception. Having said that, we believe that the development of this approach does provide a means of auditing UI design decisions and can be valuable for developing concept UI designs.

In addition to outlining the design process that is being followed in the development of UI for SPEEDD, this chapter also highlights the need to better appreciate the strategy that operators will adopt in response to UI designs (where 'strategy' means the selection and use of information provided to support decision making). This means that, in addition to considering preference for UI designs or time / error (which are common approaches to UI evaluation) we believe that it is critical to know how the operator is using the UI and this implies that there will be optimal ways of using the UI for specific types of decision. This is one of the key assumptions of the modelling work we are undertaking in SPEEDD, and this is discussed in chapter 3.

3 Modelling Decision Making Using Different User Interface Designs

3.1 Introduction

In this chapter, we explore the challenge of evaluating options in a design space through the use of modelling decision activity. We begin with the assumption that the dashboards that fraud analysts employ (as we reviewed in D7.2) tend to follow two main types: either containing a number of alphanumeric fields which show account and transaction details or a color-block (heat map) coding of transactions. Several companies, for example SL Corporation, StreamBase, Apama, Gartner Group have developed software to provide real-time monitoring and visualization services to support people making real-time decisions with large amounts of visualized data. Apama, for example, uses heat maps to visualise suspicious trading activity in real time (Figure 14) in the domain of stock market trading.

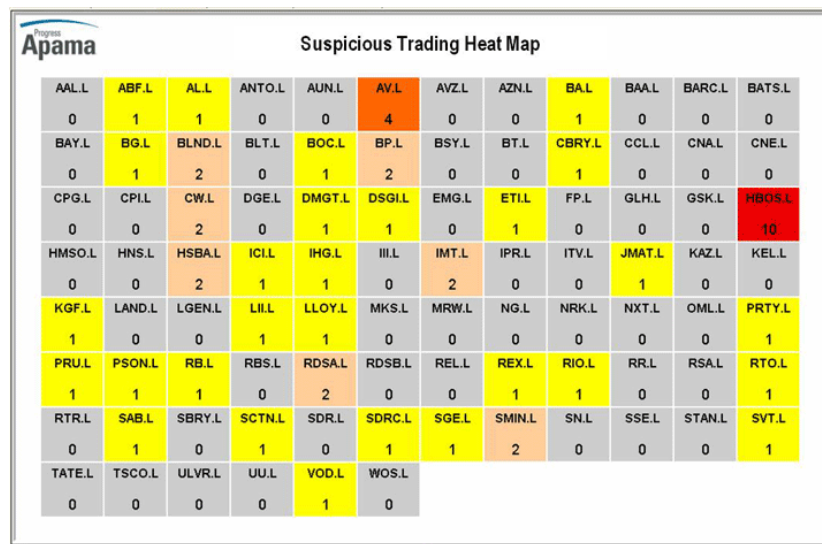


Figure 14: Heat map of suspicious trading²Apama uses a heat map to visualise suspicious trading activity in real time. The heat map is based on the event processing visualization platform by SL Corporation.

In terms of Bertin's (1983) forms of perception, heat maps can be assumed to support selective, associative and order forms of perception of information, but might be less useful for quantitative perception. Conversely, an alphanumeric display might be useful for quantitative perception but less so for the other forms. Table 3 compares a heat map with an alphanumeric display. Obviously, there are many other forms of presentation that could be considered, particularly as neither of these forms support the identification of trends in data. However, we felt that a comparison of these two extremes would

² <http://www.thecepblog.com/2008/01/02/apama-fraud-detection-and-heat-maps/>

illustrate how a design space could be described and then explored through modelling.

Table 3: Categorising alternative UI designs (for Fraud management)

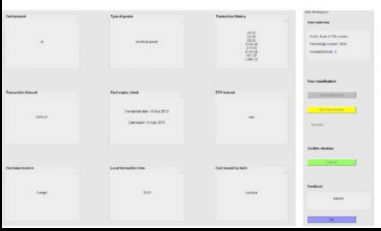
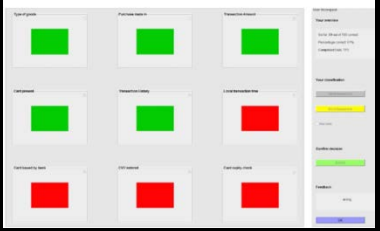
		
Associative	Difficult: no indication of target or threshold values	Easy: Colours used to define target against threshold values
Selective	Difficult: no indication of target or threshold values	Easy: targets pop out (providing colour contrast is sufficient and number of targets is not too large)
Order	Difficult: no indication of threshold values or proportion of targets	Easy: different colours indicate proportion of targets against threshold values
Quantitative	Easy: can read and interpret values for each entry	Difficult: cannot easily define the threshold values, or what boundaries between values might mean



Figure 15: Four interface variants for credit fraud detection.

The text is not intended to be readable. What is important is that information cues are represented with text (left panels) or colour (right panels) and that the information is either immediately available (bottom panels) or revealed by clicking (top panels).

In addition to compare the representation we also wished to model access costs. In other words, selecting information from a screen in which all information is permanently available is likely to have a lower access cost than one in which the user needs to click on a box to reveal its contents. These conditions are illustrated by figure 15 and explained in more detail in chapter 5.

3.2 Aims of Modelling

Our approach is based on the assumption that user strategies are adaptive to a range of constraints (Lewis et al., 2014; Payne and Howes, 2013). In this approach a sequence of user actions is predicted from an analysis of what it is rational for a user to do given an interface design and given known constraints on human cognition and given a particular type of decision to make. These analyses take into account the costs and benefits of each action to the user so as to compose actions into efficient behavioural sequences (Tseng and Howes, 2015). In addition, there are models of multi-tasking in which the time spent on each of two or more tasks is determined by their relative benefits and time costs (Zhang and Hornof, 2014). More recently, it has been suggested that menu search can be understood as a Partially Observable Markov Decision Problem (POMDP) (Chen et al., 2015). Two important properties of this kind of problem are (1) the visualised data allows only partial observations through foveated vision, which covers only 1-2 degrees of visual angle, and (2) the outcomes are partly random (the outcomes of a decision are not always known) and are partly determined by the users' information selection (which part of the visualisation the user chooses to focus on).

3.3 Defining the Context of Credit Card Fraud Analysis

D7.1 and D7.2 provide detailed discussion of the nature of credit card fraud analysis. In broad terms, machine learning is employed on a large scale across millions of transactions in order to extract fraudulent patterns from transaction attributes that can easily enumerate to 100 or more. In terms of triaging and screening, we assume that an automated detection process is running and that this process has flagged a given transaction (or set of transactions) as suspicious and a user will engage in some form of investigation to decide how to respond to the flag. Based on interviews and discussions with credit card fraud analysts and organisations, we believe that there are several ways in which the investigation could be performed. In some instances, the investigation could involve direct contact with the card-holder. In some cases, the investigation could involve the analysis of a set of transactions on an account, with the analyst seeking to decide whether or not to block the card (this is the approach assumed in this report). In this instance, the analyst would take a more forensic approach to the behaviour of the card holder and the use of the card, relative to some concept of normal activity. In some cases, investigation could be at the level of transactions, in which the analyst seeks to identify patterns of criminal activity involving several cards. In this instance, the analysis would be looking for evidence of stolen details or unusual patterns of use of several cards, say multiple transactions in different locations within a short timeframe. Other functions that people can perform in the fraud detection process include: risk prioritisation, fast closure of low risk cases, documentation of false positives, and identification of risk profiles and fraud patterns. Once a potentially fraudulent transaction is detected (which occurs in near-real-time), a decision needs to be taken as to whether or not the credit card is blocked and the transaction flagged. Such a decision is often best made given multiple sources of information pertaining to the transaction.

For the model (and the experiment using human participants reported in chapter 4) we assume that credit

card fraud analysts attempt to identify fraudulent patterns in transaction datasets, often characterised by a large number of samples, many dimensions and online updates. For this study, we defined nine credit card transaction attributes Table 4) as relevant to the detection of credit.

Table 4: Nine information sources.

The values that assigned to be fraudulent/normal for the information sources with star sign (*) were counterbalanced across participants.

Info Sources	Normal	Fraudulent	Validities
Transaction Amount	≤ 500	> 510	0.6
Transaction History	3 small amounts in a row	N/A	0.7
Card Present	YES	NO	0.65
CVV Entered	YES	NO	0.55
Card Issued Bank*	Hanford	NorthWest	0.6
Purchase Made in*	Europe	USA	0.85
Card Expiry Check	≥ 5 days	≤ 4 days	0.55
Transaction Time	6:00-20:00	20:00-6:00	0.60
Type of Goods*	Travel agent	Electrical goods	0.55

Information for each source was presented in binary terms based on rules for fraudulent and non-fraudulent behaviour (Table 4). The cues had *validities* [0.85, 0.7, 0.65, 0.6, 0.6, 0.6, 0.55, 0.55 and 0.55], where validity was defined as the probability that the cue indicated fraud given that the ground truth of the transaction is fraudulent. Validities were arbitrarily assigned to the nine cues for the purposes of this report.

3.4 Using Visualization

When confronted with a collection of data with different provenance, reliability and salience to their task, analysts need to engage in a number of processes to ascertain the most appropriate sources to use and the most appropriate tasks to apply these sources. These processes carry with them a cost, and the cost structure can be considered in terms of the resources available to the user and the methods that they could apply to exploit these resources (Kandel et al., 2012). The resources themselves could relate to the content and visual appearance of information sources (external resources) and to the knowledge, effort and ability of the user (internal resources). The methods relate to the actions that can be performed to access, interpret and collate the content of the information sources. For Russell et al. (1993) the main cost arises from data extraction. The implication is that the design of visualization can have an impact on this data extraction cost, such that it should be possible to elicit differences in extraction cost (e.g., measured by search time) from different layouts, content or features in the visualization.

3.4.1 Visual Perception

Studies of visual perception show that perceiving a pattern, involves a complex sequence of eye movements to gather information and maximize the utility of the decision (Hayhoe and Ballard, 2014; Nunez-Varela and Wyatt, 2013; Sprague and Ballard, 2007; Trommerhauser et al., 2009). This is a

process of active vision. People move their eyes to seek out items within a visual scene that are relevant to the task or question they are engaged in. Eye movements are necessary since only a very small area of what we look at is visible at high resolution at any one point in time. This area covers only 1-2 degrees visual angle and is called ‘foveal’ vision: the fovea has the highest density of daylight/color vision receptor cells. With increasing eccentricity, there is a sharp drop off in the density of these cells, and hence vision becomes rapidly blurred. To guide eye movement from one item to the next, people are generally believed to use information gathered from peripheral vision to guide saccadic eye movements. The periphery, covering a much larger area than the fovea, still contains useful information despite the reduced acuity. It is well known that peripheral vision plays a key role in guiding eye movements during visual search (Geisler, 2011; Kieras and Hornof, 2014) but less is known about the role of eye movements in the use of visualizations for decision making. Therefore, it is important to understand the strengths and limitations of designing displays that enable the use of peripheral vision in visual search.

3.5 Modelling Theory

In order to model how people use visualizations in credit card fraud detection we build on a previous model of visual search (Chen et al., 2013; Tseng and Howes, 2015). In this approach eye movement strategies, scan paths and stopping rules, are an emergent consequence of the visualization and the limits of human vision. The assumption is that people choose which cues to look at first and when to stop looking at cues informed by the reward that they receive for the decisions they make. Better decisions will receive higher rewards, which will reinforce good eye movement strategies.

This approach is known as reinforcement learning and it makes use of optimal control and Machine Learning methods (Baron and Kleinman, 1969; Russell and Subramanian, 1995; Sutton and Barto, 1998). A key contribution of this literature has been to provide a formal basis for learning a strategy, including eye movement strategies, given only a definition of the reward function, the state space, and the action space. The strategy learnt (the control knowledge) is then that which determines what-to-do-when. In the case of information gathering in service of decision making, it concerns where to get information and when to make the decision. Importantly, the aim of this behaviour work is that behaviours, such as search rules, stopping rules and decision rules, should emerge from theoretical assumptions, rather than being encoded/assumed by the researchers.

In this framework, the expected value of an action given a state is the sum of an immediate reward plus the rewards that would accrue from subsequent actions if that action were selected. This simple assumption has provided a means of deriving human visual search strategies in well-known laboratory tasks (Chen et al., 2013) and menu search tasks (Chen et al., 2015). It also provides a means by which to derive credit card fraud detection strategies given assumptions given different visualization techniques, but only if the decision making problem can be defined as a reinforcement learning problem. In the following paragraphs, we first report the learning problem and then report pilot data that tests the model predictions.

3.5.1 Problem formulation

We assume that the problem faced by a decision analyst can be modelled as a Partially Observable Markov Decision Process (POMDP). The process is Partially Observable because the true state (the

values of the cues) is unknown to the analyst and can only be observed through a noisy and uncertain foveated visual system. The process is Markovian because the outcomes are determined by the response to the display, i.e., in terms of where to look and whether or not to block a transaction.

A decision can be made by choosing from scanning and choice actions, $a \in A$. The action selections are dependent on the observations, and their history. At each moment, the environment is at one state $s \in S$. The state is not fully observed by the analyst. Instead, by interacting, observations, $o \in O$, and rewards, $r \in R$ are received from the environment, i.e., the environment is partially observable. This action-observation-reward sequence happens in cycles indexed by $t = 1, 2, 3, \dots$. The action-observation sequence is used to update the estimate of the belief about the true state using Sequential Bayesian updating (explained below). Q-learning is used to learn which action to do next (e.g., to gather more information or to make a decision) given the current belief about the state. It does so by learning the belief-action values through simulated experience. Belief-action values are updated incrementally (learned) as reward and cost feedback is received from the interaction during the simulated experience. For example, if the model looks at cue A and subsequently makes an incorrect decision, then the value of cue A to the model will decrease. With enough simulation trials, the optimal strategy will emerge and the model will take the best actions given the beliefs.

In the following sections more detail is provided about how the belief update and optimal controller work.

The basic concept in POMDPs is following: the states cannot be directly observed by the agent. Instead, the decision maker receives two signals from the environment: (1) observations, determined by the observation model, and (2) the rewards, determined by the reward function. The goal is for the agent to choose actions at each time step that maximise its expected future discounted reward. Many POMDP algorithms have been proposed to find the optimal policy to do so. In our model, we used Q-learning algorithm to learn the optimal strategy. More details are given below.

A POMDP is defined by the following elements:

- S : At each time t , the environment is in a state $s_t \in S$. A state represents a true information pattern presented on the user interface. As shown in Figure 12 and Table 4, nine attributes associated with credit card transactions are presented on the interface. According to the experiment to be modelled, the value of each attribute was discretised as two levels, representing 'fraudulent' and 'normal' respectively. For example, one of the states is a 9 element vector as follows: [F N N F F F N F N], each item of which represents the value for one attribute (F for fraudulent and N for normal). Therefore the size of the state space is $2^9 = 512$.
- A : A set of actions that the decision analyst can take. It consists of the information gathering actions (i.e., which attribute to fixate at, i.e., f_i ($i \in 1, 2, 3, \dots, 9$)) and decision making actions (i.e., block/allow transaction). Therefore, the size of the action space is 11.
- O : A set of observations that can be made. The observation at step t is defined as the information gathered up until t for all the information sources. Information is gathered through both foveated vision and peripheral vision. The observation is a 9 element vector. Each element has three levels, F (fraudulent),

N (normal) and U (unknown). For example, one of the observation is [F N U, F U U, U N N], each element of which represents the information gathered for one attribute. Therefore the upper bound of observation space is $3^9 = 19683$.

- $R(S, A)$: A set of rewards generated by the environment. At any moment, the environment that occupies in one of the states s , generates a reward $r \in R$ in response to the action taken a . The reward for the information gathering actions is the time cost. The time cost includes both the dwell time on one sub-window and the saccadic time cost travelling across sub-windows. More details about the time cost is provided in subsection ‘Time Cost’ below. The reward for a correct decision was +10; the penalty for an incorrect selection was -20.

- $T(S_{t+1} | S_t, A_t)$: This transition function describes how the state changes based on the actions taken. In the current task the information pattern (i.e. the state) across time steps (within one trial) did not change. Therefore, $T(S_{t+1} | S_t, A_t)$ equals to 1 only when $S_{t+1} = S_t$. $T(S_{t+1} | S_t, A_t)$ equals 0 otherwise. That is, the state transition matrix is the identity matrix.

- $p(O_t | S_t, A_t)$: This observational function describes how states and actions combine to yield observations. In the experiment, the information was represented on the interface by either color blocks or text. It is known that an object’s color is more visible in the periphery than the object’s text label. In our model, the observation model is based on the acuity functions reported in (Kieras and Hornof, 2014). The observational model is explained in more detail in the subsection ‘Observation Model’.

- γ : At each time step t , the agent receives the reward $R(s_t, a_t)$. Its goal is to maximize its expected long-term reward $E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$, where γ is a discount rate, $0 \leq \gamma < 1$.

Because the agent does not directly observe the environment’s state, the agent must make decisions under uncertainty of the true environment state. By interacting with the environment and receiving observations, the agent update its belief in the true state. Below we describe how this belief update is computed. An example update from $t = 0$ to $t = 1$ is given, which could be generalised to from t to $t + 1$.

At $t = 0$, the environment occupies in a state $s \in S$ we have an initial belief, $\vec{B}_{t=0}$, that is assumed to be a uniform distribution over all possible states. This means that without any evidence, the model believe that the environment is equally in one of the possible states. At $t = 1$, the agent takes an action a_1 , which causes the environment to transition to state s' with probability $T(s' | s, a)$. After reaching $s' \in S$, the agent observes o_1 with probability $P(o_1 | s', a_1)$. An agent needs to update its belief upon taking the action a_1 and observing o_1 .

$$B_1(s') = \frac{\sum_{s \in S} B_0(s) \times T(s' | s, a_1) \times p(o_1 | s', a_1)}{\sum_{s' \in S} p(o_1 | s', a_1) \sum_{s \in S} T(s' | s, a_1) B_0(s)} \quad (1)$$

As mentioned above $p(s'|s, a_1) = 1$ only if $s' = s$, and 0 otherwise, Equation (1) can be simplified as Equation (2):

$$B_1(s) = \frac{B_0(s) \times p(o_1 | s, a_1)}{\sum_{s \in S} p(o_1 | s, a_1) B_0(s)} \quad (2)$$

At each time t , a belief \vec{B}_t vector consists of a probability for each of possible states, $B_t(s_i)$, where $i \in 1, 2, 3, \dots$. Each element $B_t(s_i)$, of the belief vector is updated independently. The estimate of the state (i.e., belief) is summarised in the vector, \vec{B}_t . We use this as a prior for next update when a_{t+1} and o_{t+1} is receiving.

3.5.2 Observation Model

The observation obtained is constrained by the human visual system. In the experiment, the information was presented either by color blocks or in texts. It is known that color plays a key role in visual search (Sprague et al., 2007; Kieras and Hornof, 2014). As that it can be perceived from a wide range of eccentricity, it often serves as a guide of the eye movements (Gordon and Abramov, 1977). In contrast, the text recognition would require a fixation on the texts unless the text is very large (Kieras and Hornof, 2014). In our model, the observational is implemented based on the acuity function reported in (Kieras and Hornof, 2014).

3.5.3 Peripheral vision

Our model assumed that the semantic of the text information was obtained only when it was fixated. The color acuity was specified as a quadratic psychophysical function from (Kieras and Hornof, 2014). This function was used to determine the availability of the color in each cue given the eccentricity and the size of the item. In our model, the acuity function was represented as the probability that each visual feature of the item was recognised.

$$P(\text{available}) = P(s + X > \text{threshold}) \quad (3)$$

where $\text{threshold} = a \times e^2 + b \times e + c$; $X \sim N(s, v \times s)$; s is the item size; e is eccentricity. In the model, the function were set with parameter values of $v=0.7$, $b=0.1$, $c=0.1$, $a=0.035$ as in (Kieras and Hornof, 2014). The objection size s was chosen according to our experiment. The color blocks used in the experiment is about 4 degrees in visual angle. These parameter settings resulted in the acuity function shown in Figure 16. On each fixation, the availability of the color information was determined by these probabilities.

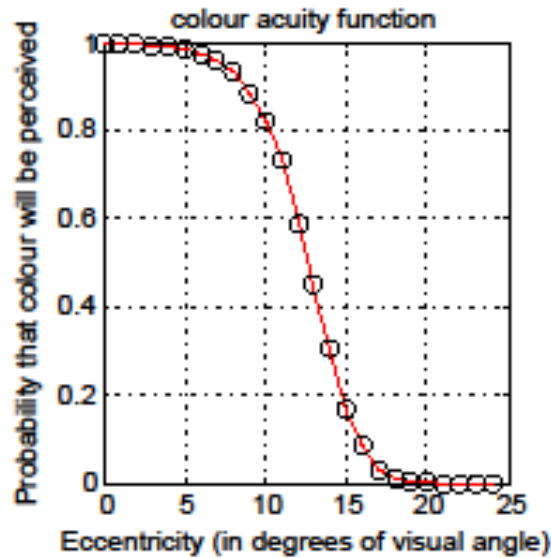


Figure 16: The color acuity function used in the optimal control model.

3.5.4 Time Cost

3.5.4.1 Saccade duration

The saccade duration D (in milliseconds) was determined with the following equation [2]:

$$D = 37 + 2.7A \quad (4)$$

where A is the amplitude (in terms of visual angle in degrees) of the saccade between two successive fixations.

3.5.4.2 Fixation duration

The fixation durations used in the model were from the measurements from the study using human participants that we report in chapter 4.

3.5.5 Learning

The control knowledge is represented as a mapping between the beliefs and actions, which is learnt with a reinforcement learning algorithm, Q-learning. Further details of the algorithm can be found in any standard Machine Learning text (e.g., Sutton and Barto, 1998).

Before learning, an empty Q-table was assumed in which the values (i.e., Q-values) of all belief-action pairs were zero. The model therefore started with no control knowledge and action selection was entirely random. The model was then trained until performance plateaued (requiring 10^5 trials). The model explored the action space using an ϵ -greedy exploration. This means that it exploited the greedy/best action with a probability $1 - \epsilon$, and it explored all the actions randomly with probability ϵ . ϵ was set to 0.1 in our model. Q-values of the encountered belief-action pairs were adjusted according to the reward and cost feedback, as shown in Equation (5).

$$Q(b,a) \leftarrow Q(b,a) + \alpha[r + \gamma \max_{a'} Q(b',a') - Q(b,a)] \quad (5)$$

where $Q(b,a)$ is the Q-value for one belief-action pair (b, a) , r is the immediate reward/cost obtained while the action a is taken, α is called learning rate, and γ is called discounted factor.

The idea is that, these Q-values are learned (or estimated) by simulated experience of the interaction tasks. The true Q-values are estimated by the sampled points encountered during the simulations. The optimal policy acquired through this training was then used to generate the predictions described below.

While we used Q-learning, any reinforcement learning algorithm that is guaranteed to converge on the optimal policy is sufficient to derive the rational adaptation (Sutton and Barto, 1998). The Q-learning process is not a theoretical commitment. Its purpose is merely to find the optimal policy. It is not to model the process of learning and is therefore used to achieve methodological optimality and determine the computationally rational strategy (Lewis et al., 2014).

In summary, Q-learning was used to learn (or estimate) the value of each belief-action pair by simulated experience of the interaction tasks. The optimal policy is then the greedy policy given the Q-values. The model was implemented in Matlab.

3.6 Results

3.6.1 Information used and accuracy

As shown in Figure 17, the model predicted that more cues should be used in the ‘visible/text’ (VT) condition than in the other conditions. This is because the information in the visible conditions is much cheaper, compared with covered/color (CC) and covered/text (CT) where it had 1.5 seconds delay for each cue (based on the experiment design). However, because of the difference between the acuity function for colour and text, fewer cues were fixated by the model in the visible/colour condition than in the visible/text condition.

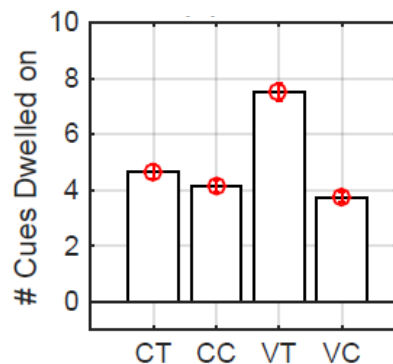


Figure 17: Information Sources used by the model

The model achieved about 79% accuracy across the four conditions (Figure 18). The accuracy level is near ceiling given the available cue validities and is likely to reflect the importance of accuracy given the high temporal cost of an incorrect decision.

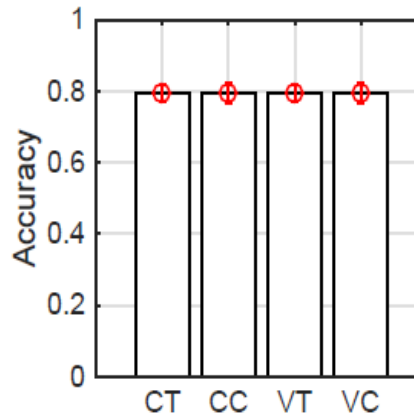


Figure 18: Accuracy achieved by the model

3.6.3 Decision time distribution

Long-tail left-skewed distributions are a signature feature of human decision times. Figure 19 shows the response time distributions for the model. The distributions are interesting because despite the fact that the no skewed distributions are assumed in the model performance they do emerge as a consequence of adaptation to the constraints.

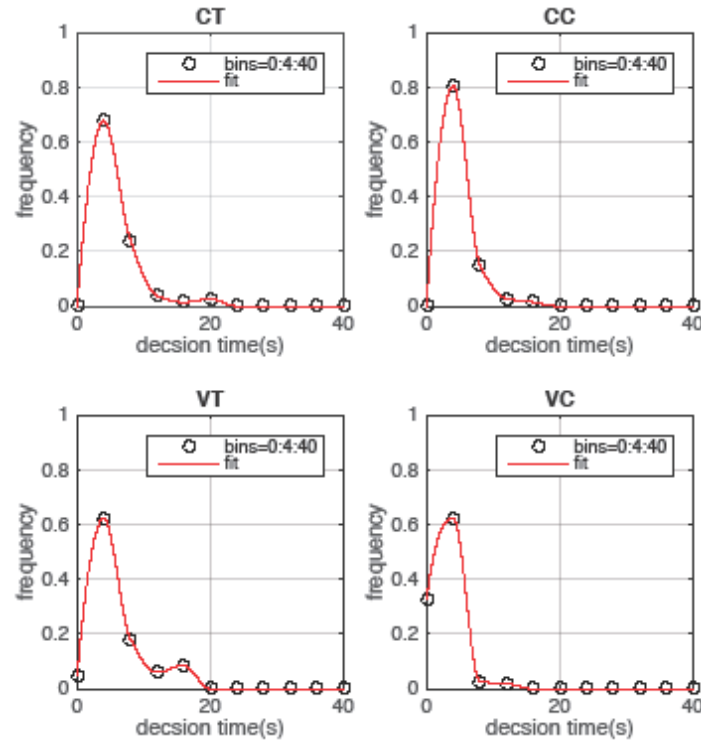


Figure 19: Model decision time

3.7 Discussion

We have reported a novel model of how people use information visualisation to make decisions. The model shows how different types of information visualisation lead to the emergence of different user strategies. In particular, it shows how when colour block visualisations were made available without the need for mouse clicks to reveal the data, more use should be made of peripheral vision to gather information. This result is a consequence of the fact that emergent strategies for information gathering can take advantage of the different human acuity functions associated with colour and with text; whereas colour information can be obtained from the periphery, text must be foveated to be understood.

In addition, the model led to the derivation of an optimal strategy given a POMDP problem formulation; this strategy involved using a weighted integration of the optimal number of the best cues. These cues provide the right information at a time cost that is best given the trade-off between time and accuracy imposed in the experiment.

The strategies that the model used are comparable to those used by humans (see chapter 4). This suggests that model was able to discover strategies to complete the task in an optimal manner. This means that, rather than being programmed with strategies (which is typical of other approaches to modelling activity, such as Critical Path Analysis that we consider in chapter 6), the strategies emerge from the performance of the model. This is important because it suggests how the theory might be developed in the future so as to rapidly evaluate the usability of a broader range of visualisations for a broader range of decision tasks. A key factor here is that the model is based on a very general modeling framework. The model is based on a specification of the decision *problem* faced by a people with foveated vision and it is not based on a specific set of decision heuristics. Strategies for a specific visualisation are then learnt through interaction. It should therefore be possible to apply the model to a broad range of visualisation technologies and automatically derive predictions for their usability.

4 Experiments on Decision Making using Different User Interface Designs: Credit Card Fraud Use Case

4.1 Introduction

In the credit card fraud use case, operators at the customer service level have to explain the reasons for blocking a transaction or card as they are speaking with the customer. At the higher level of fraud analysis, analysts need the ability to look for patterns on both, ‘local’ and ‘global’ levels, based on an extremely large number of parameters that may factor into a machine learning algorithm for fraud detection. Hence, both use cases require the accessible presentation of (at least) those data that are driving the automatic system. However, humans are known to be efficient when it comes to expending effort on information search (Pirolli and Card, 1999, Gigerenzer and Gaissmaier, 2011). This efficiency was indicated in the modeling research (chapter 3) and leads to the conclusion that simply making a large amount of data available to an operator does not mean that all of it would actually be used. In order to appreciate how people might use such information selectively, we need to consider recent work on human decision making.

4.2 Describing Human Decision Making

In the psychology literature, extensive studies have been conducted to understand how people gather information in service of decision making. One influential approach is based on the idea of fast-and-frugal heuristics (Gigerenzer and Goldstein, 1996; Gigerenzer and Todd, 1999). This approach makes a strong commitment to the heuristic theory of decision making (Gigerenzer and Gaissmaier, 2011) which assumes that people use simple rules in order to make good decisions fast, rather than computing exact probabilities based on all available information. One prominent decision heuristic has received great attention: the take-the-best (TTB) heuristic (Gigerenzer and Todd, 1999). The TTB heuristic consists of a set of rules concerning the most important aspects of information gathering: the search rule, the stopping rule and the decision rule. This requires knowing the validity of cues. Validity is the probability that the information represented by a cue will lead an observer to the correct decision. A validity of 1 will always lead to the correct decision, whereas a validity of 0.5 will result in chance performance. A person using TTB searches from the most useful information (with the highest validity) to the information with the lowest validity (the search rule). Information search terminates once a cue discriminates between the considered options or once all cues have been examined (the stop rule); at which point the model chooses the option favoured by the discriminating cue (the decision rule). For example, when people are asked to determine whether Berlin has a higher population than Neuss, they might use a high validity cue (recognition) to tell them that, as they have not heard of Neuss, Berlin must be bigger. However, if they are asked whether Stuttgart is bigger than Nuremburg, both of which are recognised, then they may move to the next best cue, say whether or not they have heard of a football team from that city.

Following Gigerenzer and Todd (1996) who showed how Take-The-Best (TTB), could describe human information gathering and decision making behaviours, a number of articles offered empirical investigations into which heuristics people choose (Broder and Gaissmaier, 2011; Broder and Schiffer, 2006; Lee and Zhang, 2012; Newell and Shanks, 2003; Newell et al., 2003). A particular concern in this research has been whether people use a heuristic, Take-The-Best (TTB), that uses information selectively, or a heuristic such as Weighted-ADDitive (WADD) that integrates all information (Rieskamp and Hoofrage, 2008; Rieskamp and Ott, 2006). While this debate has been waged in the psychology literature, it is highly relevant to understanding how information visualisations are used by people, and therefore to how visualisations should be designed. If people use TTB and not WADD then they may make use of a much smaller part of the displayed information than if they use WADD.

4.3 Experiment 5.2.1

4.3.1 Experimental Design

The design of the task for Experiment 5.2.1 (which was modeled in chapter 3) was based on the SPEEDD credit card fraud use case. We addressed two design questions: the immediate availability of data, and the representation of that data. Firstly, we focussed on data presentation using either a ‘full view’ (F) design like a dashboard and a ‘multiple-screen display’ layout in which information has to be revealed (R), such as in a menu or through data query. In terms of data representation, we focussed on the provision of exact data (D), such as numerical values for transaction amounts or details of date and time of a card transaction, and the provision of abstracted and pre-interpreted data, represented for example in a traffic light colour coding (C). This produce a combination of four experimental conditions: FD, FC, RD, RC.

The four experimental conditions were designed to examine the effect of two specific display modifications: firstly, we were interested in differences in behaviour when evaluating information sources that displayed data based on a simple colour scheme (green – possibly normal, red – possibly fraud) as compared to presenting the original data. Secondly, we were interested in how the direct availability of information affected user behaviour, hence comparing a condition in which information had to be revealed sequentially with a condition in which all information was in view straight away. Figure (20) shows the user interfaces for each of the four experimental conditions:

1. R_D – user has to reveal information, information displayed as data
2. R_C – user has to reveal information, information displayed as colour
3. F_D – full information available immediately, information displayed as data
4. F_C – full information available immediately, information displayed as colour



Figure 20: User interfaces for the four experimental conditions
 Top left: R_D, top right: R_C, bottom left: F_D, bottom right: F_C.

In the ‘reveal’ conditions, a blank screen was shown for 1.5 seconds once the ‘reveal’ button was clicked. This was done to prevent participants rapidly revealing all information and to consider the associated time cost; in an applied scenario, the time cost would arise from querying data from a sensor or database or from selecting parameters from a dropdown menu.

4.3.1.1 Scenario

Participants took on the role of a fraud analyst at a bank. The task was to screen flagged credit card transactions in order to decide whether the transaction should be blocked (prevented from being authorised) or allowed. The requirement was for participants to correctly classify 100 transactions as quickly as possible while minimising the number of mistakes. Participants were further asked to learn the usefulness of the individual information sources (or ‘cues’) in order to achieve best performance and were questioned about the cue ranking after completing the task. The ratio of normal and fraudulent transactions was 50:50, which was not disclosed to the participants. If participants were guessing, the chance level accuracy was hence 50% and the participant would be expected to complete approximately 200 transactions. The target number of correct transactions was based on comparable experiments (Newell and Shanks, 2003, Newell et al., 2003) as well as a sensitivity analysis performed prior to commencing the study, in which the effect of completed transaction number on the accuracy of the estimated usefulness of information was calculated (for more details about ‘validity’ estimates see chapter 3).

4.3.1.2 Participants.

Sixteen participants were recruited from staff and students at Birmingham University. No financial or other incentives were given to participate except for snacks before/after the study. Participants were equally and randomly assigned to four experimental groups which are detailed below. Eleven men and five women took part with a mean (SD) age of 31.6 (7.5) years. The study population reflects the pilot character of this work; in future, the population size will be increased.

4.3.2 Method

4.3.2.1 Set up and Data Acquisition

A custom user interface was created in Matlab (The MathWorks), consisting of two main panels: one large panel on the left, displaying information associated with a credit card transaction, and one small panel on the right, which served as the main user interaction unit (Figure 21, showing one of our four UI design concepts). The panel containing information associated with the transaction showed nine panels laid out in a 3 x 3 grid, each panel representing an individual information source. The panel on the right side guided the user through the workflow of each card transaction while logging decisions and providing feedback.

Figure 21: General layout of the abstracted user interface that was then adapted to the four experimental conditions.

The large panel on the left (1) displays nine attributes associated with a credit card transaction. The small panel on the right (2) served as the main user interaction unit, where decisions were logged and feedback provided. Panel 1 differed across the four experimental conditions, whereas panel 2 remained the same.

Nine card transaction attributes were selected as relevant to the detection of credit card fraud based on the literature and discussions with domain experts from FICO, FeedZai and the UK Cards Association. These are shown in table 5.

Table 5: Nine information sources used in the experiment.

Nine attributes (ROI: region of interest) associated with credit card transactions were presented to inform a participant about the likelihood of fraud. Since the different sources have a different validity (percentage of trials predicted correctly), participants have to learn which sources are most reliable in order to perform the task most efficiently and effectively. This table shows ranges for normal and fraudulent transactions; the attributes 'card issued by bank', 'purchase made in' and 'type of goods' were counterbalanced between trials.

	ROI 1	ROI 2	ROI 3	ROI 4	ROI 5	ROI 6	ROI 7	ROI 8	ROI 9
	Transaction amount	Transaction history	Card present	CVV entered	Card issued by bank	Purchase made in	Card expiry check	Local transaction time	Type of goods
Normal	≤ 500	N/A	Yes	Yes	Hanford	Europe	≥ 5 days	6:00 – 20:00	Travel agent
Fraud	≥ 510	3 small amounts	No	No	NorthWest	USA	≤ 4 days	20:00 – 6:00	Electrical goods
Validity	0.6	0.7	0.65	0.55	0.6	0.85	0.55	0.60	0.55

The information sources were independent, and information displayed for each source could be evaluated in binary terms based on rules for fraudulent and non-fraudulent behaviour (Table 5) given to the participants. Each attribute had a specific validity; validity describes the fraction of transactions which a participant would evaluate correctly if he/she always based the decision on the suggestion of this attribute. For example, for an attribute with a validity of 0.6, 20 normal transactions would comprise 12 instances correctly suggesting that the transaction is normal and 8 instances incorrectly suggesting that the transaction is fraudulent. The validities for the attributes were selected to approximate a power function. The exact values displayed for the attributes were randomly sampled from two uniform distributions, one for fraudulent transactions and one for normal transactions, which were separated by a definite threshold.

The location of each information source within the UI was assigned at random for each participant (to avoid confounds due to information position), but kept constant across all trials for that participant (to avoid confusion due to changes in position and to help participants learn strategies for searching for information). Three attributes that were ambivalent were counter-balanced (bank issuing the card, location where the purchase was made and type of goods purchased). The sequential presentation of transactions with the associated attributes was based around a pseudo-random blocked design.

Participation in the study started with reading/signing the information sheet and consent form, followed by reading the detailed written study instructions. Study instructions were modified to suit each of the four experimental conditions. A participant was then asked whether any questions remained, and after clarifying these the participant was seated in front of 22 inch screen instrumented with an eye tracker (X2-60, Tobii, Sweden) recording gaze data at 60 Hz. The eye tracker was operated through Matlab using the Tobii SDK and Matlab binding. The study started by calibrating the eyetracker using a standard 5-point system. A participant then worked through the first transaction with the opportunity to ask any clarifying questions. After this, participants worked through transactions at their own pace.

The workflow for evaluating a transaction was as follows: the participant examined as many information sources (either by revealing them in the R_D and R_C condition or by simply looking at them in the F_D and F_C condition) as he deemed necessary to make a decision. He would then indicate this decision in the right panel as 'Allow transaction' or 'Block transaction'. The tick box 'not sure' could be selected for future examination of thresholds. Following the participant was instructed to tick those information sources using the provided checkboxes that he deemed most important for making the decision. Following this, the 'submit' button had to be clicked in order to receive feedback regarding the correctness of the decision. By clicking 'OK', the UI was then saved out and the next transaction could be called up using an interim screen that had the button 'next transaction'. The button was placed in the top right corner of the UI above the workflow panel to not confound initial gaze data above any of the nine information sources. At intervals of 15 minutes, participants were offered the opportunity to take a short break if they wished.

While the participant was performing the task, the following data were saved out for each transaction: raw gaze data and associated attributes for both eyes; time between starting the transaction and making the decision; cues revealed (in the R_D and R_C condition) and the time at which they were revealed; decision made, ground truth, transaction attributes and correctness of the decision; status of all checkboxes; timings of interacting with the user panel; and a screenshot of the final state of the UI before proceeding to the next transaction.

At the end of the study, participants were given a questionnaire asking them to (1.) rank the nine cues in order of their perceived usefulness; (2.) describe their strategy; (3.) indicate whether they noticed any patterns; (4.) provide any general comments on the experiment; and (5.) rate the perceived difficulty of the experiment.

From the recorded data, the following features were calculated for each transaction:

- (1.) the actual decision time, from which the time that the participant spent looking at the user interaction panel was subtracted. Hence the time reflected only the time the participant spent looking at the information sources. For the 'reveal' conditions ('R_D' and 'R_C'), the time taken to reveal cues was further subtracted to make times comparable between all four conditions (number of revealed cues * 1.5 s).
- (2.) The total number of revealed cues for the 'R_D' and 'R_C' condition based on the mouse interaction.
- (3.) The number of cues dwelled on (fixed with the eyes for at least 200 ms) based on the gaze data.
- (4.) The number of visual transitions ('switches') between the available UI panels based on the gaze data.
- (5.) Using network analysis, the number of 'edges' in a viewing network, which is the number of any pair of information sources between which a participant switches the gaze.
- (6.) The maximum dwell time per information source. The Kruskal-Wallis test was used to examine whether there were significant differences for any of these parameters between the four experimental conditions.

Across all trials, the accuracy of the participant was calculated as the number of correct decisions divided

by the total number of trials. Based on the gaze data and network analysis, viewing networks were constructed across the last 15 trials based on the scanning behaviour of the participant, where the number of edges (or switches) between the same two information sources was counted across the trials. This visual representation allows to examine preferred scan patterns across participants and conditions to evaluate whether scanning became very systematic both within and across participants.

From the questionnaires, cue rankings were extracted and compared with the order in which information sourced were attended to. For this purpose, the median order of attending an information source for the first time within a scan pattern was calculated across all trials per participant. Answers to the conceptual questions were qualitatively summarised across participants.

4.3.4 Results

4.3.4.1 General performance.

The mean decision time (figure 22) for a single transaction ranged from 3.5 to 21.1 s across all participants, with a mean (SD) decision time of 6.1 (3.0) to 12.3 (5.9) s across conditions. There was no significant effect of the experimental condition on the decision time ($p = 0.206$).

The accuracy of participants (figure 22) ranged from 52.1% (chance level) to 79.4% across all participants, with a mean (SD) accuracy of 67.5% (10.7%) to 73.8% (6.0%) across conditions and no significant difference between conditions. For the last 15 transactions, mean accuracy stayed at a comparable level both across and within conditions, however the accuracy range across participants widened to 46.7% to 86.7%.

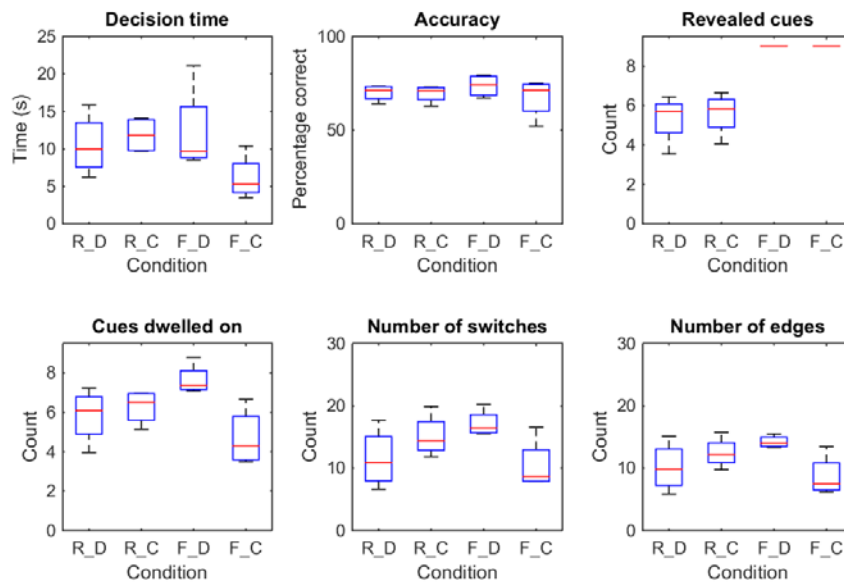


Figure 22: Boxplots summarising six outcome measures for each condition.

Abbreviated conditions: R_D / R_C – reveal information, data shown / colours shown; F_D / F_C – full information available immediately, data shown / colours shown.

4.3.4.2 Use of information sources.

In the Full conditions, use of information sources (figure 22) as indicated by the number of revealed cues could only be assessed for the two ‘reveal’ conditions, as in the ‘full’ conditions all cues were accessible straight away. In the ‘reveal’ conditions, the ‘data’ condition triggered a mean (SD) of 5.3 (1.3) cues and the ‘colour’ condition triggered 5.6 (1.1) cues to be revealed per transaction and participant (Figure x). There was no significant effect of the data representation on the number of revealed cues between these two conditions ($p = 0.686$, Mann-Whitney U-Test).

In the Reveal conditions, use of information sources as indicated by the number of cues that a participant dwelled on for longer than 200 ms could be assessed for all conditions and also compared to the figures for the cues that were eventually revealed. The mean (SD) number of cues dwelled on ranged from 4.7 (1.5) to 7.6 (0.8) cues per transaction and participant. The mean number of cues dwelled on was 5.8 and 6.3 in both ‘reveal’ conditions, 7.6 in the ‘full availability with data’ condition and 4.7 in the ‘full availability with colours’ condition (Figure 22). The effect of the condition on the number of cues dwelled on was significant ($p = 0.025$).

4.3.4.3 Visual scanning behaviour.

The mean (SD) number of gaze switches between cues was 13.6 (4.4) across all conditions, ranging from 6.6 to 20.2 switches across participants. Across conditions, the switch count ranged from 10.4 (4.2) to 17.1 (2.1) switches per transaction and participant. There was no significant effect of the condition on the number of switches ($p = 0.147$). For the two ‘reveal’ conditions, number of cues revealed and dwelled on was very similar.

The mean (SD) number of edges per trial in network analysis was 11.4 (3.4) across all conditions, ranging from 5.9 to 15.8 edges across participants. Looking at each source once in any order would have resulted in 9 edges (10 edges counting the return to region of interest 10, the user interaction panel). Across conditions, the number of edges ranged from 8.7 (3.3) to 14.2 (1.0) edges per transaction and participant. There was no significant effect of the condition on the number of edges ($P = 0.122$).

The constructed viewing networks showed that each participant developed a very unique approach to combining information sources visually. Figure 23 shows viewing networks across the last 15 trials. While the scanning behaviour of some participants was very varied and likely unsystematic across trials (e.g. participant 3, 6 and 13), other participants developed a very specific scanning approach (e.g. participant 9, 10 and 12). There was no clear trend for the experimental condition to cause differences in systematic scanning behaviour.

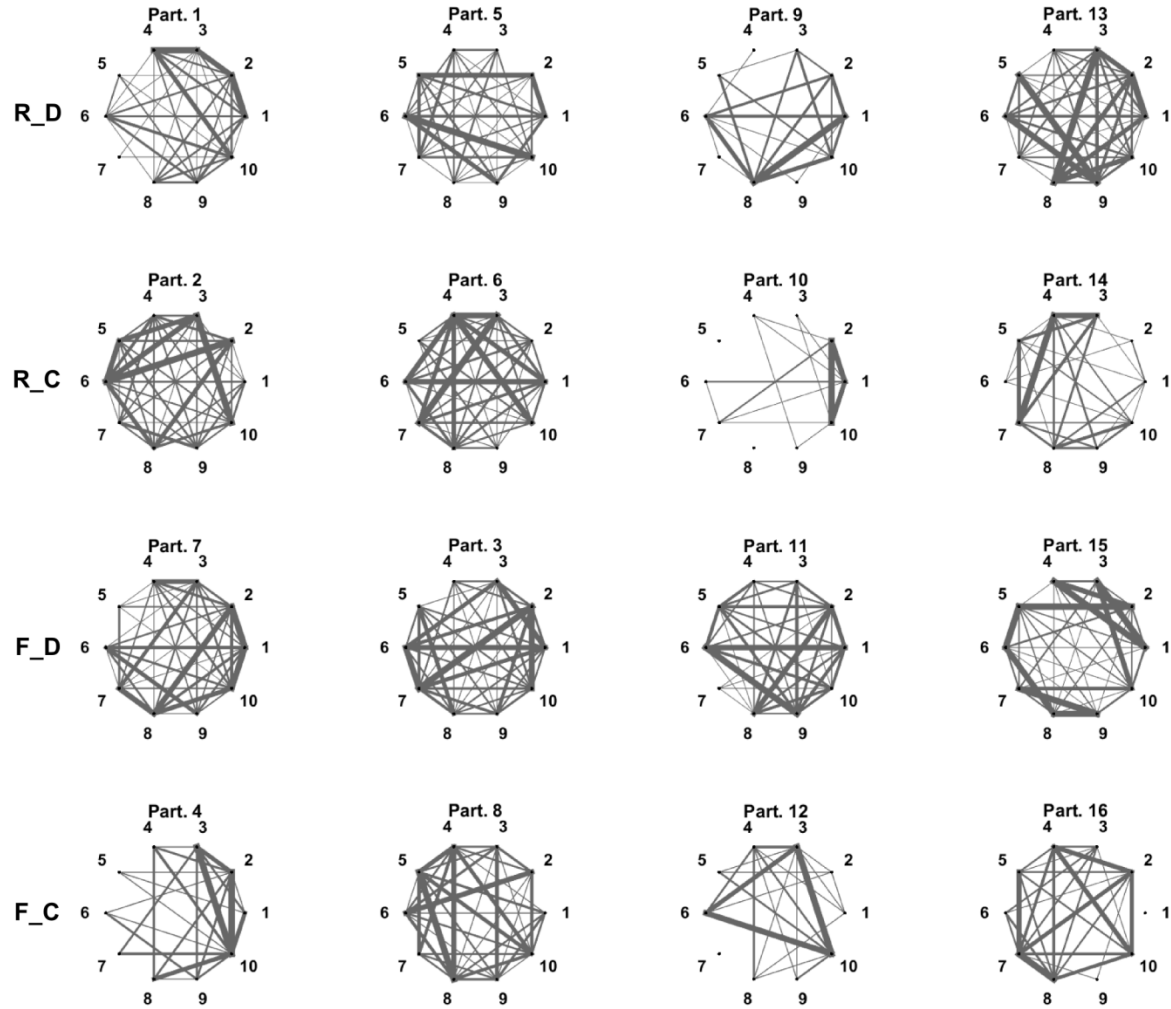


Figure 23: Viewing networks sorted by the experimental condition.

The thickness of lines represents for the number of trials during which a participant moved gaze between the two information sources connected by that line. Code of information sources: 1 - Transaction amount, 2- Transaction history, 3 - Card present, 4 - CVV entered, 5 - Card issuing bank, 6 - Purchase made in, 7 - Card expiry check, 8 - Local transaction time, 9 - Type of goods; region 10 corresponds to the user interaction window on the right of the UI.

4.3.4.4 Dwell times.

The effect of the experimental condition on dwell times varied across information sources (figure 24). For ‘transaction history’, ‘card present’ and ‘card expiry check’, the experimental conditions had a significant effect on dwell times ($p \leq 0.009$). For the information sources ‘transaction history’ and ‘card expiry check’, dwell times were longer in the two conditions showing the actual data compared to the corresponding conditions displaying colour. For the remaining six information sources, dwell times were very similar across conditions with no significant differences detectable ($p \geq 0.068$). Across the two ‘data’ conditions, mean dwell times across the four participants of each group ranged from 0.4 to 1.3 s. Across

the two ‘colour’ conditions, mean dwell times across the four participants of each group ranged from 0.4 to 0.9 s.

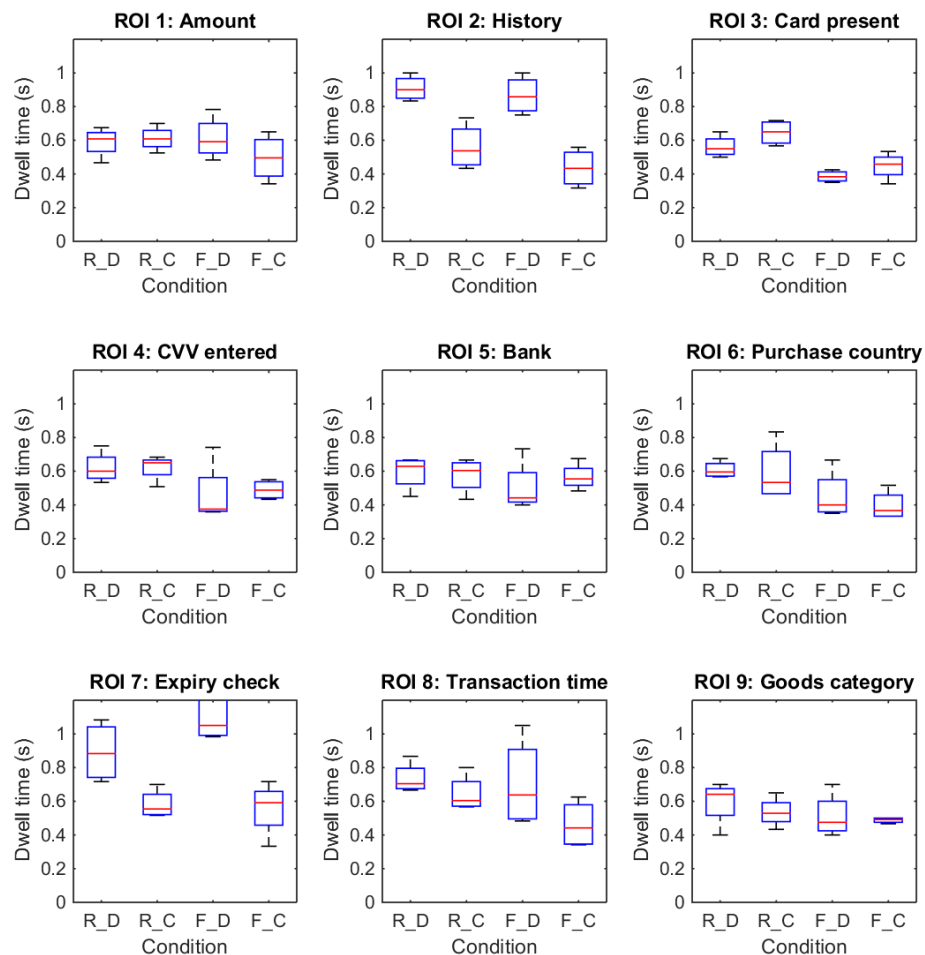


Figure 24: Dwell times for the nine information sources for each of the four experimental conditions.

The average maximum dwell time per trial generally ranges between 0.3 and 1.5 s. The largest differences between conditions are visible for ROI 2 and ROI 7, which require thought about the data while colours can be interpreted ad hoc. Abbreviated conditions: R_D / R_C – reveal information, data shown / colours shown; F_D / F_C – full information available immediately, data shown / colours shown.

4.3.3.5 Cue ranking.

The ranking of cues as provided on the questionnaires showed large variation in those cues considered most important. Four participants, one from each experimental condition, ranked the most important cue (‘Purchase made in, ROI 6, validity 0.85) as #1, and six participants ranked the second most important cue (‘Purchase history’, ROI 2, validity 0.7) as either #1 or #2. Overall, all but four participants ranked at least one of ‘Purchase made in’ or ‘Purchase history’ within the top 3. However, there was no obvious relationship between correctly identifying important cues and accuracy or decision time on the task. In

many cases, the most highly ranked cues corresponded to those cues which participants looked at the earliest. However, for some participants this match was not observed at all.

4.3.3.6 Questionnaire responses.

The mean (SD) difficulty rating of the task was 5.7 (1.6) across all participants and for each condition as follows: 4.8 (1.7) for R_D, 6.5 (1.3) for R_C, 5.0 (1.8) for F_D and 6.5 (1.3) for F_C. The answers of the participants to the three conceptual questions are summarised below for each question:

What was your strategy? If it changed during the experiment, what changed?

This question aimed to get an understanding of the participant's reasoning during the task. Strategies depended in part on the UI design, as for example overview scans were only possible in the two scenarios where all information was readily available. For example, one participant quickly scanned all categories to get a general idea regarding the case and then looked at detail, over time developing a 'gut feeling' after the initial scan. This strategy would only be possible in the 'reveal' condition if all information was revealed, requiring a long waiting time. Most participants reported evolving strategies as the task progressed, often starting with a few cues and then changing the focus on other cues or those cues which they thought were most informative. For example, one participant stated to reduce the number of used cues to be able to check them better and increase accuracy. Some participants combined cues to form a story, for example: "I also looked at card present and CVV (to see if this was an internet purchase)" or "[...] for example, has the person stayed out late at night or are they purchasing a one-off strange item. In these cases where there was a possible innocent explanation I generally let the transaction through". Some participants also employed assessment of cues following feedback: one participant tried to guess what would have been the correct cues given the current case and feedback. A second participant reported to, after some time of randomly revealing colour cues, pay attention to those cues that matched the outcome and then reveal those straight away on the next trial. A third participant tried at first to correlate individual cues with the outcome, and after feeling that this method failed to work tried to correlate groups of cues with the outcome.

Did you spot any patterns in the data or relationships between individual cues?

This question aimed to determine whether participants observed false correlations and were conscious of the independent nature of the cues. Responses to this question showed that a large number of participants (12 out of 16) felt that they had noticed relationships between cues. Again, some participants created stories involving the available information. For example, the relationship between card not present, CVV and local transaction time was related to shops being open or an online purchase being made; the transaction amount was related to the type of goods; and if the card was present and the CVV was entered, this was perceived as a likely online purchase and legitimate. A large number of participants reported false correlations without an explicit story, such as card expiry date and CVV, local transaction time and country of purchase, card present and country of purchase; Card expiry, CVV and transaction amount; transaction history and card issuing bank; transaction time and transaction history, purchase location and transaction amount, card issuing bank and CVV and card issuing bank and card present.

What are your general thoughts about the task?

Comments in this section were very varied. Several participants noted that they had enjoyed the task, for example stating “good”, “interesting”, “very interesting”, “could do it again; fun!”. Two participants highlighted that it was difficult to maintain accuracy for a given speed, one of them stating that it was hard to “progress at a reasonably fast pace while retaining focus and taking in the necessary amount of detail”. Three participants reported that they sometimes thought to have found useful relationships using a certain number of cues, but to then find that this was not reliable; this was perceived as “confusing” or “frustrating”. One of these participants sometimes started to guess if that happened. A further participant reported that “sometimes I could not find the logic in some of the trials, and couldn’t use the information learnt”. One participant found it “challenging to remember the results from previous runs and therefore was not able to use it in making my decision in the current run”. Another participant commented that, if real money was involved, he/she would have been likely to block more transactions.

4.4 Discussion

In Experiment 5.2.1, we examined how two distinct manipulations of UI design – data representation using values or colours and information availability based on reveal or full view – affect viewing behaviour in a fraud management task. Further, we investigated whether it is possible for humans to learn the importance of cues when faced with a large number of information sources of which many are not very useful. This task was very demanding, as even the highest cue validity was only 0.85 (meaning that in 15 out of 100 trials the cue would point to the wrong decision) and we wondered whether humans could detect important cues at all, given that these cues are evaluated in context of other cues, and have to be filtered down to the correct combination of few cues.

The decision times which we observed in this study ranged from 3.5 s to 21.1 s, with an average time of approximately 11 s for conditions ‘R_D’, ‘R_C’ and ‘F_D’, and approximately 6 s for condition ‘F_C’. This result fits in with the common observation that information which requires interpretation results in longer dwell times. We suggest that a system which uses colour coding to pre-classify data may allow human operators to spend minimal time looking at individual information sources. The question of whether this is an advantage or not depends on the nature of the decision to be made; if we wish people to screen the data and focus their attention on information which has been classified (e.g., by the automated system) then colour coding could be beneficial; if we wish people to check all of the data in order to ensure that the classification has been appropriately applied (e.g., where an automated system might perform below acceptable levels during to noise or uncertainty in the data) then colour coding might encourage a less effective strategy.

Nevertheless, it is worth noting that the decision times matched the times from experienced analysts from FeedZai (see previous D7.2), who took on average approximately 15 s to complete a task very similar to the present one. We will repeat a variation of this experiment with those analysts in order to check their performance and also to gauge their reaction to the validity of the task. We hence deduce that decision times by trained staff or untrained lay-people are comparable when evaluating UI prototypes. Decision times were substantially longer for FeedZai staff when working through cases using their proprietary software, often taking time in the order of minutes. It hence has to be highlighted that performance in ‘lab’

conditions has to be tested in situ in order to discriminate between effects in a lab scenario and a real-world scenario, where users may have different priorities when interacting with software.

We did not find evidence that any of the four UI concepts resulted in systematically superior performance of participants. This was reflected in the comparable accuracy of around 70% across designs. The UI design that showed full information but in form of data appeared to attract attention to more information sources whereas the design using colours attracted attention to the fewest number of information sources. This suggests that information either varied between these conditions (with the latter condition resulting in participants using less information, but this should have been reflected in the accuracy) or that information sampling using peripheral vision could be effective (as hinted at in the modeling work in chapter 3). One explanation of this behavior is that, having learned the location of information sources, it may be possible to rely on pattern recognition in order to integrate multiple sources. The role of peripheral vision in fast information integration will be of importance for future work and UI design.

The viewing networks showed noticeable variability in scanning the UI across even for only the last 15 trials. This highlights that participants did not converge on a single scan pattern, even after such a large number of training and learning trials, but rather stuck to a strategy where they tried different options until satisfied with the evidence to make a decision. If a consistent visual workflow is desired in future, we hence expect that a guided assessment of the evidence will result in more repeatable scan patterns than just training a participant on the target of work and then hoping that he/she will figure out the best way to do it alone. Guidance could be given in the form of required interaction with the UI, highlighting or eye tracking-based feedback.

The questionnaires revealed that some participants liked to make stories or tried to find patterns between multiple cues, where participants tried to find logic in individual card transactions and the relationship between cues. In the present experiment, cues were independent and any correlation between cues would have been due to similarly high cue validities. In our scenario, we would hence have expected ‘country that purchase was made in’ and ‘transaction history’ to point into the same direction most often. However, participants also reported associations between pairs of cues which were not statistically related (on our measures) but which they believed ‘made sense’ logically. This resulted in false correlations and incorrect decisions. In previous work (in intelligence analysis), it has been shown how users engage in parallel, overlapping explorations of data and often work with minimal and sketchy frames to explain these data (Baber et al., 2015). This work shows that we might anticipate that users will use a subset of the of available information, partly in an effort to reduce the cost associated with information access, partly as a result of the incremental construction of an explanatory frame to model the data, and partly as a result of the methods that they apply. When interpreting big data, ‘storifying’ the evidence could either become a confounding factor, especially when integrating independent information sources, due to people seeking to impose a structure on the data or could, with experience, be a way of identifying the relevance of data to analysis. The question of how sensemaking and storifying are used to respond to big data is significant to the development of future visualizations and visual analytics.

We found that accuracy on the task was substantially higher than chance, indicating that participants used the information sources provided to their advantage. Most participants learned to determine at least one of the most two important cues, however this did not necessarily lead to superior performance, not even

towards the end of the study. This finding illustrates how the selection of several inferior cues may still lead to performance comparable to using the best cues; this is a very relevant finding in context of the reliability of decision making when faced with big data. The results also provide some support for the notion that, at a population level at least, decision making seems to follow a ‘Take-the-Best’ approach, with the most relevant information be attended to more often than less relevant information, as show in figure 22. We examined the action sequences generated by the model to determine whether they corresponded to a Take-the-best strategy (TTB) or a weighted-additive strategy (WADD). TTB would be indicated by the participants selecting just the very best cue (which in our interface always discriminates) and then making a fraud/no-fraud decision. WADD would be indicated by the participants using all of the available cues. In fact, the number of decision cues used (Figure 20) is at neither of these extremes and, in addition, our informal inspection of the action sequences suggests that people use a range of intermediate strategies; they examine a few of the best cues, integrate information and then make a decision. Rather than following TTB or WADD, it would seem that our participants behaved more like the model reported in chapter 3, i.e., participants selected cues which appeared to provided that most information for their decision but limited their selection based on the time-cost (effort) of accessing the information.

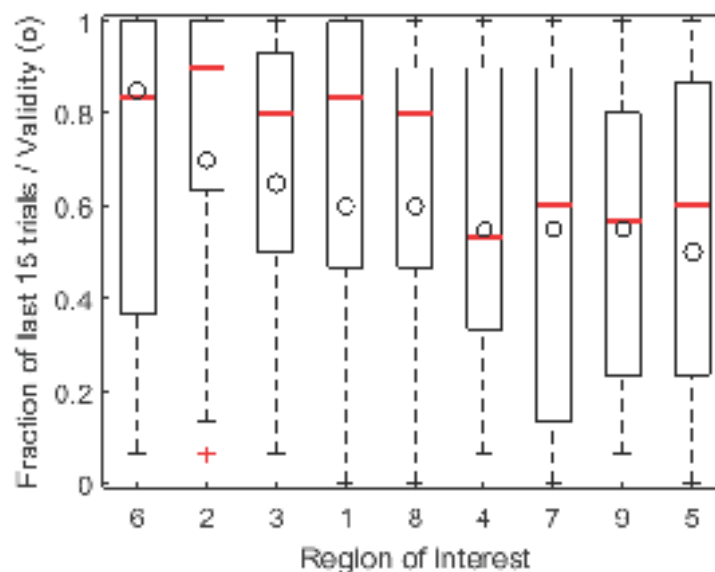


Figure 25: Frequency of cue use ordered by cue validity.

Region of interest (ROI) 6 had the highest cue validity, ROI 2 the next highest validity etc. Box plots represent frequencies and circles represent cue validities.

5 Experiments on Decision Making using Different User Interface Designs: Traffic Management

5.1 Introduction

In this section, two experiments conducted under the scope of the traffic management use-case are presented. Some of the questions we were looking to answer are:

- a) What is the information sampling strategy that humans apply and how does that affect their performance?
- b) Is it safe to assume that information availability is enough to ensure correct responses?
- c) How well do humans perform at spotting automation failure and what could influence their accuracy in spotting those failures?
- d) Does ‘en detail’ reporting lead to better performance in spotting failures?

Experiment 5.2.2 looks at questions a) and b), while the second study concerns questions c) and d). Both experiments use an adaptation of a real-life Traffic Management task. Of course, some simplifications had to be made in order to be able to reduce the reliance on subject-matter knowledge, however task fidelity has been kept. The experiments implement the ramp rate (rate of change of traffic lights on inbound ramps) management scenario that had been developed for the first prototype of the SPEEDD architecture (see Figure 23). The simplifications previously mentioned consist of: 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.

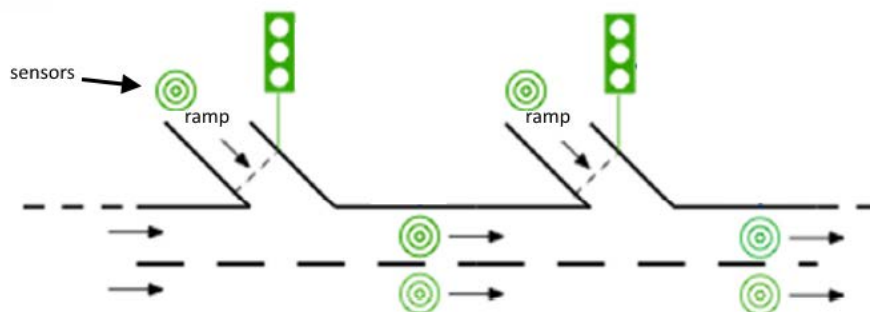


Figure: Experiment Scenario (adapted from [Deliverable 6.1, Figure 2.11])

5.2 Study 5.2.2

5.2.1 Experimental Design

This first experiment was designed to explore how users might respond to automation failure in a traffic management scenario. The two types of automation failure that have been considered are given by either an incorrect suggestion of the recommender system or by an incongruence (disagreement) in the information displayed by the different panels (ROIs or information sources). The first failure is intended to simulate situations where the automated system produce a suggestion based on incorrect or incomplete data; the second failure is intended to simulated situations where a sensor is lost or its ID erroneously replaced by that of another due to communications problems.

5.2.1.1 Scenario

As the automated system computes recommendations based on sensors placed along the road network, it relies on these sensors functioning correctly. In case a sensor failure occurs, it either stops functioning or it produces erroneous data, based on which the computer (recommender system) might come to a wrong suggestion. Furthermore, due to the high complexity of the architecture, timing issues may occur as a result of unexpected bottlenecks in the system. This could lead to a suggestion that, even though correct, has been computed on redundant data. A failure such as this could affect only a small number of system modules and not all. Therefore, the user will need to monitor for both types of failure in order to arrive to a correct decision.

5.2.1.2 Participants

For this experiment, a total of 30 participants were recruited: three traffic management experts from the DIRCE Grenoble and 27 University of Birmingham students. For a subset (18) of the student participants, eye-tracking data was recorded. All participants gave their consent prior to taking part in the study.

5.2.2 Method

A user interface has been designed in JavaScript for the purpose of this experiment (Figure 26). It consists of 4 information sources (panels, or ROIs – regions of interest). ROI 1 is comprised of the computer suggestion and an indication of which ramp the suggestion refers to. This panel is also the place where the user has to make the decision whether to “Challenge” (reject) or “Accept” the displayed computer suggestion. This is done by clicking the “Challenge” or “Accept” button, respectively. The second panel consists of a map on top of which the ramp controller that it refers to is overlaid. The third panel contains a display of sensor data in a graphical form. It also indicates which ramp controller it refers to. The fourth and final panel consists of a list of all the ramp controllers (traffic lights) available. The selection of a specific controller is indicated by a black border around it.

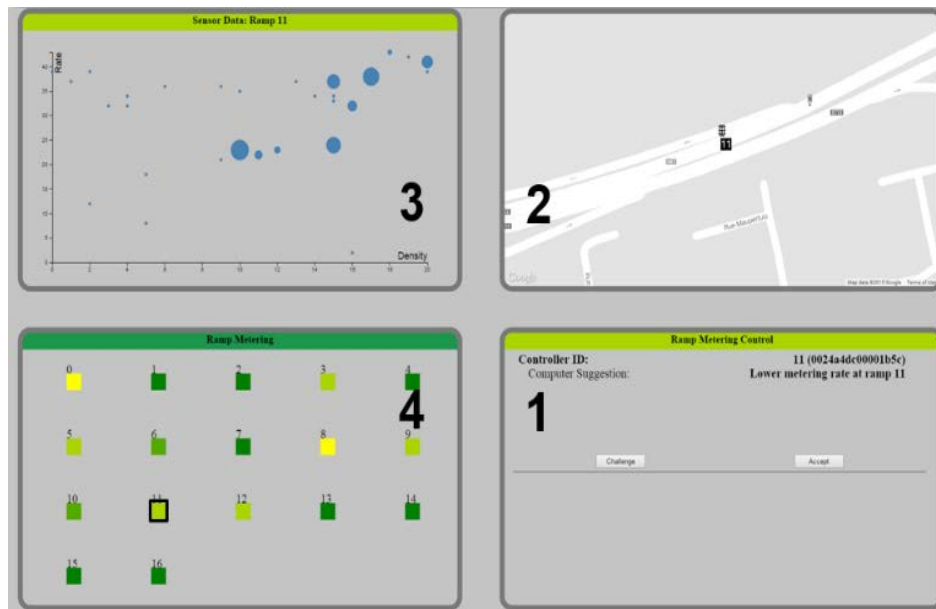


Figure 26: User Interface for the Traffic Management Experiment

Two automation failures have been introduced – erroneous computer suggestion and information source incongruence (as described in the section on scenario definition). In order to determine whether the computer suggestion is correct the user had to look at the sensor data (ROI 3). The information source incongruence could be diagnosed by checking all the four panels for the controller they each refer to. The participants were instructed to accept the computer suggestion if and only if it was correct based on the sensor data graph and the information sources were referring to the same ramp controller.

The experiment consisted of 32 events. In each event the user had to determine whether to accept or reject the computer suggestion depending on the presence or absence of automation failures. The events were split into four categories, each comprising of 8 events:

- 1) information sources agree and computer suggestion correct (TT)
- 2) information sources agree and computer suggestion incorrect (TF)
- 3) information sources disagree and computer suggestion correct (FT), and
- 4) information sources disagree and computer suggestion incorrect (FF).

5.2.2.1 Setup and Data Acquisition

The UI was displayed on a 22'' monitor (1080p resolution). For all participants, decision time data (time to participant response) was recorder in a CSV (comma-separated variable) file. For a subset of 18 participants, eye-tracking data was also gathered. Everything was stored on University of Birmingham servers.

After the participants were given the instructions for the experiment, they were give 2 practice trials after

which they were able to ask any clarifying questions. In between each of the 32 trials, the participants were shown a white screen with a real-time clock. They could proceed to the following trials on their own accord by pressing on the timer.

5.2.3 Results

For the entire cohort of participants we looked at decision times in terms of response they gave (accept or challenge), the experimental design categories (TT, TF, etc. – explained previously) and in terms of the signal detection categories (true positive (TP), false positive (FP), false negative (FN) and true negative (TN)) (see figure 27). A Kruskal-Wallis test was performed on these decision time data, but the results showed no significant difference between the user response types ($p = 0.931$), or between event categories ($p = 0.674$).

Based on decision time performance of the expert participants, the student participants were split into a “high-performing” and a “low-performing” group at the 95% mark. The “high-performing” group consisted of 5 students, while the “low-performing” group was formed of 22 students. Due to calibration issues, eye-tracking data was gathered from 4 participants in the former group and from 14 in the latter.

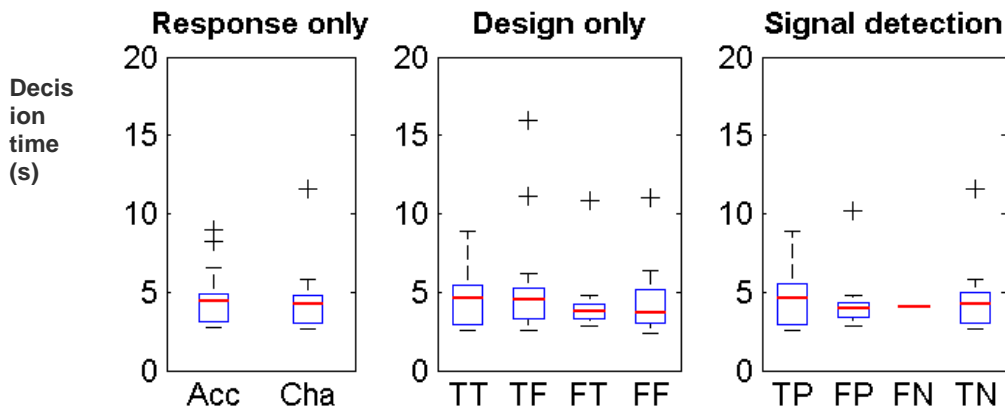


Figure 27: Boxplots for decision times for the different response categories

The performance data illustrated in Figure 28 reveals that, while both groups performed similarly in the TT, TF and FF scenarios, there is a clear difference in the case of FT. The “low-performing” student group was either not able to stop the information source incongruence, or to identify it as a failure. The group scored a mean of 1.84 correct responses ($\sigma = 2.4$), while all students in the “high-performing” group successfully spotted all instances (8/8) of information source incongruence.

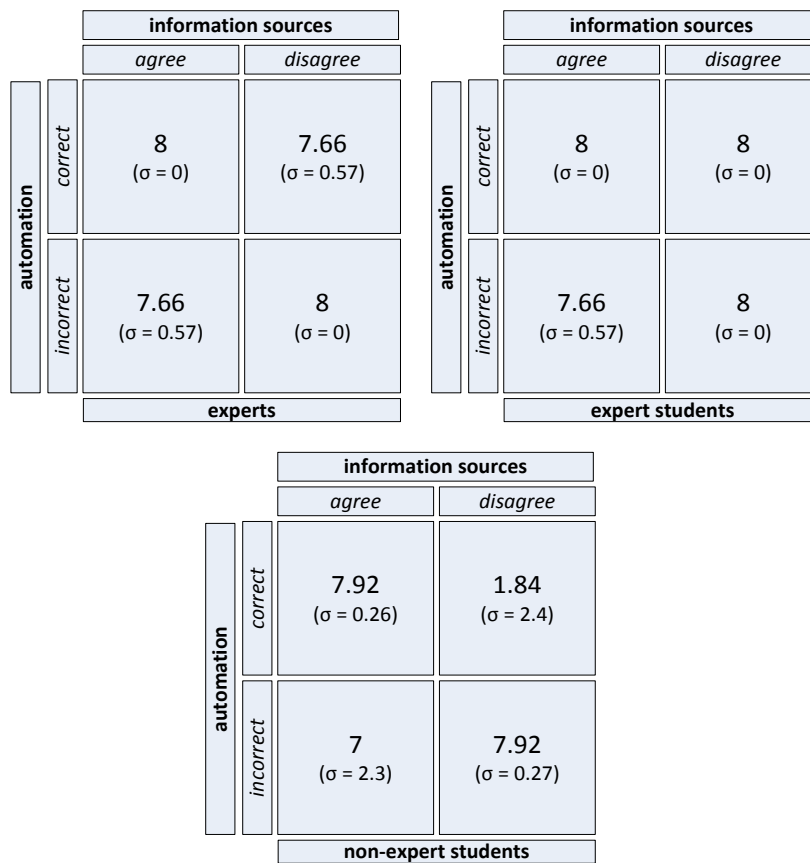


Figure 28: Mean correct responses in terms of scenario for each group

In terms of the eye-tracking data, we looked at four metrics: dwell times per region of interest (figure 29a), number of attended panels (ROIs) per response category (figure 29b), gaze switch count per response category (figure 29c) and percentage view time per region of interest (figure 30). No significant differences could be observed in the first three cases, however, an interesting effect can be noticed from the graph of % View time per region of interest (Figure 30). The “low-performing” student group spent more time looking at ROIs 1 and 3 (~40%) than ROIs 2 and 4 (~10%), while the “high-performing” group spent a similar amount of time looking at ROIs 2, 3 and 4 (10-15%) and a larger amount of time at ROI 1 (~55%).

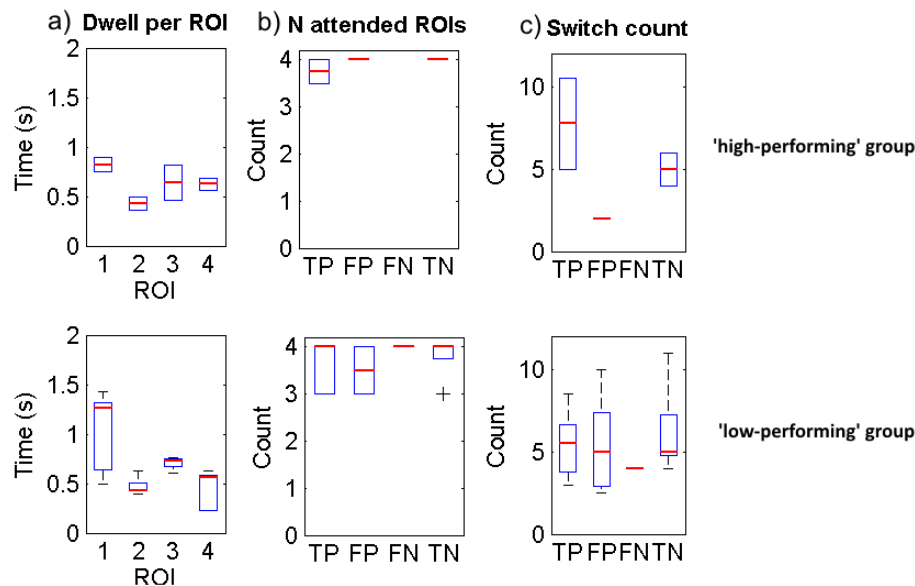


Figure 29: Boxplots for decision times different response categories

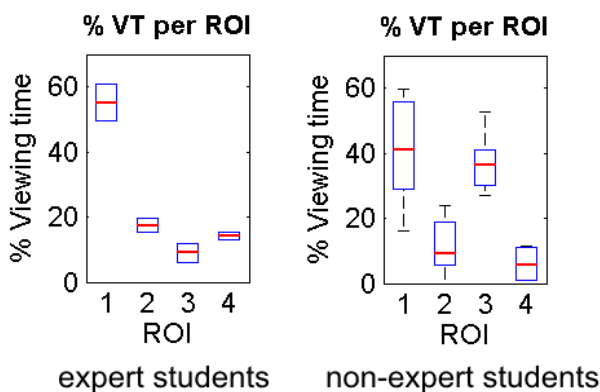


Figure 30: Percentage view time per region of interest (ROI)

5.2.4 Conclusions to Experiment 5.2

This experiment shows that Subject Matter Experts perceive the experimental task differently to the non-experts. Experts are able to correctly diagnose both automation failure types, while non-experts seem to miss the information source incongruence situation. The eye-tracking data (figure 28) suggest that the non-expert group fail to check for this error type, considering the time they spend on the information sources relevant to this task.

This finding poses an interesting question relating to the Proximity Compatibility Principle (PCP) (Carswell and Wickens, 1987; Wickens and Carswell, 1995). This experiment shows that even though a display is constructed in such a way that its elements 'look' to be correctly placed in terms of spatial

proximity, without an understanding of the task demands, they might lack task proximity. In other words, factors that might influence the process of determining the display's 'worth' or diagnosticity is not simply a response to the content of the displayed information. This underlines the need for some sort of highlighting of 'relevant' (however relevance may be defined) information sources. In terms of design, this could be achieved by an increase in saliency of the information source (panel) such that it becomes more 'visible' (it demands more attention), or by integration of the information into another, more salient, information source (panel). In terms of training, this could be achieved by ensuring that task-relevant knowledge is exercised by participants. Another consideration could be that, considering that the majority of participants (~80%) missed the incongruence of information sources, the diagnosis failure in this experiment is due to the response paradigm that the experimental design assumes. The interface is designed in such a way that the user is required to respond to a computer suggestion rather than to a situation (system state). It could be that, because this paradigm is enforced through the user interface, people perceive their task to involve mainly the validation of automation suggestions, leading to, in some cases, to a complete disregard of possibility of incongruence.

5.3 Study 5.2.3

5.3.1 Experimental Design

During discussion with Subject Matter Experts in the evaluation of prototype 1 (Deliverable 7.2), it became apparent that the need to log the decisions made by the Operator was an important consideration in Traffic Management. We considered whether reporting would have an impact on performance. The impact could be negative, i.e., an increase in performance time, or could be positive, i.e., improved decision making. An experiment was designed to explore this.

5.3.1 Scenario

A traffic management task was performed under different conditions of task demand and automated support. The task demands were either make a decision (figure 31) or complete a form and make a decision (figure 32). The automated support would either appear before the user response, i.e., the computer suggestion field would be filled in before the user made a response, or 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. Finally, system reliability was either low (25% correct) or high (80%) correct. Reliability was defined by two factors: (i.) whether the identity of the 'ramp' was the same in all windows (to simulate a failure 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 (i.e., by selecting the 'error' decision option).

5.3.2 Participants

In this study, 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. Participation was for course credit. Participation was through a purpose-built web interface.

5.3.3 Method

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

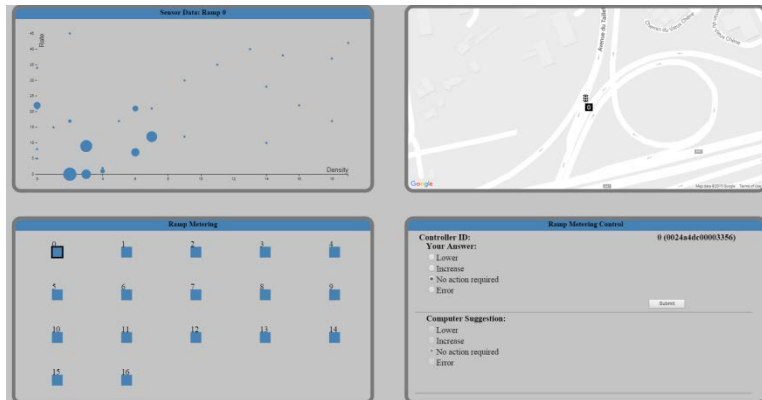


Figure 31: User Interface for Decision Only condition

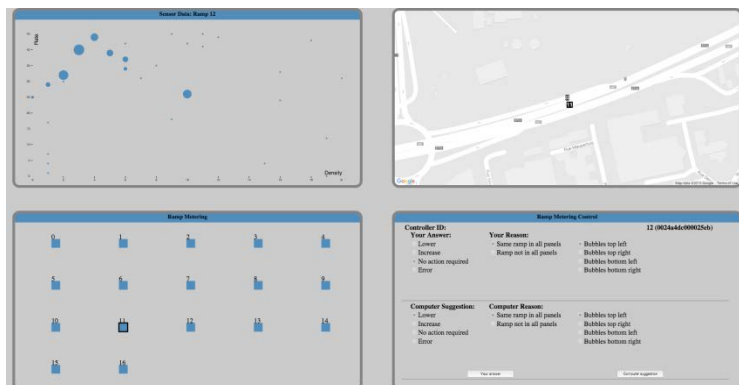


Figure 32: User Interface for Complete Form and Make Decision condition

5.3.4 Results

The performance of the participants was analysed using repeated measures Analysis of Variance (ANOVA) for decision time, correct responses and solution source (i.e., did the participant believe that the solution came from them or from the computer).

5.3.4.1 Decision Time

The time from the presentation of the problem to the participant hitting the 'submit' button was defined as the decision time for each problem. A 4-way ANOVA was performed (Reliability (Low vs. High); Task (Decision only vs. Form-filling plus decision); Solution turn (user vs. computer); Match (information in displays and / or correct solution)).

For the Match variable, there are two letters. The first letter describes the match between information cues (T = all cues match, F = cues do not match) and the second letter describes the correctness of the computer solution (T = true, correct and F = false, incorrect).

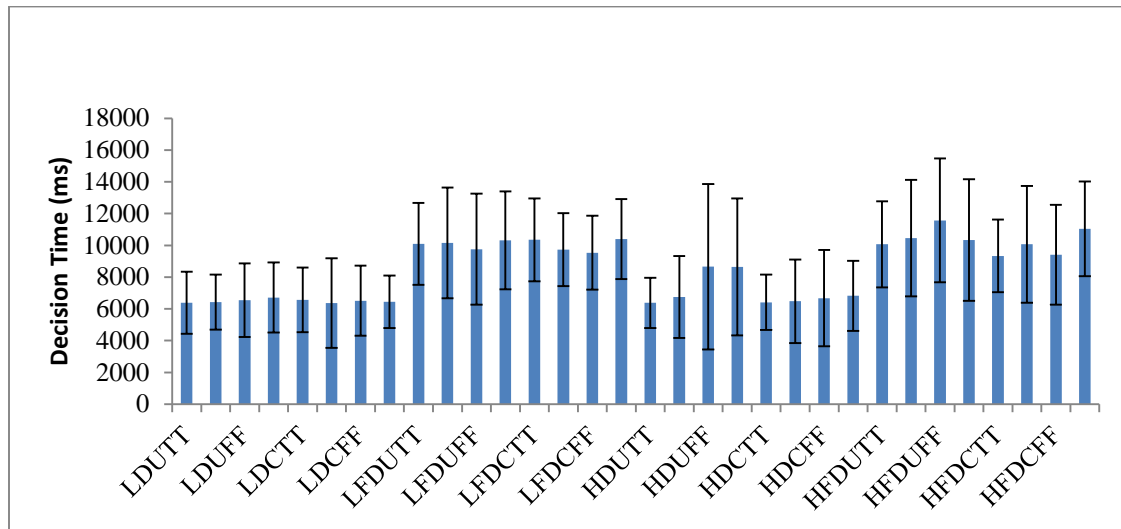


Figure 33: Mean Decision Time across all conditions

ANOVA revealed no effect of Reliability of decision time [$F(1,19) = 0.41$, $p = 0.53$]. There was a significant main effect of task [$F(1,19) = 391.466$, $p < 0.0001$] and of match [$F(3,57) = 2.77$, $P < 0.05$]. Given the lack of 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. For Low reliability conditions, there was significant main effect of task [$F(1,19) = 137.238$, $p < 0.0001$] but no other effects. For the High reliability condition, there were significant main effects of task [$F(1,19) = 82.452$, $p < 0.0001$], turn [$F(1,19) = 4.304$, $p < 0.05$] and match [$F(3,57) = 2.678$, $p < 0.05$] but no interaction effects.

Figure 33 suggests that, for the Low reliability conditions, decision time varies between tasks but is not affected by turn or match. For the High reliability conditions, in addition to task, decision time seems to vary with match when the User responds first and when the match is true. That is, for the Decision Only, HDUTT and HDUTF are faster than HDUFF and HDUFT but this is not apparent when the Computer responds first (all decision times look similar). When the task involves form-filling as well as decision making, there is a less obvious effect (although it does look as if time is slightly faster when the Computer responds first and there is some effect of match)

5.3.4.2 Percentage Correct Response

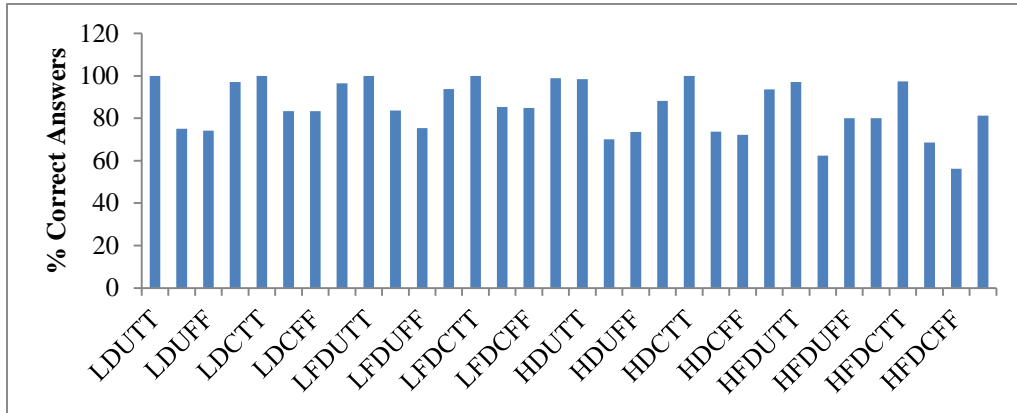


Figure 34: Mean Correct Response for conditions

ANOVA revealed a significant main effect of match [$F(3,57) = 10.032$, $p < 0.0001$] but no other effects. There was a significant interaction between reliability x turn x match [$F(3,57) = 3.787$, $p < 0.05$]. This suggests a complex relationship between the performance of the computer system and the participant's ability to correctly define a solution. For the Low reliability condition, there was only a main effect of match [$F(3,57) = 7.812$, $p < 0.0001$]. However, for the High reliability condition, there were main effects of task [$F(1,19) = 9.333$, $p < 0.05$] and match [$F(3,57) = 7.7777$, $p < 0.0001$] and an interaction between turn and match [$F(3,57) = 3.147$, $p < 0.05$]. Figure 34 suggests that participants were more strongly affected by errors in match in the High reliability condition.

5.3.4.3 Solution Source

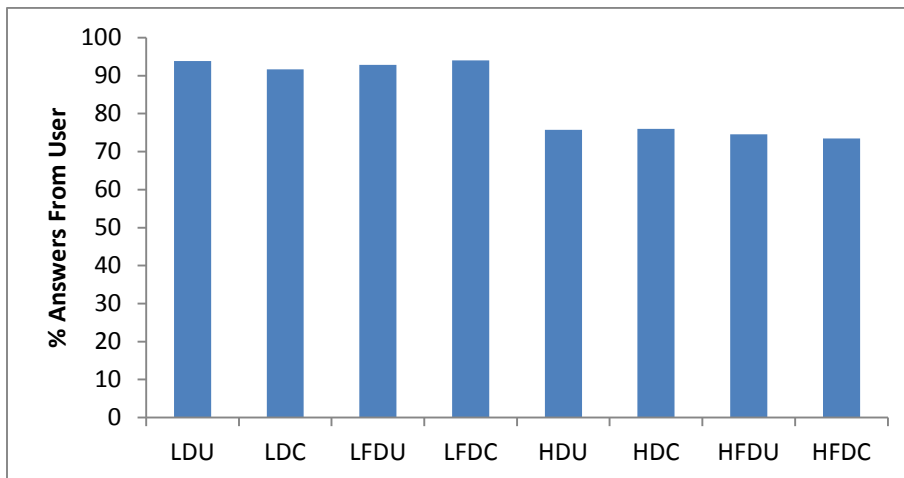


Figure 35: Comparison of Solution Source

There was a main effect of reliability [$F(1,19) = 5.459$, $p < 0.05$]. As figure 35 shows, participants were

more likely to select ‘Self’ (than Computer) as the source of the solution on the Low reliability condition, but this was lower in the High reliability condition.

5.3.5 Conclusions

While experiment 5.2.3 shows that completing a form in addition to making a decision incurs a time cost, there are some less obvious findings here. First, the reliability of the automated system has an impact on user activity. When the system has low (25%) reliability, then users are likely to rely on their own interpretation on the system state (and so, prefer to have themselves as the solution source). When the system has higher (80%) reliability, then users will accept advice from the computer (and so, see the computer as a viable solution source on some of the trials). In cases where the system has high reliability and the computer suggestion is presented first, decision time tends to be constant – as if the users are monitoring system recommendation and then accepting this. Thus, there is a significant effect of turn, but only in High reliability conditions. Furthermore, the impact of an incorrect match only had an impact on time in the High reliability conditions – as if the users take time to reconcile errors on the display when they are assuming a reliable system. Second, the reliability of the system has an impact on correctness of response. In both reliability conditions, user performance seems to follow the correctness of matching on the display, i.e., where there is a mismatch, then correctness drops. However, this impact seems to be more noticeable in the High reliability condition, i.e., when the mismatch occurs, users might be less likely to notice this.

Overall, this suggests that system reliability has little impact on the time taken to perform a task but does impact on the likelihood that users will accept computer advice (and thus be less likely to see themselves as the only solution source) and does impact on the likelihood of errors persisting. In other words, if users regard the system as having High reliability, they might be less likely to intervene when the system has made an error. What is particularly noteworthy in this experiment is that participants had to check the content of the displays in order to determine whether the computer was likely to be correct (i.e., Tx). If we combine the Match data across all conditions, we can see that participants tend to be faster for the TT situation than for the others. The highest time comes under FT – this is where the information sources on the display do not match but the computer suggestion is correct. Presumably this leads to a level of incongruity that requires the participant to decide whether they trust the solution even if there is an error on the display. When the Computer solution is incorrect (TF and FF) then decision time is somewhere between the two extremes.

Table 6: Comparison of Average Times across different forms of Match under all conditions

TT	TF	FF	FT
8196.047	8302.513	8580.865	8836.837

5.4 Discussion

Considering the questions formulated in the beginning of this chapter the two experiments presented offer clarifications:

- a) Experiment 5.2.2 shows that ~80% of the (non-expert) participants failed to spot automation failure illustrated through the incongruence of information sources. The ‘low-performers’ in this study seem to have adopted a sampling strategy that differed from the one adopted by ‘high-performers’ in terms of the proportion of time accorded to information sources relevant to spotting incongruence. This difference in sampling strategy lead to a significant difference in decision accuracy between the two groups.
- b) Experiment 5.2.2 also shows that information availability does not guarantee information usage. This is to say that just because the information is there it doesn’t ensure that it will be taken into account. The design of the user interface ‘made sure’ that in terms of information placement (task and proximity compatibility) the information sources are presented equally. The hypothesis was that this would lead to participants equally ‘weighing’ the different information sources. However we have seen that this is not the case, and some other factors might influence the ‘perceived importance’ of information sources (see Discussions of this study).
- c) Experiment 5.2.3 shows that changing the response paradigm can influence human decision accuracy. In this study participants had to give their own answer based on a system state that could be determined from the information sources shown. Even though the same information sources were available to the participants in the first experiment, the difference comes from the fact that, here, the participant’s response is not triggered by a computer suggestion and the decision is not in relation to the computer suggestion, but rather to the issue at hand. While in the first experiment we saw that for ~80% of participants performances were around 65%, here 10% of participants show performances lower than 80%.
- d) Experiment 5.2.3 also shows that reporting does not lead to a higher decision accuracy, while at the same time increasing the time cost of making decisions. Subsequent discussions with some participants have led us to believe that humans might perceive the task of ‘form filling’ or reporting as being different unrelated to the main task decision-making. This would mean that the reporting task is done in an automatic manner and the time spent for performing it does not contribute to decision-making. This hypothesis would explain the fact that ‘reporting’ did not result in a difference in performance.

6 Defining Performance Baselines

6.1 Introduction

A core Human Factors problem for the SPEEDD project arises from the need to quantify change in operator performance with the introduction of the proactive, event-driven decision support that the project is developing. The scale of this problem relates to two issues. The first is that the scenarios which are being developed in the use cases do not necessarily reflect current activity of operators. For example, the prototype for Road Traffic Management involved ramp metering, but the Grenoble operators were not involved in ramp metering in their day-to-day activity. Having said this, Grenoble has been investing in new systems to monitoring ramps onto the ring road and this will eventually result in the DIR CE control room operators having automated assistance in managing traffic lights to control flow through these ramps. A visit back to the control room is planned to discuss these changes. In the Credit Card Fraud use case, the role of the analyst will vary between organisations and between stages of the fraud investigation process (as detailed in D7.3). However, we believe that we have a reasonable description of this role and the model (chapter 3) and experiment 5.2.1 (chapter 4) illustrate the manner in which fraud analysis will draw on disparate pieces of information to reach a decision. The second, and more significant issue, arises from the fact that introducing new technology inevitably alters the nature of the task being performed. Consequently, changes in performance *could* be a result of the technology but could equally well arise from adaptations of the person in response to either the technology or to changes in the nature of the job that occurred in parallel with the introduction of the technology. In response to both of these issues, we feel that it is essential to identify invariant aspect of the work that would apply across different forms of technology. In the language of Cognitive Work Analysis, these invariant aspects are the Values and Priorities of the human-automation system. The question then becomes are matter of defining these and, more importantly, understanding the strategies that operators apply in order to fulfil these purposes.

6.2 Can time be a useful measure of baseline performance?

A great deal of Human-Computer Evaluation (HCI) evaluation relies of the timing of human activity to provide a measure of system performance. An assumption is that faster performance is often desirable and that redesigning a user interface to result in reduced performance time is an improvement on the previous design. During the 1990s, it was fashionable to develop time-based models to make predictions about potential improvements in performance time. These models took the Keystroke-Level Models (KLM) that were popular in the 1980s (e.g., Card et al., 1983) and recast these to recognize the fact that human activity contained a degree of parallel activity, using the concept of Critical Path Analysis (John and Kieras, 1996; Olson and Olson, 1990; Baber and Mellor, 2001; Stanton and Baber, 2008).

6.2.1 A Critical Path Model of Experiment 5.2.3

Given the two UIs from experiment 5.2.3, a reasonable question would be to ask whether it is possible to predict the performance time that someone using these might be expected to achieve. One benefit of such a prediction could be to provide a comparison of the UIs, while another could be to define a target for performance time (say, as an indication of whether training and practice has allowed an operator to reach

criterion). Recall that the motivation for experiment 5.2.3 was to ask whether the task of ‘reporting’ (which the Subject Matter Experts in DIR CE identified as core to their work in the evaluation of initial prototypes in D8.3) had a significant bearing on performance, and whether removing the need to report would free-up time. If time was freed up (with the removal of reporting), then the obvious question would be to what benefit this ‘free’ time might be put?

In order to develop a Critical Path Model (CPM) of experiment 5.2.3, we begin with a task analysis of the actions that can be performed using the UI. This involves defining tasks at the level of moving the cursor from one Region of Interest to another, clicking the button on the mouse, moving the eye from one Region of Interest to another, reading or checking the display content at a specific Region of Interest, and making decisions or choices. In order to develop the model, we create a sequence of tasks that could reflect interaction during completion of the activity. It is important to note that this sequence is defined solely in terms of plausible task ordering and no claims can be made as to whether this is the best (or even the only) way of performing this activity. We will return to this point in the next section, but for now figure 31 gives an indication of how the model is constructed.

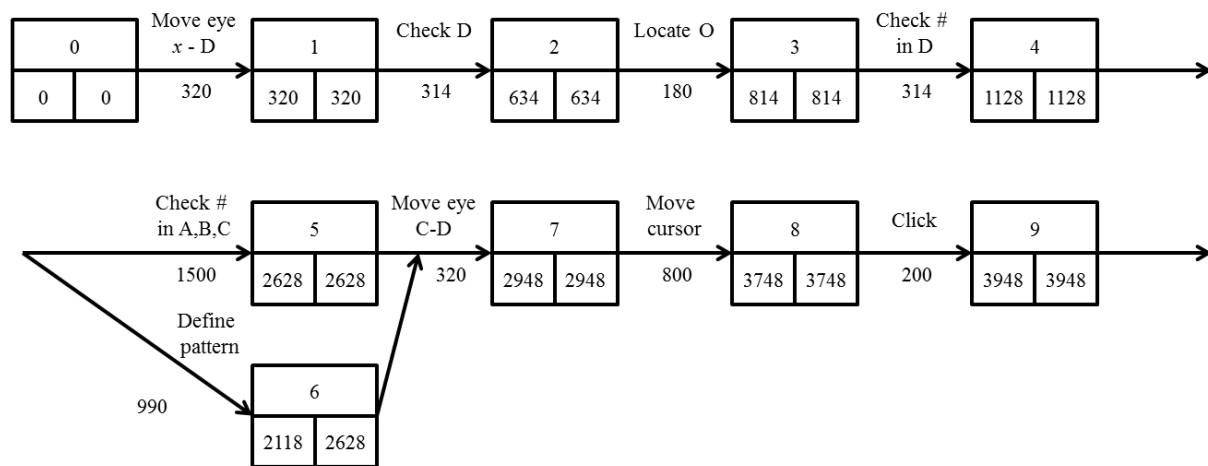


Figure 36: Extract from CPM for Experiment 5.2.3

In each cell in figure 36, the top number indicates the cell number in the sequence, the bottom left number indicates the earliest start time and the bottom right indicates the latest finish time; between each cell there is an arrow labelled with the task that is performed in order to make a transition from one cell to the next, and the standard time (in milliseconds) for that task. For our modelling, we prefer to take these standard times from published sources. The CPM is constructed by summing the standard times from left to right; if there is more than one task feeding into a cell, then one uses the highest time. Once this forward pass has been completed, a backward pass is performed in which one subtracts the standard times from right to left; where there is a more than one time, one takes the lowest time. In this manner, cells with multiple inputs are likely to have differences between earliest start and latest finish times. This difference is known as the float and indicates a period of time during which the preceding task could be performed.

One advantage of this modelling is that it is a simple matter to modify aspects in order to compare different UI designs or task sequences. For the purposes of this exercise, we modelled the two UIs for experiment 5.2.3 (i.e., decision with and without reporting) under two conditions – when the user had to check all of the data (low reliability) and when the user could rely on the system to do the checking (high reliability). The results of the models are compared with the results from the experiment in table 7. In these models, we assume that the radio buttons (in the UI) all need to be changed to produce the correct response and this is then submitted. We assume that the operator makes no mistakes and responds to all information correctly.

Table 7: Comparison of Predicted and Observed times for Experiment 5.3

	Total time	Visual inspection	Decision	Manual input
Model A: ramp numbers do not agree; select ‘error’; select ‘submit’				
	4860ms	50%	17%	33%
Model B: complete full report; update all entries; select ‘submit’				
Experiment (low)	10040ms			
Model (low)	10108ms	52%	24%	24%
Experiment (high)	10284ms			
Model (high)	8250ms	35%	21%	44%
Model C: no report; update all entries; select ‘submit’				
Experiment (low)	6492ms			
Model (low)	6252ms	44%	24%	32%
Experiment (high)	7100ms			
Model (high)	3638ms	35%	21%	44%

As table 7 illustrates, the predicted times accord quite well with the average times observed in the experiment. While this might imply a degree of ‘accuracy’ in the modelling, the intention is not to provide ‘correct’ predictions of times so much as to highlight discrepancies between actual and assumed (modelled) performance. There are three key points to note from table 8. First, when there is low reliability, one can assume that the user will ignore system recommendation and check information, the model’s predictions concord with the data from the experiment. Second, when the system has high reliability, then there are discrepancies between observed and predicted times. Third, the proportion of time devoted in each model to the activities of visual inspection, decision and movement show some variation in the models, i.e., while ‘decision’ activities seem consistent across the model, ‘visual inspection’ and ‘manual input’ show changes. In other words, the models assume that (for high reliability), users will be less likely to look at the screen and more likely to simply click on the radio buttons suggested by the system. While the results from experiment 5.2.3 do support this latter assumption, the differences in time suggest that performance is not as neatly separated as the model assumes.

6.2.2 What’s wrong with Critical Path Models?

A core problem with Critical Path Models lies in the nature of their construction. It is not obvious that the sequence of tasks which the analyst has assumed either reflects the *actual* sequence followed by individuals, or that this sequence is optimal or that this sequence is unique. Thus, the sequence-as-

modelled reflects the best guess estimate of the analyst for a particular way of performing the task. If the resulting times accord with the observed times (as they do in the models for the low reliability UI) this implies some fit with the data but even this does not guarantee that the fit arises from the accuracy of the model or from some confounding factor. Thus, we can model tasks in terms of time, we can produce estimates of performance which accord with observed performance outcomes and this could provide us with a means of calculating baseline data. However, the fact that the models appear to be subject to variability outside the model parameters suggest that this is quite a crude approach and that it misses the salient aspects of strategy which underpin operator performance. Therefore, we believe that it is necessary to clearly and precisely define strategy in terms of the operator's ability to correctly determine those information sources which are essential to their decision making, and to understand how operators sample from the space of available information sources in order to make these decisions. This is the approach which we follow in chapter 3. It is also the approach which underlies CWA, in terms of appreciating how the values and priorities that constrain operator performance are balanced.

7 Discussion

7.1 Introduction

This report presents an approach to design (with an illustration of the UI design that we are developing on the basis of this approach), in chapter 2. While this approach aligns CWA with information requirements and then with representation, we require a means of evaluating the possible designs. Alternative ways of conducting such evaluation range from canvassing the opinion of end-users (as we did in D7.2 and D8.3), to produce time-based models (chapter 6) to decision models (chapter 3) to conducting controlled user trials (chapters 4 and 5). Of these approaches, the decision models can be applied to concept designs early in the development process, while user trials would be most appropriate when there are functioning prototypes, and end-user opinion would inform design assumptions and decisions throughout the design process.

Comparing the results reported in chapters 3 and 4, one can see that the decision model ranks the validity of the information sources in a manner that is comparable to the users. While the model reaches a ceiling in its accuracy, and the participants in chapter 4 show some variability, the model is able to identify the optimal strategy and we see some of the participants (after the duration of the experiment) are able to reach this level.

The results of experiment 5.2.1 suggest that the decision strategy is closer to Take-The-Best than Weighted-Additive decision models. However, a better explanation might be that the participants are adopting neither of these strategies but, as the model suggests, are selecting cues which appear to provide the most relevant information to a given decision but which involve a finite effort to retrieve. In order words, the participants appear to be sensitive to both the validity of the information sources and also to the cost of accessing them. It can also be inferred that peripheral vision plays an important role in information acquisition, at least in the Full display conditions of experiment 5.2.1.

Experiment 5.2.2 (chapter 5) showed that non-experts missed information source incongruence (even though their eye-tracking data suggests that they fixated on this information). This indicates suboptimal performance, even when sampling seems appropriate. A plausible explanation of this is that the validity of source congruence was not responded to by the non-experts (but it was by the expert participants). Consequently, presentation of information will not guarantee optimal decision response, and the people who use this information need to be able to accommodate the validity of information sources with an understanding of the domain.

Experiment 5.2.3 showed that understanding the domain could be encouraged by simply indicating that the task involved two decisions (one for congruence and one for decision). This led to very few participants missing the incongruence (even though the information sources were the same as in experiment 5.2.2). While the reliability of the automation had little impact on performance time in experiment 5.2.3, there did seem a tendency for participants to accept automated suggestion in the high reliability condition (even when this was wrong). The longest average performance time in this

experiment occurred when the automated suggestion was correct but the displays were incongruent, suggesting this led to the participants being confused by the mismatch.

Experiment 5.2.3 shows, as one would expect, that form-filling simply extends the performance time. However, it was interesting to note that participants tended to regard form-filling as separate from the decision task rather than either integral, or at least related, to it. Thus, despite the observation that indicating that the task involved two decisions improved performance when compared with experiment 5.2.2, participants did not recognize a benefit in this ‘administrative’ task for their decision making. This raises further questions for the Road Traffic Use Case, where the need to maintain records of activity is important to the role of the operators.

The model and the experiments all explore, in different ways, the manner in which people recognize the validity of information sources to the decisions that they seek to make. This suggests that it is important to understand user activity (and hence the performance baselines for the SPEEDD systems) not only in terms of activity (particularly when performance time might be a poor measure of difference between UI) but also in terms of the understanding of the validity of information sources and the trade-off that people make between seeking valid information sources and the (time) cost of making use of these sources. As experiment 5.2.2 showed, when participants were not able to appreciate the validity of specific information sources they made errors in the task, even when their search strategy appeared comparable to the experts. This suggests that the validity of the information is defined not only by the content of the information but also by the decision task and the operator’s knowledge of the domain in which they are working.

From experiment 5.2.1, during debriefing it was apparent that people tell stories to help them appreciate interactions in multiple data sets. From model, key information sources identified and we assume that optimal decision making involves assessing the value of information sources in an active and interactive exploration of the options. This means that, rather than people following a TTB (in which an overarching heuristic dictates the information selection strategy) people revise their estimates of the value of information sources. It might, therefore, be the case that the stories themselves provide useful insights into the experience of this adaptation to, and optimization of, the selection of information sources in support of decision making. In other words, rather than being something peripheral to the decision task or something performed in parallel with the optimization, the stories are a symbiotic means of both guiding the underlying reinforcement learning process (through which the value of information sources are estimated and applied) and an awareness of the development of the unfolding optimization (through which those information sources with highest values become more likely to influence the developing stories). From a sensemaking perspective, this suggests that ‘value’ is partly a verbalisable account of the decision process and partly the crystallization of the importance of information sources to specific decisions. From a reinforcement learning perspective, this suggests that ‘stories’ is partly a definition of hyper-parameters which constrain the optimization processes and partly a meta-awareness (in the decision maker) of the defining of these parameters.

8. References

Baber, C., Attfield, S., Conway, G., Rooney, C., and Kodagoda, N. (2016) Collaborative sense-making during simulated intelligence analysis exercises. *International Journal of Human-Computer Studies*, 86, 94-108.

Baber, C., and Mellor, B. (2001). Using critical path analysis to model multimodal human-computer interaction. *International Journal of Human-Computer Studies*, 54(4), 613-636.

Baron, S., and Kleinman, D. L. The human as an optimal controller and information processor. *Man-Machine Systems, IEEE Transactions on* 10, 1 (1969), 9-17.

Bennett, K. B., and Flach, J. M. (2011). *Display and interface design: Subtle science, exact art*. Boca Raton, FL: CRC Press. Carswell, C. M. (1992). Choosing specifiers: An evaluation of the basic task model of graphical perception. *Human Factors*, 34, 535-554.

Bertin, J. (1983) *Semiology of Graphics*, Madison, WN: University of Wisconsin Press.

Bisantz, A.,M., Roth, E.M., Brickman, B., Gosbee, L.L., Hettinger, L. and McKinney, J. (2003) Integrating cognitive work analysis in a large-scale system design process, *International Journal of Human-Computer Studies*, 58, 177-206.

Bröder, A., and Gaissmaier, W. *Heuristics: The foundations of adaptive behaviour*. Oxford University Press, 2011.

Bröder, A., and Schiffer, S. Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32, 4 (2006), 904.

Canudas de Wit, C., Ojeda, L.R.L. and Kibangou, A.Y. (2012) Graph constrained CTM observer design for the Grenoble south ring, *13th IFAC Symposium on Control in Transportation Systems*, Sofia, Bulgaria

Carswell, C. M., and Wickens, C. D. (1987) Information integration and the object display: An interaction of task demands and display superiority. *Ergonomics*, 30, 511-527.

Chen, X., Bailly, G., Brumby, D. P., Oulasvirta, A., and Howes, A. (2015) The emergence of interactive behaviour: A model of rational menu search. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM , 4217-4226.

Chen, X., Lewis, R., Myers, C., Houpt, J., and Howes, A. (2013) Discovering computationally rational eye movements in the distractor ratio task. In *Reinforcement Learning and Decision Making*.

- Geisler, W. S. (2011) Contributions of ideal observer theory to vision research. *Vision research* 51, 7, 771–781.
- Gigerenzer, G., and Gaissmaier, W. (2011) Heuristic decision making. *Annual review of psychology* 62, 451–482.
- Gigerenzer, G., and Goldstein, D. G. (1996) Reasoning the fast and frugal way: models of bounded rationality. *Psychological review* 103, 4, 650.
- Gigerenzer, G., and Todd, P. M. (1999) *Fast and frugal heuristics: The adaptive toolbox*. Oxford University Press.
- Gordon, J., and Abramov, I. (1977) Color vision in the peripheral retina. ii. hue and saturation. *JOSA* 67, 2, 202–207.
- Hayhoe, M., and Ballard, D. (2014) Modeling task control of eye movements. *Current Biology* 24, 13, R622–R628.
- John, B. E. and Kieras, D. E. (1996). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(4), 320-351
- Jungk, A., Thull, B., Hoeft, A., and Rau, G. (1999). Ergonomic evaluation of an ecological interface and a profilogram display for hemodynamic monitoring. *Journal of clinical monitoring and computing*, 15(7-8), 469-479.
- Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998) Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1, 99–134.
- Kandel, S., Paepcke, A., Hellerstein, J. M., and Heer, J. (2012) Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18, 12, 2917–2926.
- Kieras, D. E., and Hornof, A. J. (2014) Towards accurate and practical predictive models of active-vision-based visual search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 3875–3884.
- Lee, M. D., and Zhang, S. (2012) Evaluating the coherence of take-the-best in structured environments. *Judgment and Decision Making* 7, 4.
- Lewis, R., Howes, A., and Singh, S. (2014) Computational rationality: linking mechanism and behaviour through bounded utility maximization. *Topics in Cognitive Science* 6, 2, 279–311.
- Lintern, G. (2005) Integration of cognitive requirements into system design, *Proceedings of the Human*

Factors and Ergonomics Society 49th Annual Meeting, Santa Monica, CA: HFES, 239-243.

Lintern, G. (2012) Work-focused analysis and design. *Cognition, Technology and Work*, 14, 71-81.

Newell, B. R., Weston, N. J., and Shanks, D. R. (2003) Empirical tests of a fast-and-frugal heuristic: Not everyone “takes-the-best”. *Organizational Behaviour and Human Decision Processes* 91, 1, 82–96.

Newell, B. R., and Shanks, D. R. (2003) Take the best or look at the rest? factors influencing “one-reason” decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29, 1, 53.

Nunez-Varela, J., and Wyatt, J. L. (2013) Models of gaze control for manipulation tasks. *ACM Transactions on Applied Perception (TAP)* 10, 4, 20.

Olson, J. R., and Olson, G. M. (1990). The growth of cognitive modeling in human-computer interaction since GOMS. *Human-computer interaction*, 5(2-3), 221-265.

Paassen, M. M. van. (1995). New visualisation techniques for industrial process control, In *Proceedings of the 6th ifac symposium on analysis, design and evaluation of man-machine systems*. Boston.

Payne, S. J., and Howes, A. (2013) Adaptive interaction: A utility maximization approach to understanding human interaction with technology. *Synthesis Lectures on Human-Centered Informatics* 6, 1, 1–111.

Pirolli, P. and S. Card. (1999) Information foraging. *Psychological Review* 106(4): 643-675.

Rasmussen, J.(1983) Skills, rules and knowledge: signals, signs and symbols, and other distinctions in human performance models, *IEEE Transactions of Systems, Man and Cybernetics*, 13, 257-266.

Rieskamp, J., and Hoffrage, U. (2008) Inferences under time pressure: How opportunity costs affect strategy selection. *Acta psychologica* 127, 2, 258–276.

Rieskamp, J., and Otto, P. E. (2006) Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General* 135, 2, 207.

Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993) The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, ACM, 269–276.

Russell, S., and Subramanian, D. (1995) Provably bounded-optimal agents. *Journal of Artificial Intelligence Research* 2, 575–609.

Sanders, M. S., and McCormick, E. J. (1987) *Human factors in engineering and design* . McGRAW-HILL book company.

- Simons, D.J., Levin, D.T. (1997). Change blindness. *Trends Cogn. Sci.* 1, 261–267
- Sprague, N., Ballard, D., and Robinson, A. (2007) Modeling embodied visual behaviours. *ACM Transactions on Applied Perception (TAP)* 4, 2, 11.
- Stanton, N. A., and Baber, C. (2008). Modelling of human alarm handling response times: a case study of the Ladbroke Grove rail accident in the UK. *Ergonomics*, 51(4), 423-440.
- Stevens, S.S. (1946) On the theory of scales and measurement, *Science*, 103, 677-680.
- Sutton, R. S., and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 1998. [35]
- Trommershäuser, J., Glimcher, P. W., and Gegenfurtner, K. R. (2009) Visual processing, learning and feedback in the primate eye movement system. *Trends in Neurosciences* 32, 11, 583–590.
- Tseng, Y.-C., and Howes, A. (2015) The adaptation of visual search to utility, ecology and design. *International Journal of Human-Computer Studies* 80, 45–55.
- Upton, C. and Doherty, G. (2008) Extending ecological interface design principles: a manufacturing case study, *International Journal of Human Computer Studies*, 66, 271-286.
- Wickens, C. D. and Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design, *Human Factors*, 37, 473-494.
- Zhang, J. (1997). Distributed representation as a principle for the analysis of cockpit information displays, *International Journal of Aviation Psychology*, 7 105-121
- Zhang, J. and Norman, D.A. (1994). Representations in Distributed Cognitive tasks, *Cognitive Science* 18 87-122
- Zhang, J.) Norman, D.A. (1995). A representational analysis of numeration systems, *Cognition* 57 271-295