2018-9

Analysing online user activity to implicitly infer the mental workload of web-based tasks using defeasible reasoning

Paul Mara
Technological University Dublin

Analysing online user activity to implicitly infer the mental workload of web-based tasks using defeasible reasoning

Paul Mara

D16124441

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Advanced Software Development)

2018

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Advanced Software Development), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: _____

Date: **16 September 2018**

ABSTRACT

Mental workload can be considered the amount of cognitive load or effort used over time to complete a task in a complex system. Determining the limits of mental workload can assist in optimising designs and identify if user performance is affected by that design. Mental workload has also been presented as a defeasible concept, where one reason can defeat another and a 5-layer schema to represent domain knowledge to infer mental workload using defeasible reasoning has compared favourably to state-of-the-art inference techniques. Other previous work investigated using records of user activity for measuring mental workload at scale using web-based tasks

For this research, a solution design and experiment were put together to analyse user activity from a web-based task to determine if mental workload can be inferred implicitly using defeasible reasoning. While there was one promising result, only weak correlation between inferred values and reference workload profile values was found.

Key words: *Mental Workload; Defeasible Reasoning; User Interaction; Cognitive Load; Human-Computer Interaction*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my research supervisor, Dr Luca Longo, for both his guidance and teaching. Thanks to Lucas Rizzo who shared his knowledge and expertise. Thanks to Joaquim Filipe Romero who undertook the experiments which resulted in the creation of the dataset used in this research.

I wish to thank all the staff of DIT, particularly those in the School of Computing and especially all involved in the MSc Computing.

I also wish to thank all my work colleagues for their interest and assistance in my studies.

I would like to thank my family and friends for all their good will.

Finally, I would like to thank my parents. Thanks Mam and Dad!

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1 INTRODUCTION

1.1 Background

The adoption of computing is now widespread throughout society and there is a focus on moving computers into the background while human behaviour in the field of Human-Computer Interaction (HCI) has gained foreground attention (Pantic, Nijholt, Pentland, & Huanag, 2008). A significant body of research exists in the field of HCI with key areas being investigated including the improvement of interface design, user experience and the communication between humans and computers (Karray, Alemzadeh, Abou Saleh, & Nours Arab, 2008; Hartmann, Sutcliffe, & Angeli, 2008; Chao, 2009; Luca Longo, 2015b; Hornbæk & Hertzum, 2017). The optimisation of human performance as they interact with computing interfaces is essential and MWL is used for this.

The concept of Mental workload (MWL) has gained importance in HCI as use of the World Wide Web has grown (P. A. Hancock & Caird, 1993; Wästlund, Norlander, & Archer, 2008; Luca Longo, 2011; L. Longo & Dondio, 2015). Measuring and determining the limits of MWL can assist in optimising designs for human performance and to identify if a design is possibly causing them confusion, frustration or an increase in mistakes made (Loft, Sanderson, Neal, & Mooij, 2007; Gwizdka, 2010; Luca Longo, 2011; L. Longo, 2015; Luca Longo, 2016b, 2017). It can also assist in the assessing, prediction, design and operation of tasks and interfaces to minimise the amount of human information processing needed.

Interface designs can fail if they include tasks for humans which require high cognitive load such as relying on users learning many complex commands too quickly or needing them to remember too much (Balogh, Cohen, & Giangola, 2004). These human limits need to be respected and multiple methods of measuring cognitive load and MWL (which can be considered to be the amount of cognitive load or effort used over time to complete a task in a complex system) have been developed to help in resolving these issues (Xie & Salvendy, 2000; Tracy & Albers, 2006; Luca Longo, 2016b; Moustafa, Luz, & Longo, 2017; Contreras, 2018).

User interaction tracking is another HCI area shown to be assistive in website analysis, identification of elements raising the cognitive load of websites and understanding web

user behaviour (Atterer, Wnuk, & Schmidt, 2006; Guo & Agichtein, 2012). Tracking has been done using logs of web user interaction indicators such as number of mouse clicks and amount of scrolling. Recently this form of analysis of web user interaction logs has shown to be of potential assistance in the assessment of MWL in the context of website design (Romero, 2017). Individual indicators were found to correlate to MWL and further investigation on whether the value of one indicator is related with the value of two or more indicators was recommended.

Defeasible reasoning is another area which has recently been found promising for applying MWL in different HCI areas (Luca Longo, 2011, 2014, 2015a). Defeasible reasoning is considered to be a form of nonmonotonic reasoning (NMR) which occurs in situations where conclusions are drawn while it is known that further information may result in changing to another conclusion (Ford & Billington, 2000). MWL can be considered as defeasible as knowing different components of MWL result may change the conclusion arrived at. For example, it may be reasonable to assume a low MWL conclusion if a low amount of time was spent on a task whereas a human reporting a high amount of effort may change that conclusion to a high amount of MWL.

Recently, defeasible reasoning was used to infer MWL using a 5-layer argument-based framework, built upon Argumentation Theory (AT), see Figure 2.1, to represent domain knowledge compared favourably to state-of-the-art MWL inference techniques (Rizzo & Longo, 2017). Argumentation theory is a research subject for representing, supporting or discarding arguments for defeasible reasoning (L. Longo, Kane, & Hederman, 2012; Luca Longo & Hederman, 2013). The five layers detail what is considered in an argumentative system, what are the main components used and how they are strictly connected.

1.2 Research problem

As mentioned, MWL and its measurement is important for areas of HCI such as the analysis and design of interfaces and websites. Three types of MWL measurement method have been identified. These are subjective (e.g. survey after performing a task), physiological (e.g. tracking eye activity during a task) and performance (e.g. analysis of error rates following a task) (Cain, 2007; Luca Longo, 2015a).

The recent work on the assessment of MWL through the analysis of web user interaction logs and user activity indicators included the capability to be performed on a large scale ("data from hundreds of users") as an advantage as alternative measurement methods were found to be more costly to scale (Romero, 2017). For example, previous studies with experiments assessing MWL in the HCI field were found to be "invasive", requiring controlled and planned experiments, and had costs which increased as the number of participants increased. This gives motivation to pursuing the recommended further work to assess whether the value of one indicator is related with the value of two or more indicators. The indicators dataset which was created for that work contains data of user activity indicators, related MWL measures and MWL questionnaire answers.

Separately, the 5-layer argument-based framework used to represent MWL domain knowledge to infer MWL used explicit knowledge gathered from student questionnaires designed for the widely used subjective MWL measures: the NASA-Task Load Index (NASA-TLX) (Hart & Staveland, 1988) and the Workload Profile (WP) (Tsang & Velazquez, 1996). The 5-layer argumentation system compared favourably to state-of-the-art MWL inference techniques and is capable of using a knowledge base with multiple indicators, as "arguments" and their interactions as "relations". This opens the possibility of using the system with the previously mentioned indicators of user activity.

1.3 Research Objective

This project aims at contributing to the body of knowledge related to the research of the concepts of Mental Workload and Defeasible Reasoning in the field of Human-Computer Interaction. This research seeks to investigate the creation of a defeasible-reasoning based knowledge-base of web user activity indicators to infer the MWL of web-based tasks using the 5-layer argumentation system. A single research objective has been defined.

The research objective consists in representing mental workload by employing defeasible reasoning to create a model from a knowledge base of rules (arguments) using user activity features tracked online, while performing a web-based task. The model will infer an index of MWL from the data and this index will be compared against other baselines indices (subjective NASA-TLX and WP Mental Workload measurements)

which use different features such as survey answers completed by participants following their performance of the web tasks.

The research question and research hypotheses defined in order to accomplish this objective are:

- Can the mental workload of web-based tasks be represented and inferred implicitly from user activity using defeasible reasoning in a multi-layer argument-based framework, built upon argumentation theory?
- Null hypothesis (H0): the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly unrelated to baseline NASA-TLX or WP mental workload measures for the task.
- Alternative hypothesis (H1): the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly correlated to baseline NASA-TLX or WP mental workload measures for the task.

1.4 Research Methodologies

This work started with secondary research of the literature to understand state of the art studies and solutions related to the research problem in the areas of: Human-Computer Interaction, Design and Usability; Mental Workload; User Interaction Tracking; User Interaction Indicators; and Defeasible Reasoning and Argumentation Theory.
Also, it is a secondary analysis as the dataset used was obtained from an external source.
The experiment involved a computational AT approach for implementing Defeasible Reasoning to infer Mental Workload where both the input parameters and output results can be quantified, thus a quantitative research methodology is adopted.
Correlation was assessed by using statistical correlation coefficients, capable of measuring the extent to which those inferred and existing values related to each other, therefore, testing the hypotheses previously set for the project empirically.
The study described the environment and the design details of the experiment, and then continues deductively, from the general problem to the specific details.
For quick reference, the categories of the research methodologies used are shown in Table 1-1 Categories of the research methodologies used.

Type	Objective	Form	Reasoning
Secondary	Quantitative	Empirical	Deductive

Table 1-1 Categories of the research methodologies used

1.5 Scope and Limitations

The scope of the study as defined by the research objective was to investigate the possibility of assessing a possible correlation between User Activity and Mental Workload inferred using AT and Defeasible Reasoning.

Setting up the AT approach, while time consuming, was straight forward and involved creating the arguments and the relationships between the arguments called conflicts or attack.

The 108-row dataset for this research was created by (Romero, 2017) as part of an web-based user activity experiment including infographics which had 145 participations and includes MWL, performance and user interaction data.. The size of the dataset which was refined from 108 to 49 rows may be considered a limitation as strong correlations were not achieved.

1.6 Document Outline

This section provides a brief summary of the five chapters in this document:

- **Chapter 2** gives details from the **Literature Review** performed in the fields of Human-Computer Interaction, Design and Usability, Mental Workload, User Interaction Tracking, User Interaction Indicators and Defeasible Reasoning and Argumentation Theory which are relevant to this research.

  It begins giving some detail on Human-Computer Interaction, Design and Usability, how they have a focus on optimising performance for humans and how this can be found in common in areas of Mental Workload research. It goes on to give background and understanding of Mental Workload, the three type of measurement used for it and the main methods used in the subjective measurements type. It goes on to highlight the benefits of user interaction tracking. That leads into details about specific under interaction indicators. Following that, an examination of the fields of Defeasible Reasoning and Argumentation Theory is detailed. Then it summarises a review of the work

which was performed, presents the gap in the literature and concludes by detailing the research hypotheses and research question.

- **Chapter 3** gives details the **Design and Implementation** of a solution and series of experiments to test the research hypothesis. It starts with a definition of the hypotheses necessary to answer the research question. It also includes discussion of the planned work, software and components used for the experiment as well as and strengths and weaknesses of the design and implementation approach.

- **Chapter 4** discusses the **Results and Evaluation** of the experiments to test the research hypothesis. It presents the results and associated results data with the aim of analysing them and check for correlations between inferred MWL and subjective measures already in the dataset. It finishes with a look at the strengths and limitation of the results and evaluation approach.

- **Chapter 5** presents the C**onclusions** of the research. It summarises the research performed, how it contributes to the body of knowledge, proposes potential future areas of research and details the conclusions found from the research.

2 LITERATURE REVIEW

This chapter gives details on related mental workload, human-computer interaction, user activity and defeasible reasoning literature and literature providing the basis for the work being performed for this research. The research hypothesis and research question are also defined.

2.1 Overview

First, the Human-Computer Interaction, Design and Usability section is presented, which begins with an introduction to element of HCI, and looks at how it links to Usability and Mental Workload through their focus on improving or optimising design for use by humans.

The Mental Workload section gives foundation details including the main categories of Mental Workload measurements followed by the details on the subjective measures NASA Task Load Index and Workload Profile.

The User Interaction Tracking section details its benefits in areas such as usability, rapid experimentation and how it has been used in the field of Mental Workload. The User Interaction Indicators section details web-based indicators and how their measurement or interaction with others can be used to draw conclusions.

The Defeasible Reasoning and Argumentation Theory section introduces and explains the basis of defeasible reasoning. It explains how argumentation theory is used to implement defeasible reasoning and details the five layers generally used in argumentation theory systems. It also discusses Mental Workload research which has been done with Defeasible Reasoning and Argumentation Theory.

Finally, the Summary and Gap in Literature section reviews the work done based and the presents the gap that motivates the formulation of the research question.

2.2 Human-Computer Interaction, Design and Usability

Computing is becoming pervasive to the extent that it is seen to be moving into the background with human behaviour and human-computer interaction (HCI) taking on

more focus (Pantic et al., 2008). There has been much advancement in the field of HCI with a focus on interface design and user interaction aiming to make technology feel natural to use while also having it as hidden to the user as possible (Karray et al., 2008). A key goal for this is for computers systems to adapt and serve human needs to most efficiently facilitate communication between the two (Chao, 2009; Luca Longo, 2015b). It has been argued that usability issues such as cognition and satisfaction can lead to abandonment in the use of a variety of technologies, even those intended to be assistive to the user (Phillips & Zhao, 1993; Dawe, 2006). It has also been found that better design aesthetics do not necessarily correlate with better usability and that users give them differing priorities depending on the importance of their tasks (Hartmann et al., 2008). A recent review into what influences user adoption of information technology discussed how better user experience increases likelihood of technology adoption, that user enjoyment perceptions have influence and implied that designers need to be aware of the link between the features of their systems and adoption as well as how psychological needs analysis would be beneficial (Hornbæk & Hertzum, 2017).

Mental workload (MWL) is another area in which assessing, predicting and designing tasks as part of system design and operation is looked at in relation to human performance and has been found to have practical benefits such as reducing error rates (P. A. Hancock & Caird, 1993). This also applies to information technology where it has been shown, for example, in experiments looking to design website interfaces to minimise human information processing found that MWL can be used to compare different presentation mediums (Wästlund et al., 2008; L. Longo & Dondio, 2015; Luca Longo, 2017).

The next section goes into more detail about MWL, how it can relate to design and interaction as well as recent research in the area.

2.3 Mental Workload

While not having a clear definition, a review of the MWL literature found that workload tended to be concerned with the demands of tasks on their operators and the associated costs and performance of the tasks (Cain, 2007). The review provides a board outline of how the concept of MWL has arisen in many research areas, such as Interaction design, and has multiple means of measurement, representation and analysis.

For example, work which looked at using a roles approach for MWL argued that interface and interaction designs should be optimised to avoid 'mental overload' (Zhu & Hou, 2009). This has been supported in related work which showed that MWL measurement is also important alongside usability measurement in the field of web design and human computer interaction and was found, for example, that the higher the MWL of a website design, the lower the level of usability that may be associated with it (Luca Longo, Rusconi, Noce, & Barrett, 2012; Luca Longo, 2017). In a separate investigation of MWL for vehicle drivers, it was shown to be a highly important consideration for optimising design for assisting decision making by users (Silva, 2014). More recently, user activity recorded during the performance of web-based tasks has been investigated for its potential in measuring MWL at scale with interfaces using differing designs (Romero, 2017).

Subjective, physiological and performance methods have been identified as the main types of measurement of MWL (Cain, 2007; Luca Longo et al., 2012; Rizzo, Dondio, Delany, & Longo, 2016; Rizzo & Longo, 2017; Contreras, 2018).

- **Subjective**

  Subjective methods involve the analysis of feedback from task participants gathered through methods such as survey questionnaires answered following the task.

- **Physiological**

  Physiological methods involve the analysis of physiological data of task participants recorded during the task through methods such as eye tracking and heart rate monitoring.

- **Performance**

  Performance methods involve the analysis of participants performance estimations for the tasks such as error rates, time taken, and actions taken.

A novel method using data-driven computational models of MWL created by a selection of supervised Machine Learning (ML) classification techniques has also shown promise as a predictor of **objective** performance (Moustafa et al., 2017).

There have been multiple methods used for subjective measurement, however three of the primary methods NASA-TLX, Workload Profile (WP) and also the Subjective Workload Assessment Technique (SWAT) which measures across three dimensions:

time, mental and pyschological which trying to minimise subjectivity using controls have been compared (Rubio, Díaz, Martín, & Puente, 2004). The conclusions outlined that depending on the measurement goal, that together NASA-TLX and WP would cover all tasks while SWAT would only cover some goals.

NASA-TLX is used to measure the MWL of human-machine equipment across six dimensions: mental, physical, temporal, performance, effort and frustration (Hart & Staveland, 1988). Previous work has examined the relationship between 'cognitive load' and the design of web sites arguing that designs can modified reduce high cognitive load (Tracy & Albers, 2006). They used the NASA-TLX score as a post-event subjective measures along with other real-time measures.

The WP method measures multiple dimensions: solving and deciding, task and space, verbal material, auditory attention, speech response, response selection, visual attention and manual activity and is considered reliable and based on the multiple resource theory (Tsang & Velazquez, 1996).

The next section goes into detail about gathering data from users interactions on websites, how it can relate to the MWL of web based tasks and how it has pros and cons.

2.4 User Interaction Tracking

User Interaction Tracking has been shown to be assistive in website usability analysis. For example, specific website elements raising cognitive load have been identified by monitoring users capacity to continue finger tapping as a secondary task while performing a primary web-based task. (Albers, 2011). Monitoring does not have to be physiological and other research has shown that monitoring user activity such as cursor movements and scrolling in web-based search tasks was beneficial for understanding user judgements during the tasks and could be used for improving search based technologies (Guo & Agichtein, 2012).

Gathering data using online websites can have upsides like being able to rapidly perform experiments and downsides like not being able to control for when information appears for participant or for varying participant computer hardware (Woods, Velasco, Levitan, Wan, & Spence, 2015). However, it is argued that online tracking is acceptable for experiments not measuring precise response timing or requiring specific hardware

output like a particular tone outputted from a hardware speaker (van Steenbergen & Bocanegra, 2016).

The dataset being used in this research was created through user interaction tracking of participants in web based experiments to gather user activity indicators and was used in an attempt to determine the MWL of web-based tasks (Romero, 2017). The experiment used different web based interfaces and recording indicators such as scroll, mouse clicks, time taken etc recorded using Javascript and stored in log files. Successful collection of data was performed with 145 participations in the web-based experiments.

The next section gives details on user interaction indicators in web-based tasks, how they can lead to task completion or success as well as how they may influence each other.

2.5 User Interaction Indicators

Detailed user interaction tracking can be performed on websites using the standard web technology Javascript and multiple indicators have been examined by putting tracking code into websites (Atterer et al., 2006). While scrolling and mouse movements were not always being recorded, the following indicators were more consistently monitorable:

- Page load and resize events
- Browser window width and height
- Page focus, blur and unload events
- Mouse click and hover events
- Mouse position coordinates
- Key press events

More recently, similar embedding of client side logging into websites has been used including for click, movement and keyboard events with the focus on combining indicator information and other relevant information with rules to determine if interface elements are usable or if the websites are having accessibility issues (de Santana & Baranauskas, 2015).

Looking to the literature for the relevance of indicators, both single and multivariate indicators have been identified for use in predicting the relevancy of a webpage found

as a result of a web-search (Shapira, Taieb-Maimon, & Moskowitz, 2006). The most promising indicator was found to be mouse movement against reading time. A general finding was that more interaction with a webpage was an indicator of higher relevancy. Other research in the same domain reinforces the view that mouse movement against reading time as noteworthy as it was found to indicate better quality web browsing results than using simple measures of cursor travel distance or time spend on web pages (Huang, White, & Dumais, 2011).

Another indicator found was that users who spending longer times scanning a webpage may be struggling to find relevant information (Guo, Lagun, & Agichtein, 2012). Negative correlations for success were found when there was higher numbers of mouse clicks and where users dwell times on their webpages were higher. This focus on time as an indicator in research which suggested that providing user dwell time as a simple heuristic to assisted judges in their assessment of whether documents are high versus low effort for finding and using information (Yilmaz, Verma, Craswell, Radlinski, & Bailey, 2014).

The measurement of time as a web-based indicator has been investigated in an experiment comparing Javascript response times to those of a an established tool MATLAB's Psychophysics Toolbox (Brainard, 1997) which is capable of synchronising its own response time to that of a computer display device (de Leeuw & Motz, 2016). Javascript was found not to significantly alter the distribution of results and was able to be used in an experiment to determine if response times increased depending on the number of items on screen during a visual search task.

A 2016 paper looked at the number of user clicks in the different contexts of either being inside or outside a knowledge module they had added to a webpage as clicks are considered a prominent feature of user engagement. (Arapakis & Leiva, 2016).

The next section introduces and gives details about defeasible reasoning and argumentation theory.

2.6 Defeasible Reasoning and Argumentation Theory

Drawing a conclusion when it is known that further information may result in changing to another conclusion is known as nonmonotonic reasoning (NMR) and multiple complex systems of NMR have been developed (Ford & Billington, 2000). Defeasible

reasoning is considered to be a form of NMR and many approaches have been taken to build automated reasoning systems which include defeasible reasoning capabilities (Baroni, Guida, & Mussi, 1997). It arose from the field of artificial intelligence and can arise in situations where a premise or multiple premises which lead to one conclusion may no longer be justified when additional information is considered (Pollock, 1987). The premise, or "argument", itself is considered defeasible.

An argument can be considered as having three parts, a set of premises, a conclusion and an inference from the premises to the conclusion and argumentation can be seen as building a chain of arguments where the conclusion of one inference is a premise in the next (Walton, 2009). Argumentation theory (AT) is a research subject which has investigated how defeasible reasoning can be used to represent, support or discard arguments and it shows promising results compared against ML tools (L. Longo et al., 2012; Luca Longo & Hederman, 2013; L. Longo & Dondio, 2014).

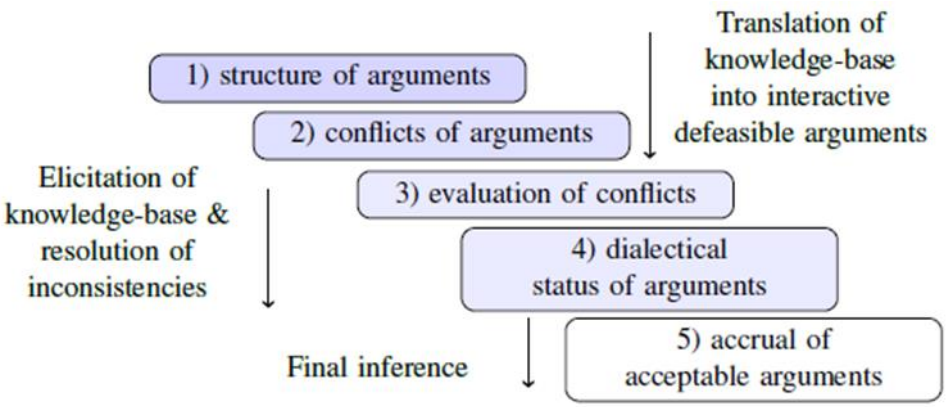

Figure 2.1 Five layers which argumentation systems are generally built upon (Rizzo & Longo, 2017)

AT systems generally contain five layers (Prakken & Vreeswijk, 2001) as seen in Figure 2.1. They are:

- **Layer 1**: definition of the internal structure of arguments

  Monological models are used for to represent the internal structure of arguments. This can be done by representing a set of premises and the conclusion to follow them with the application of a rule. Premises can be different and take on differing roles as in the argument model introduced in the Toulmin argument model (Verheij, 2009) as follows:

- Claim (C): the original assertion. It is the starting point of an argument. It can also be considered a claim (conclusion) with a potentially controversial nature;
- Data (D): the basis of the claim, i.e. the statements of facts or beliefs
- Warrant (W): the connection between the data and the claim, i.e. the statement entitles the conclusion or claim to be drawn from the data;
- Backing (B): the trustworthiness information of a warrant, i.e. why the warrant holds;
- Qualifier (Q): the strength of the moving from the data to the claim, i.e. statement expressing the level of certainty related to the claim;
- Rebuttal (R): a circumstance in which the warrant may be set aside, i.e. a situation statement which defeats the claim or conclusion.

An example illustration of Toulmin's arguments layout can be seen in Figure 2.2

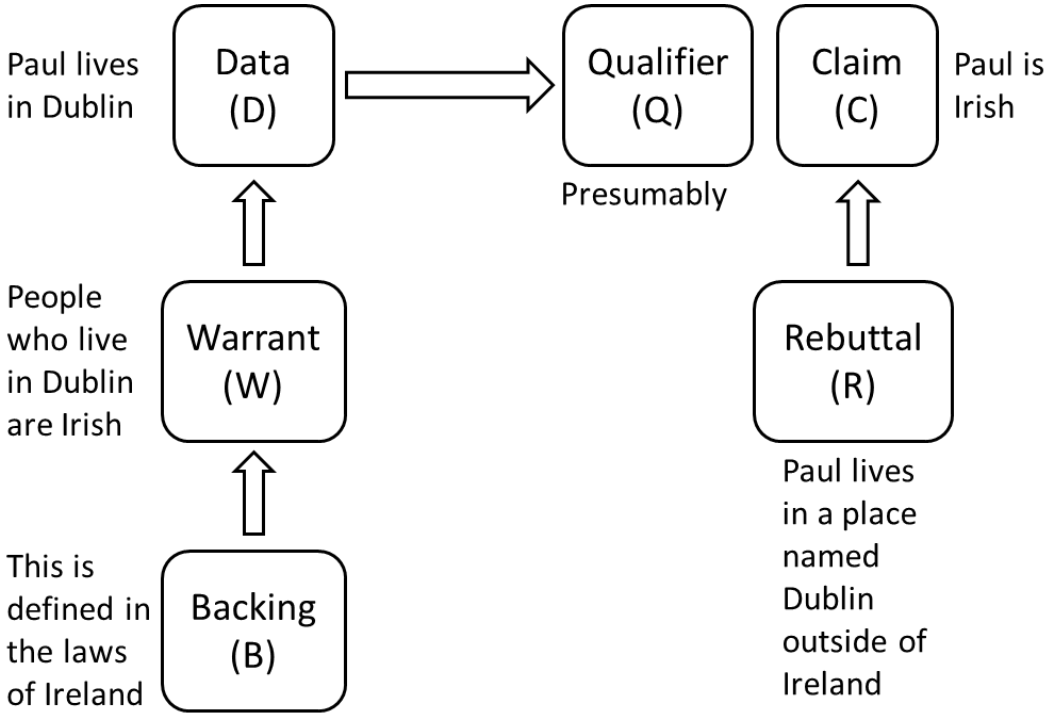

Figure 2.2 Flowchart example of Toulmin's arguments layout#

- **Layer 2**: definition of the conflicts between arguments

  The monological models are complemented by dialogical models to represent relationships typically known as a conflict (or as an attack or a defeat) between arguments. Three classes of conflict have been categorised as follows:

- Undermining attack: an argument premise is attacked by another argument where the conclusion negates the premise
- Rebutting attack: an argument conclusion is contradicted by the conclusion of another argument
- Undercutting attack: an argument defeasible inference rule attacked by another argument with a special case which does not allow the rule to be applied.

- **Layer 3**: evaluation of conflicts and definition of valid attacks

  While the conflicts between arguments are important, they do not evaluate whether an attack is successful (Luca Longo, 2016a). Attacks are generally binary relations between two arguments and two forms, a weaker 'defeat' and a stronger 'strict defeat', have been recognised. The strength can depend on a variety of considerations such as the test reliability or the observer expertise.

  Establishing if an attack can be considered a successful defeat can happen using strength of arguments, for example based on the preferentiality among arguments. This is based on the inequality of the strength of arguments which must be accounted for in the computation of sets of arguments and counterarguments.

  Another approach can be using preferentiality associated to attack relations instead of arguments. This is based on associating weights to attack relations instead of arguments. This approach allows the representation of the degree to which an argument attacks another as well allowing the assignment of probabilities to arguments and defeat relations referring to the likelihood of their existence in order to capture the inherent uncertainties in the argumentation system.

- **Layer 4**: definition of the dialectical status of arguments

  The strength of the arguments in their interactions with one another is important but does not tell which argument is justifiable and so a definition of their dialectical status is needed.

  This is usually done by splitting the set of arguments in two classes, arguments which support the decision or action and arguments which do not. A further split can be made of arguments an undecided status. These splits are not strict as

multiple actions or decisions can be accounted for in a defeasible reasoning process, so the number of classes can increase.

Modern implementations to calculate the dialectical status are based on Phan Minh Dung's abstract Argumentation Framework (AF) where the underlying idea is that given a set of abstract arguments (without considering the internal structure) and a set of defeat relations, the arguments to accepted can be decided. This includes not only looking at whether an argument is defeated by another "defeater" argument but also whether the defeater has itself been defeated. Arguments can be represented as nodes and attacks as arrows (Luca Longo, 2016a) as seen in Figure 2.3.

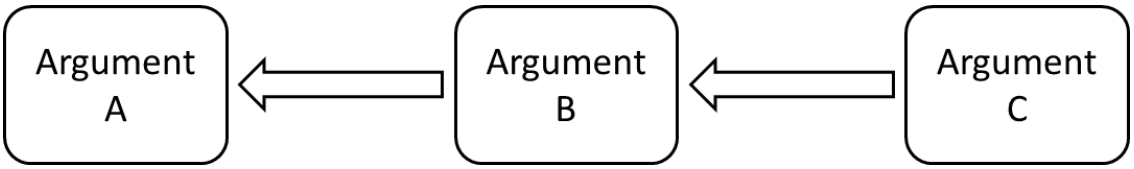

Figure 2.3 Visual representation of arguments and attacks

Argument node A is attacked by B which itself is attacked by C. As the is no attacker for C, it is not rejected but B is rejected. A is then accepted as B is no longer valid. Therefore, C reinstates A.

To decide which arguments are accepted, a formal criterion known as acceptability semantics is used which, given an AF, results in zero or more extensions (sets of acceptable arguments) being specified and labelled (Wu & Caminada, 2010) as follows:

- **Label: in** if it is accepted (only if all its defeaters are labelled out).
- **Label: out** if it is rejected (only if it has at least one defeater labelled in).
- **Label: undec** if it cannot be accepted or rejected.

A set of arguments:

- can only be called conflict-free if it does not contain any argument A and B such that A defeats B.
- can only be said to defend an argument C if each defeater of C is defeated by an argument in the set.

Together these present the concept called "**Complete Semantics**" which is used to compute complete extensions of acceptable arguments. More than one

complete extension can exist with complete semantics, but other semantics exist as follows:

- **Grounded Semantics**: this takes a more sceptical approach where exactly only one extension exists. For grounded semantics and complete semantics, the in-labelled and the out-labelled arguments are minimised while those labelled undec are maximised and also can be empty.
- **Preferred Semantics:** this takes a credulous approach which maximises the in-labelled and the out-labelled arguments, based on the idea of admissibility. A set is admissible if and only if it is conflict-free and defends at least itself. An empty set would always be admissible as due to being conflict free and having no defeaters. At least one preferred extension (accepted argument) always exists.

Every grounded and preferred extension is a complete extension.

The set in the example in Figure 2.3 are:

- **Admissible:** {C} and {A,C}
- **Not admissible:** {B} and {A} (as they are attacked)

Only one preferred extension exists: {A,C}.

Many other notions of argumentation semantics exist and a state of the art overview has been put together (Baroni, Caminada, & Giacomin, 2011).

- **Layer 5**: accrual of acceptable arguments

  This layer is considered one of the less developed layers in literature (Luca Longo, 2016a). Sometimes it is added to the argumentative schema aimed at extracting the most credible or consistent point of view for informing a decision or action. For example, it includes a strategy for computing a degree of credibility of each extension that can be used for comparison purposes.

  Layer 4 may result in multiple acceptable extensions of arguments being computed coinciding with possible consistent points of view that can be considered for describing the knowledge being modelled. These can be used for decision-making and defeasible inference. For practical purposes, sometimes a single decision must be taken, or a single action must be performed.

  The most credible for supporting decision-making can eventually be selected. A variety of strategies, such as considering the strength of arguments, a preference list among them, or the extension with higher cardinality have been proposed.

To summarise, the five layers detail what is considered in an argumentative system, what are the main components used and how they are strictly connected. Layer 1 deals handles monological argumentation and the other layers work with dialogical argumentation. Sometimes the layers can be missing or merged.

In the literature, an AT based framework aimed at encouraging MWL research has been implemented for defining, measuring, analysing, explaining and applying MWL in different HCI areas as a defeasible concept through interactions of defeasible arguments (Luca Longo, 2015a). The arguments used to calculate MWL are defeasible where one reason can defeat another, when determining a 'numerical usable index' as an assessment metric (Luca Longo, 2015a). An expert rules deductive model has been used to measure MWL and demonstrating that a knowledge base can be used in the calculation of MWL (Rizzo et al., 2016). Subsequently using a 5-layer AT based schema to represent domain knowledge to infer MWL using defeasible reasoning, as seen in Figure 2.1, was found to compare favourably to state-of-the-art MWL inference techniques (Rizzo & Longo, 2017).

2.7 Summary and Gap in Literature

The literature review indicates that promise has been shown in MWL research in the field of User Interaction Tracking of User Interaction Indicators as well as in the field of Defeasible Reasoning using AT systems.

The 5-layer argumentation system compared favourably to state-of-the-art MWL inference techniques and is capable of using a knowledge base with multiple indicators, as "arguments" and their interactions as "relations".

It seems like a 5-layer argumentation system has not been used with User Interaction Indicators to infer MWL. This opens the possibility of using the system with the indicators of user activity dataset.

Thus, the research gap being investigated in this research is whether the use of this defeasible reasoning based argumentation system using measures of user interaction indicators has the potential to infer MWL and so determine MWL implicitly for web-based tasks.

2.8 Research Question

This study covers many fields including Mental Workload and Defeasible Reasoning. The research seeks to investigate the creation of a defeasible-reasoning based knowledge-base of web user activity indicators to infer the MWL of web-based tasks using the 5-layer argumentation system.

A single research objective has been defined for studying the possible correlation between the Mental Workload inferred using Defeasible Reasoning from web user activity indicators recorded during a web task and the subjective NASA-TLX and WP Mental Workload measures. As a result, the research question that this study purposes start answering is:

- Can the mental workload of web-based tasks be represented and inferred implicitly from user activity using defeasible reasoning in a multi-layer argument-based framework, built upon argumentation theory?

Linked to this main research question, the following hypotheses and objectives were defined:

- Null hypothesis (H0): the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly unrelated to baseline NASA-TLX or WP mental workload measures for the task.
- Alternative hypothesis (H1): the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly correlated to baseline NASA-TLX or WP mental workload measures for the task.

3 DESIGN AND IMPLEMENTATION

This chapter outlines the design and implementation of a solution and experiment put together to analyse user activity to determine if mental workload can be inferred implicitly using defeasible reasoning.

An outline design of the solution can be seen in Figure 3.1. An online tool[1] for 'implementing argumentation theory in practice' by (Rizzo & Longo, 2017) is used. The user activity dataset created by (Romero, 2017) can be seen feeding into a five-layers based argumentation system to generate MWL measures based on defeasible reasoning rules. The measures are then compared with the NASA-TLX and WP MWL measures from the dataset which serve as baselines.

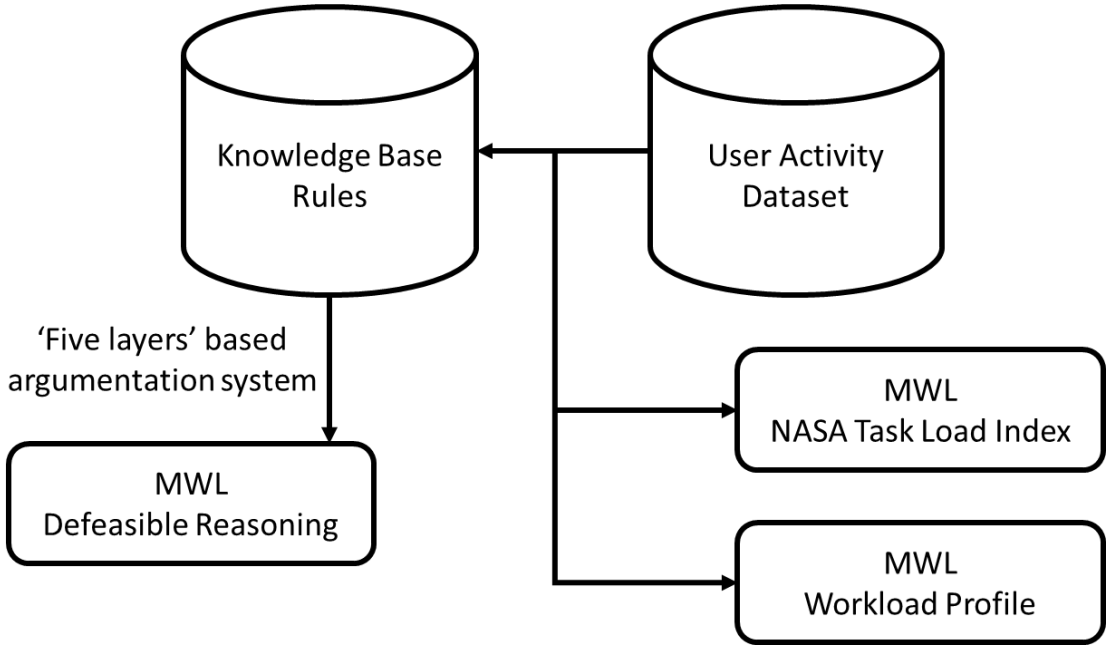

Figure 3.1 High level architecture of a solution to measure MWL using defeasible reasoning which incorporates NASA-TLX and WP measures for comparison

[1] http://lucalongo.eu/lucas/index.php

3.1 Hypotheses definition

As an experiment, six instantiations of the 5-layer schema were planned. Starting from building a simple argumentation system defeasible reasoning structure and progressing to more complex structures.

- The first was planned to use a single variable representing a single user activity indicator.
- The second was planned to use two variables representing user activity indicators and to include simple defeat rules.
- The third was planned to use three indicators and to include a complex set of defeat rules.
- A fourth was planned to use five indicators and to include an even further complex set of defeat rules.
- The fifth was planned to use all available user activity indicators and structure them hierarchically and using a complex set of defeat rules.

For use in the experiment, the original web-based task dataset of user interactions was refined to contain only:

- values to represent measurements of indicators of user activities
- NASA-TLX and WP values
- rows which had no blank entries

Based on that, the following hypotheses are as follows:

- Null hypothesis (H0): the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly unrelated to baseline NASA-TLX or WP mental workload measures for the task.
- Alternative hypothesis (H1): the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly correlated to baseline NASA-TLX or WP mental workload measures for the task.

3.2 Software

This section gives details of software used for the experimental approach. First, it describes the operation of the online tool for 'implementing argumentation theory in practice' by (Rizzo & Longo, 2017) used to represent the defeasible reasoning rules. Then, it details about the components of The R Project for Statistical Computing used for correlation analysis.

3.2.1 Argumentation framework (Online Tool)

An online tool[2] for 'implementing argumentation theory in practice' by (Rizzo & Longo, 2017) has been created. The implementation of the solution devised for this work was assembled by creating defeasible user activity structures using this online tool. Structures are created using this online argumentation system as per the five AT layers in Figure 3.2. An example could be as follows:

- Layer 1 - Definition of the structure of arguments
  Arguments are represented as nodes as seen in Figure 3.2

Figure 3.2 Argument represented visually as a node

Multiple nodes can be used to represent different arguments. The premise and conclusion of each argument are embedded within the node and defined using a dialog as seen in Figure 3.3

[2] http://lucalongo.eu/lucas/index.php

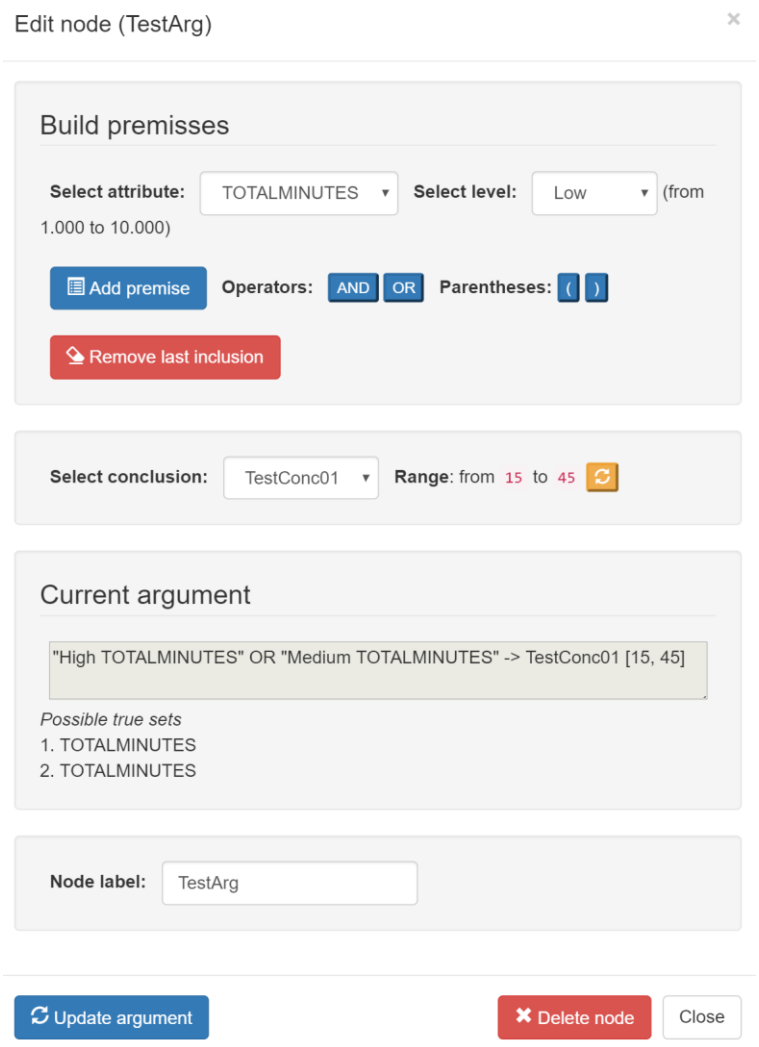

Figure 3.3 Selecting the premise and conclusion of an argument

Premises and conclusions are defined using another dialog as seen in Figure 3.4

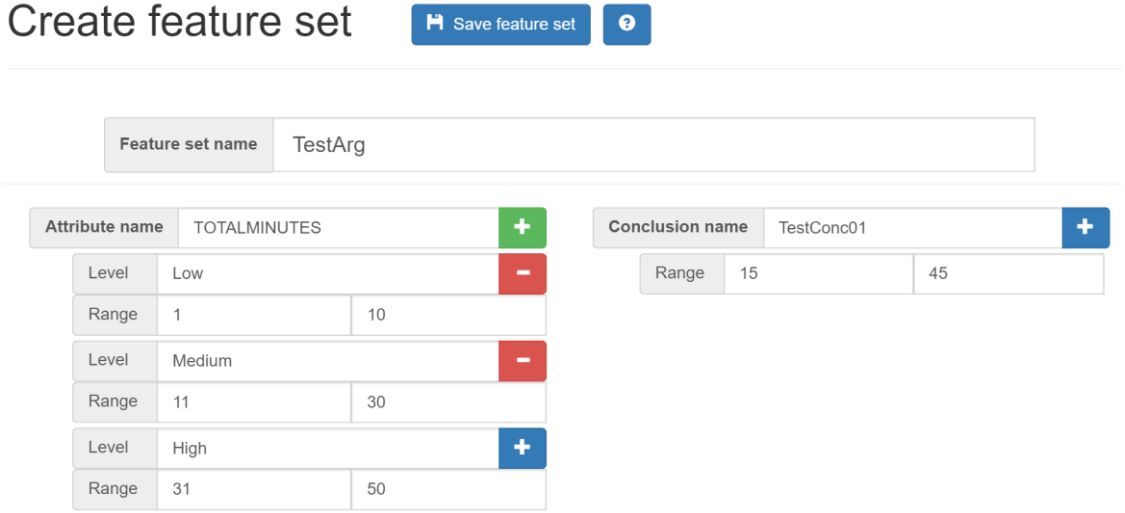

Figure 3.4 Defining premises and conclusions as a feature set

- Layer 2 - Definition of the conflicts of arguments

  Conflicts between arguments allow for one to attack another. An argument node attacks another by connecting an arrow pointing to the node being attacked as seen in Figure 3.5

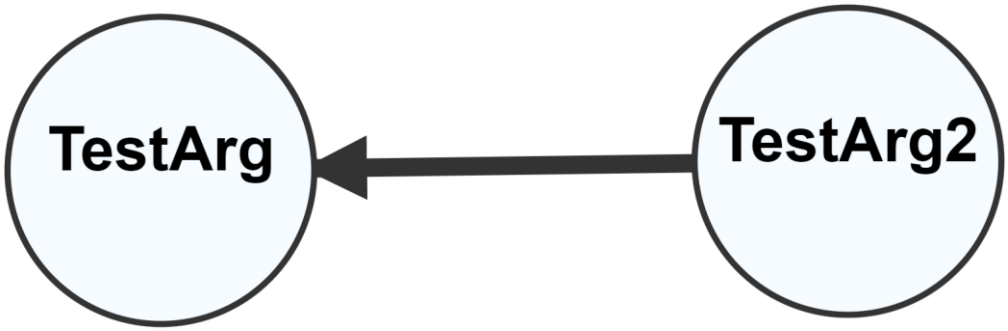

Figure 3.5 Argument node TestArg2 attacking argument node TestArg

The TestArg2 node is attacking the TestArg node as indicated by the arrow connecting the two and pointed toward the TestArg node. An argument node can be used to attack multiple argument nodes. It is also possible for an argument node to be attacked by multiple other nodes.

- Layer 3 - Evaluation of the conflicts of arguments

  Once the argument nodes and their attacks have been put into place, the knowledge base of the designer is represented in the argumentation system. Input data can then be used to activate the arguments. For example, TestArg in Figure 3.5 would be activated for attribute TOTALMINUTES data in the range of to 10 as per the premise and conclusion definition seen in Figure 3.3. Other argument nodes including attacking nodes such as TestArg2 can be activated depending on the input data. Argument nodes which are not activated are discarded.

- Layer 4 - Definition of the dialectical status of arguments

  At this layer, a variety of semantics can be applied to determine how arguments are accepted. For example, inconsistencies can arise where an argument node may be defeated due to an attack, but the attacker may also be attacked. Sets of acceptable arguments, known as extensions, are produced. As described in the

Literature Review section on Defeasible Reasoning and Argumentation Theory, Grounded semantics produce unique extensions whereas preferred semantics can produce one or more extensions. Multiple extensions are quantified by credibility using cardinality of the extensions. Higher cardinality is seen as more credible and lower is seen as less credible.

- Layer 5 - Accrual of acceptable arguments and computation of MWL
  A value for MWL is inferred following the extension generation in Layer 4. MWL values are calculated for each of the arguments in an extension and a result is generated through the aggregation of these values.

3.2.2 Correlation Analysis using R (Statistics Software)

The R software environment for statistical computing and graphics (R Core Team, 2018). was chosen for the experimental correlation analysis of the reference NASA-TLX and Workload Profile mental workload measurements and those represented and inferred from the defeasible reasoning knowledge base created using indicators of user interactions from a web-based task with the Argumentation framework (Online Tool). It was selection due to the way it provides simple and fast analysis with a range of techniques to conduct descriptive statistics, test of normality, and statistical tests to determine the relationship between variables.

The R package GGally (Schloerke et al., 2018) was also selected for concise visualisation and simple examination of the correlation coefficients using correlation matrixes plotted with the ggcorr function of the package. A correlation matrix plotted using a reduced 108 row dataset created as part of this work can be seen in Figure 3.6

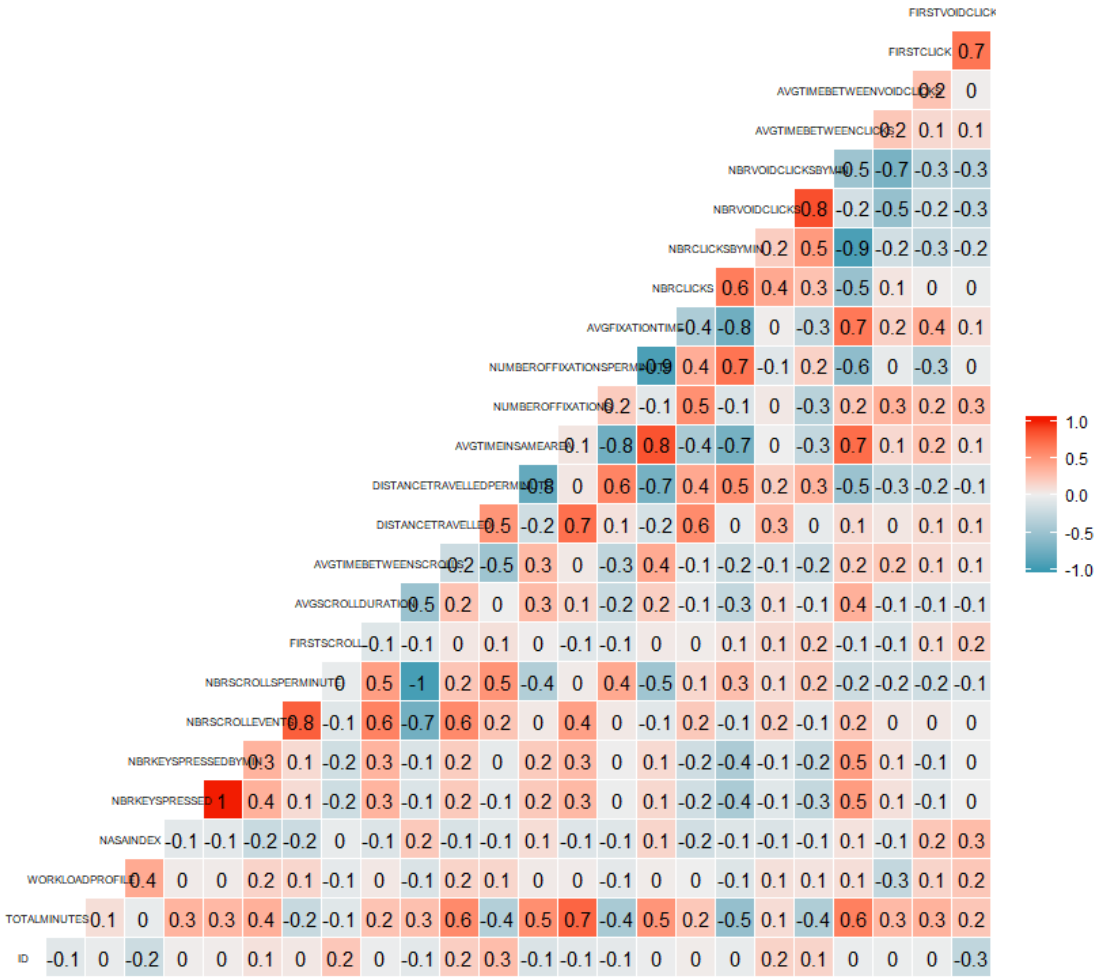

Figure 3.6 Correlation matrix of the 108 row reduced dataset

3.3 Data Understanding

This section gives details of the analysis of the dataset for used in the experiment.

3.3.1 Dataset

The dataset came in the form of a file in a comma separated variable format. When opened in a spreadsheet program such as LibreOffice Calc, it presents as Rows and Columns of data. Analysis of the file should:

- Begin with identifying the columns related to indicators of user activity.
- Identify data quality problems, specifically any rows containing indicators with blank data.

- Identify the columns needed for correlation analysis to verify the argumentation system results.

In parallel and to assist with the analysis of the dataset, a characterisation of the data should be performed. This can be done by creating a table representing the headers of each column and to which category they were identified.

3.4 Data Preparation

This section gives details of how the dataset should be prepared for use in the experiment.

3.4.1 Dataset

With rows and columns identified following analysis of the dataset, actions should be taken as follows:

- Columns related to indicators of user activity. These should be retained for use with the experiment argumentation system.
- Data quality (specifically rows containing indicators with blank data): These should not be retained to ensure maximum utility of the experiment argumentation system structures which is to use the indicators.
- Determine the columns needed for correlation analysis to verify the argumentation system results. These should be retained for the Correlation Analysis using R (Statistics Software).
- Remaining rows and columns left unidentified: These should not be retained to ensure maximum utility of the experiment argumentation system structures which is to use the indicators.

3.5 Argumentation System Modelling

Instantiations of the defeasible reasoning based 5-layer schema can be modelled using the argumentation system as described in the example in the Argumentation framework (Online Tool) subsection of the Software section. It can be summarised as:

- Layer 1 – (Multiple) Arguments are represented as nodes. Premises and conclusions are embedded within the nodes.
- Layer 2 - Argument nodes attacks one another by connecting an arrows pointing to the nodes being attacked
- Layer 3 – The knowledge base of the designer is represented in the argumentation system. Input data can then be used to activate the arguments.
- Layer 4 - Semantics can be applied to determine how arguments are accepted.
- Layer 5 – A value for MWL is inferred following the extension generation in Layer 4. MWL values are calculated for each of the arguments in an extension and a result is generated through the aggregation of these values.

3.5.1 Common definitions for Premises and Conclusions

The standardisation of calculating premises and conclusions for each of the experiment instantiations would allow a more for focus to be put on to arguments and their conflicts. Premises attributes will represent the user activity indicators and premises levels will represent values within each attributes range. Conclusions will represent values of the range of MWL.Therefore Common definitions for calculating Premises and Mental Workload Conclusions would be beneficial.

3.5.2 Instantiations of the 5-layer schema

Six instantiations of the 5-layer schema were planned as follows:

- a single variable representing a single user activity indicator.
- use two variables representing user activity indicators and to include simple defeat rules.
- use three indicators and to include a complex set of defeat rules.
- use five indicators and to include an even further complex set of defeat rules.
- to use all available user activity indicators and structure them hierarchically using NASA-TLX associations and include a complex set of defeat rules.
- to use all available user activity indicators and structure them hierarchically using WP associations and include a complex set of defeat rules.

For the non-all variable instantiations, intuitively selected user indicators can be represented as arguments and have conflicts created. For the all variable instantiations, related indicators should be hierarchically placed closer to each other, again in an intuitive manner.

3.6 Evaluation and Hypothesis Testing

To assess the level of correlation between the mental workload represented and inferred from the defeasible reasoning knowledge and the baseline NASA-TLX or WP mental workload measures for the task, the Spearman correlation coefficient was used.

The previous work with this same dataset considered using both the Pearson and Spearman correlation coefficients (Romero, 2017). However, following scatter plots analysis it was found that outlier would affect the values of the Pearson coefficient and so only the Spearman coefficient was used. Therefore, only the spearman correlation coefficient is used in this work.

The Spearman coefficient looks at the monotonic relationship where one variable change may be associated with another variable changing, but not necessarily at a constant rate or proportionally. The direction and strength of the relationship is indicated by the result with a negative result indicating the variable increases while other decreases while a positive result indicates the variable increases while other increases.

The correlation coefficient will be generated using the R software environment for statistical computing and graphics (R Core Team, 2018). The evaluation will involve face validity and convergent validity tests (Tsang & Velazquez, 1996) The face validity test looks at summaries of the correlation coefficients using correlation matrixes. The tests for convergent validity were performed by comparing correlations mental workload scores against baseline, NASA-TLX and WP, measures from the dataset (Romero, 2017).

3.6.1 Accepting or rejecting H1

The following acceptance evaluation criteria will be used to test the hypothesis H1:

- Performance of face validity tests (simple analysis looking at a representation of the correlation coefficients).

- Performance of convergent validity correlation tests (analyse full generation of Spearman correlation coefficient output) if any of the results of the face validity tests were above 0.2 or below -0.2.
- Acceptance testing of the research hypothesis where a moderate (greater than 0.5) or high (greater than 0.7) strength correlation coefficient was found as per previous work in the field of MWL (Romero, 2017).

3.7 Strengths and weaknesses of the approach

This section details strengths and weaknesses identified in the design and implementation of the research experimental approach. The design, components and implementation of five instantiations of the 5-layer schema have been outlined.

3.7.1 Strengths

The following strengths are in the proposed design and implementation:

- The online argumentation system being used is fully capable of representing all defeasible reasoning capabilities required for this work.
- The chosen methods for measuring MWL were identified, by literature, as being reliable and having good subjects' acceptance.
- The original dataset author claims the user activity indicators provided were strongly supported by literature (Romero, 2017).
- The 6 instantiations of the 5-layer schema detailed will provide analysis of single, multi and all variable scenarios.

3.7.2 Weaknesses

The following weaknesses are in the proposed design and implementation:

- The size of the original dataset is only 108 rows and is subject to a data quality analysis which may reduce it. This is likely to reduce the capability to generate enough result data to get statistically significant Spearman correlation coefficient results.

- Intuition is being used in the design of the instantiations of the 5-layer schema. A more data or research grounded approach would be preferable.
- There are many user activity indicators details available in the dataset and so potential other instantiation designs have not been investigated, designed or explored.

4 RESULTS AND EVALUATION

This chapter details the dataset analysis and preparation, the argumentation system rules and representations created, the output results and the generation of correlation coefficients for the series of instantiations of the 5-layer schema. Results were generated using the dataset and argumentation system. Generation of the correlation coefficients was performed using analysis of the defensibly reasoned MWL measures and the NASA-TLX and WP measures.

4.1 Introduction

The analysis of the dataset was straightforward and resulted in the creation of tables or text regarding dataset rows and columns.

The creation of the argumentation system rules and representations created was straightforward and resulted in the creation of tables and online resources captured as images in this document.

The analysis the of correlation coefficients results was not complex and involved face validity and convergent validity tests (Tsang & Velazquez, 1996) The face validity test looked at summaries of the correlation coefficients using correlation matrixes. The tests for convergent validity were performed by comparing correlations mental workload scores against baseline, NASA-TLX and WP, measures from the dataset (Romero, 2017).

The following acceptance evaluation criteria were used to test the research hypothesis:

- Performance of face validity tests.
- Performance of convergent validity correlation tests if any of the results of the face validity tests were above 0.2 or below -0.2.
- Acceptance testing of the research hypothesis where a moderate (greater than 0.5) or high (greater than 0.7) strength correlation coefficient was found as per previous work in the field of MWL (Romero, 2017).

4.2 Data Understanding

This section gives details the finding of the analysis of the dataset for the experiment.

4.2.1 Dataset

Analysis of the original dataset file show that it contained:

- values to represent measurements of indicators of user activities
- answers from subjective MWL survey questionnaires completed by users
- NASA-TLX and WP values calculated using the MWL survey questionnaires
- details on the traits, such as gender and age, of the users from the experiments

The dataset also contained blank entries. 67 columns are in the dataset and during analysis, they were divided into the following categories:

- Task Description
- Participant Traits
- Participant State
- Participant Task Survey
- Participant MWL Survey
- Participant Calculated MWL measurements
- User activity indicators

These categories and their associate column names can be seen in :Table 4-1

Category	**Column Names**
Task Description	ID, TASKTYPE, INTERFACE, START_TIME, END_TIME, PERFORMANCE_SCORE
Participant Traits	GENDER, AGE, NATIONALITY, EXPERTISE, EDUCATION
Participant State	PRE_FRUSTRATION, PRE_MOTIVATION, PRE_AROUSAL, PRE_EMOSTATE
Participant Task Survey	TASK1_SURVEY_Q1, TASK1_SURVEY_Q2, TASK2_SURVEY_Q1, TASK2_SURVEY_Q2
Participant MWL Survey	MWL_MENTAL, MWL_PHYSICAL, MWL_TEMPORAL, MWL_PERFORMANCE, MWL_EFFORT, MWL_FRUSTRATION, MWL_MWL, MWL_PARALLELISM, MWL_MANUALACT, MWL_VISUALACT, MWL_SOLVEDEC, MWL_CONTEXT, MWL_MOTIVATION, MWL_SKILL, MWL_KNOWLEDGE, MWL_ALERTNESS, MWL_TASKSPACE, MWL_VERBALMAT, MWL_AUDITORYATT, MWL_SPEECHRESP, MWL_RESPONSESEL
Participant Calculated MWL measurements	WORKLOAD_PROFILE, NASA_INDEX

Category	Column Names
User activity indicators	TOTAL_MINUTES,NBR_KEYS_PRESSED, NBR_KEYS_PRESSED_BY_MIN, AVG_TIME_BETWEEN_KEYUSE, FIRST_KEYBOARD_USE, NBR_SCROLL_EVENTS, NBR_SCROLLS_PER_MINUTE, FIRST_SCROLL, AVG_SCROLL_DURATION, AVG_TIME_BETWEEN_SCROLLS, DISTANCE_TRAVELLED, DISTANCE_TRAVELLED_PER_MINUTE, AVG_TIME_IN_SAME_AREA, NUMBER_OF_FIXATIONS, NUMBER_OF_FIXATIONS_PER_MINUTE, AVG_FIXATION_TIME, NBR_CLICKS, NBR_CLICKS_BY_MIN, NBR_VOID_CLICKS, NBR_VOID_CLICKS_BY_MIN, AVG_TIME_BETWEEN_CLICKS, AVG_TIME_BETWEEN_VOID_CLICKS, FIRST_CLICK, FIRST_VOID_CLICK, HAS_CLICKS

Table 4-1 Dataset categories and associated column names

The identified columns related to indicators of user activity are listed in Table 4-2

TOTAL_MINUTES,NBR_KEYS_PRESSED	FIRST_KEYBOARD_USE
NBR_KEYS_PRESSED_BY_MIN	NBR_SCROLL_EVENTS
AVG_TIME_BETWEEN_KEYUSE	FIRST_SCROLL
NBR_SCROLLS_PER_MINUTE	AVG_SCROLL_DURATION
NUMBER_OF_FIXATIONS_PER_MINUTE	HAS_CLICKS
AVG_TIME_BETWEEN_CLICKS	DISTANCE_TRAVELLED
AVG_TIME_BETWEEN_VOID_CLICKS	AVG_TIME_IN_SAME_AREA
DISTANCE_TRAVELLED_PER_MINUTE	NUMBER_OF_FIXATIONS
AVG_FIXATION_TIME	NBR_CLICKS
NBR_CLICKS_BY_MIN	FIRST_CLICK
NBR_VOID_CLICKS_BY_MIN	FIRST_VOID_CLICK
AVG_TIME_BETWEEN_SCROLLS	NBR_VOID_CLICKS

Table 4-2 Columns related to indicators of user activity

In analysing the dataset file for any data quality problems, 60 rows containing indicators blank data were identified.

These were the columns identified as needed for correlation analysis to verify the argumentation system results: WORKLOAD_PROFILE, NASA_INDEX

4.3 Data Preparation

This section gives details of how the dataset was prepared for use in the experiment.

4.3.1 Dataset

For the purpose of this research, a reduced dataset was created containing only:

- Columns to represent measurements of indicators of user activities
- NASA-TLX and WP values calculated using the MWL survey questionnaires

The reduced dataset contained 108 rows of data and it continued to contain blank entries.

A refined reduced dataset was created containing only rows which had no blank entries.

The refined dataset contained 48 rows of data.

4.4 Argumentation System Modelling

This section details common sets of conclusions and attributes used in all the instantiations of the 5-layer schema. Following that, it outlines structure of each of the five experiments.

4.4.1 Mental Workload Conclusions

In. the instantiations of the 5-layer schema which were created, conclusion ranges for MWL were used as per the definitions in Table 4-3

MWL Conclusion	**MWL Range**
LowMentalWorkload	0 to 33
MedMentalWorkload	34 to 66
HighMentalWorkload	67 to 100

Table 4-3 Mental Workload Conclusion Levels

This was created using a simple approximate division of the range of MWL values (from 0 to 100) into three parts; low, medium, and high mental workload.

4.4.2 Premises Attributes and Levels

In. the instantiations of the 5-layer schema which were created, user activity indicators were assigned to be attributes with attribute levels. Three levels for each indicator were calculated using the refined dataset with 49 rows. It was decided to use three so that they could be assigned to the three conclusion ranges used for MWL. The attribute levels were calculated as follows:

- A low level was created from 0 to the mean.
- A medium level was created from the mean to one standard deviation above the mean.
- A high level was created from one standard deviation above the mean to the maximum value recorded for the indicator.

For example, the levels for the DISTANCETRAVELLEDPERMINUTE indicator can be seen in Table 4-4

Attribute Level	Attribute Level Range
low_distrav_per_min	0 to 4934
med_distrav_per_min	4935 to 7814
high_distrav_per_min	7815 to 16877

Table 4-4 Attribute levels for the DISTANCETRAVELLEDPERMINUTE user activity indicator

4.4.3 Single variable instantiation of the 5-layer schema

An initial instantiation of the 5-layer schema using a single variable representing a single user activity indicator was planned. The indicator chosen was DISTANCETRAVELLEDPERMINUTE which represents the distance the mouse travelled per minute during an online web task. The indicator is seen in Figure 3.6 as having both a negative correlation with NASA-TLX and a positive correlation with WP MWL. Both are low weak correlations and non-significant. The arguments used in the instantiation are defined in Table 4-5

Argument	Argument Definition
lowDistTravPerMin	When DISTANCETRAVELLEDPERMINUTE is low Then LowMentalWorkload
medDistTravPerMin	When DISTANCETRAVELLEDPERMINUTE is medium Then MedMentalWorkload
highDistTravPerMin	When DISTANCETRAVELLEDPERMINUTE is high Then HighMentalWorkload

Table 4-5 Arguments defined using the DISTANCETRAVELLEDPERMINUTE user activity indicator

No conflicts would be present as the levels of the indicator were separate and resulted in separate conclusions.

4.4.4 Two-variable instantiation of the 5-layer schema

A second instantiation of the 5-layer schema using two variables representing user activity indicators was planned. The indicators chosen were:

- NBRCLICKS which represents the number of clicks a user made during an online web task. The indicator is seen in as having a negative correlation with NASA-TLX and a non-existent correlation with WP MWL. It is low weak correlation and non-significant.
- TOTALMINUTES which represents the time it took a user to complete an online web task. The indicator is seen in as having a non-existent correlation with NASA-TLX and a positive correlation with WP MWL. It is low weak correlation and non-significant.

The arguments used in the instantiation are defined in Table 4-6 and Table 4-7.

Argument	**Argument Definition**
lowNumClks	When NBRCLICKS is high Then LowMentalWorkload
medNumClks	When NBRCLICKS is medium Then MedMentalWorkload
highNumClks	When NBRCLICKS is low Then HighMentalWorkload

Table 4-6 Arguments defined using the NBRCLICKS user activity indicator

Argument	**Argument Definition**
lowTotalMins	When TOTALMINUTES is low Then LowMentalWorkload
medTotalMins	When TOTALMINUTES is medium Then MedMentalWorkload
highTotalMins	When TOTALMINUTES is high Then HighMentalWorkload

Table 4-7 Arguments defined using the TOTALMINUTES user activity indicator

Four conflicts were planned:

- Two undercutting attacks from lowNumClks attacking medTotalMins and highTotalMins
- Two undercutting attacks from lowTotalMins attacking medNumClks and highNumClks

4.4.5 Multi-variable instantiation of the 5-layer schema

A third instantiation of the 5-layer schema was planned with the indicators NBRCLICKS, TOTALMINUTES and DISTANCETRAVELLEDPERMINUTE described for the previous instantiations.

A hierarchical approach using undercutting attacks was planned:

- Arguments from a single indicator would attack arguments from only one other indicator.
- LowMentalWorkload Arguments would attack MedMentalWorkload and HighMentalWorkload arguments.
- MedMentalWorkload Arguments would attack LowMentalWorkload and HighMentalWorkload arguments.
- HighMentalWorkload Arguments would attack MedMentalWorkload and LowMentalWorkload arguments.
- TOTALMINUTES arguments would be attacked by NBRCLICKS arguments
- DISTANCETRAVELLEDPERMINUTE arguments would be attacked by NBRCLICKS arguments

4.4.6 Multi-variable instantiation of the 5-layer schema with rebuttal attacks

A fourth instantiation of the 5-layer schema was planned with the indicators NBRCLICKS, TOTALMINUTES and DISTANCETRAVELLEDPERMINUTE described in the previous instantiations as well as another two indicators. The first additional indicator was FIRSTVOIDCLICK which represents the first time a void click was made during an online web task. These were clicks which were not associated with a button or a link. The second was AVGTIMEBETWEENVOIDCLICKS which

represents the average time between void clicks. FIRSTVOIDCLICK is seen as having positive correlations while AVGTIMEBETWEENVOIDCLICKS is seen as having negative correlations with both NASA-TLX and WP MWL. Both are low weak correlations and non-significant.

The additional two arguments used in the instantiation are defined in Table 4-8 and Table 4-9.

Argument	**Argument Definition**
lowNumClks	When FIRSTVOIDCLICK is low Then LowMentalWorkload
medNumClks	When FIRSTVOIDCLICK is medium Then MedMentalWorkload
highNumClks	When FIRSTVOIDCLICK is high Then HighMentalWorkload

Table 4-8 Arguments defined using the FIRSTVOIDCLICK user activity indicator

Argument	**Argument Definition**
lowTotalMins	When AVGTIMEBETWEENVOIDCLICKS is high Then LowMentalWorkload
medTotalMins	When AVGTIMEBETWEENVOIDCLICKS is medium Then MedMentalWorkload
highTotalMins	When AVGTIMEBETWEENVOIDCLICKS is low Then HighMentalWorkload

Table 4-9 Arguments defined using the AVGTIMEBETWEENVOIDCLICKS user activity indicator

A complex set of interconnected undercutting attacks as well as the use of a rebuttal attack was planned:

- TOTALMINUTES arguments would be attacked by DISTANCETRAVELLEDPERMINUTE arguments:

    - LowMentalWorkload Arguments would attack LowMentalWorkload and MedMentalWorkload arguments
    - MedMentalWorkload Arguments would attack LowMentalWorkload and HighMentalWorkload arguments
    - HighMentalWorkload arguments would attack MedMentalWorkload and HighMentalWorkload arguments
- NBRCLICKS arguments would be attacked by TOTALMINUTES arguments:
    - LowMentalWorkload Arguments would attack LowMentalWorkload and MedMentalWorkload arguments
    - MedMentalWorkload Arguments would attack LowMentalWorkload arguments
    - HighMentalWorkload arguments would attack MedMentalWorkload and HighMentalWorkload arguments
- TOTALMINUTES arguments would be attacked by NBRCLICKS arguments:
    - LowMentalWorkload would attack HighMentalWorkload arguments
    - MedMentalWorkload Arguments would attack HighMentalWorkload arguments (this is the single **rebuttal** attack)
    - HighMentalWorkload would attack LowMentalWorkload arguments
- NBRCLICKS arguments would be attacked by FIRSTVOIDCLICK arguments:
    - LowMentalWorkload Arguments would attack LowMentalWorkload and MedMentalWorkload arguments
    - MedMentalWorkload Arguments would attack LowMentalWorkload and HighMentalWorkload arguments
    - HighMentalWorkload arguments would attack MedMentalWorkload and HighMentalWorkload arguments
- NBRCLICKS arguments would be attacked by AVGTIMEBETWEENVOIDCLICKS arguments:
    - LowMentalWorkload Arguments would attack MedMentalWorkload and HighMentalWorkload arguments
    - MedMentalWorkload Arguments would attack LowMentalWorkload and HighMentalWorkload arguments
    - HighMentalWorkload arguments would attack LowMentalWorkload and MedMentalWorkload arguments

- FIRSTVOIDCLICK arguments would be attacked by AVGTIMEBETWEENVOIDCLICKS arguments:
    - LowMentalWorkload Arguments would attack MedMentalWorkload and HighMentalWorkload arguments
    - MedMentalWorkload Arguments would attack LowMentalWorkload and HighMentalWorkload arguments
    - HighMentalWorkload arguments would attack LowMentalWorkload and MedMentalWorkload arguments

4.4.7 All-variable instantiations of the 5-layer schema

Instantiations of the 5-layer schema using variables representing all available user activity indicators were planned. Two sets of arguments were made, one for NASA-TLX and one for WP, using the correlations between the indicators and the MWL measures NASA-TLX and WP as seen in Figure 3.6. Low, medium and high level arguments were defined for each indicator.

The conditions used to assign arguments levels for NASA-TLX are defined in Table 4-10

Argument Level	**Argument Definition**
Low	• If the correlation between the indicator and NASA-TLX is positive and the value is low Or • If the correlation between the indicator and NASA-TLX is negative and the value is high Or • If the correlation between the indicator and NASA-TLX is zero and the correlation between the indicator and WP is positive and the value is low Or • If the correlation between the indicator and NASA-TLX is zero and the correlation between the indicator and WP is negative and the value is high Then LowMentalWorkload
Medium	When the value is medium Then MedMentalWorkload

Argument Level	Argument Definition
High	• If the correlation between the indicator and NASA-TLX is positive and the value is high Or • If the correlation between the indicator and NASA-TLX is negative and the value is low Or • If the correlation between the indicator and NASA-TLX is zero and the correlation between the indicator and WP is positive and the value is high Or • If the correlation between the indicator and NASA-TLX is zero and the correlation between the indicator and WP is negative and the value is low • Then HighMentalWorkload

Table 4-10 Conditions used to assign arguments levels for the NASA-TLX All Variables instantiation

The conditions used to assign arguments levels for arguments levels for WP are defined in Table 4-11

Argument Level	**Argument Definition**
Low	• If the correlation between the indicator and WP is positive and the value is low Or • If the correlation between the indicator and WP is negative and the value is high Or • If the correlation between the indicator and WP is zero and the correlation between the indicator and NASA-TLX is positive and the value is low Or • If the correlation between the indicator and WP is zero and the correlation between the indicator and NASA-TLX is negative and the value is high Then LowMentalWorkload
Medium	When the value is medium Then MedMentalWorkload

Argument Level	Argument Definition
High	• If the correlation between the indicator and WP is positive and the value is high Or • If the correlation between the indicator and WP is negative and the value is low Or • If the correlation between the indicator and WP is zero and the correlation between the indicator and NASA-TLX is positive and the value is high Or • If the correlation between the indicator and WP is zero and the correlation between the indicator and NASA-TLX is negative and the value is low Then HighMentalWorkload

Table 4-11 Conditions used to assign arguments levels for the WP All Variables instantiation

The argument names used to represent the indicators are listed in Table 4-12

Argument Name	**User Activity Indicator**
TotalMins	TOTALMINUTES
NumClks	NBRCLICKS
NumClksPerMin	NBRCLICKSBYMIN
AvgTimeBetClks	AVGTIMEBETWEENCLICKS
FirstClick	FIRSTCLICK
VdClks	NBRVOIDCLICKS
VdClksPerMin	NBRVOIDCLICKSBYMIN
AvgTimeBetVdClks	AVGTIMEBETWEENVOIDCLICKS
FirstVdClk	FIRSTVOIDCLICK
NumKeys	NBRKEYSPRESSED
NumKeysPerMin	NBRKEYSPRESSEDBYMIN
DistTrav	DISTANCETRAVELLED
DistTravPerMin	DISTANCETRAVELLEDPERMINUTE
AvgTimeSameArea	AVGTIMEINSAMEAREA
NumScrolls	NBRSCROLLEVENTS
NumScrollsPerMin	NBRSCROLLSPERMINUTE
AvgScrollDur	AVGSCROLLDURATION
AvgTimeBetScrls	AVGTIMEBETWEENSCROLLS
FirstScroll	FIRSTSCROLL
Fixations	NUMBEROFFIXATIONS
AvgFixTime	AVGFIXATIONTIME
FixationsPerMin	NUMBEROFFIXATIONSPERMINUTE

Table 4-12 Argument names used to represent user activity indicators in the All Variables instantiation

A hierarchical approach for conflicts using undercutting attacks was planned:

- Arguments from a single indicator would attack arguments from only one other indicator.
- LowMentalWorkload Arguments would attack MedMentalWorkload and HighMentalWorkload arguments.

- MedMentalWorkload Arguments would attack LowMentalWorkload and HighMentalWorkload arguments.
- HighMentalWorkload Arguments would attack MedMentalWorkload and LowMentalWorkload arguments.

The attacks used as conflicts were placed in layers as per the illustration in Figure 4.1 with the outermost layers attacking inward from Layer 5 towards the Core.

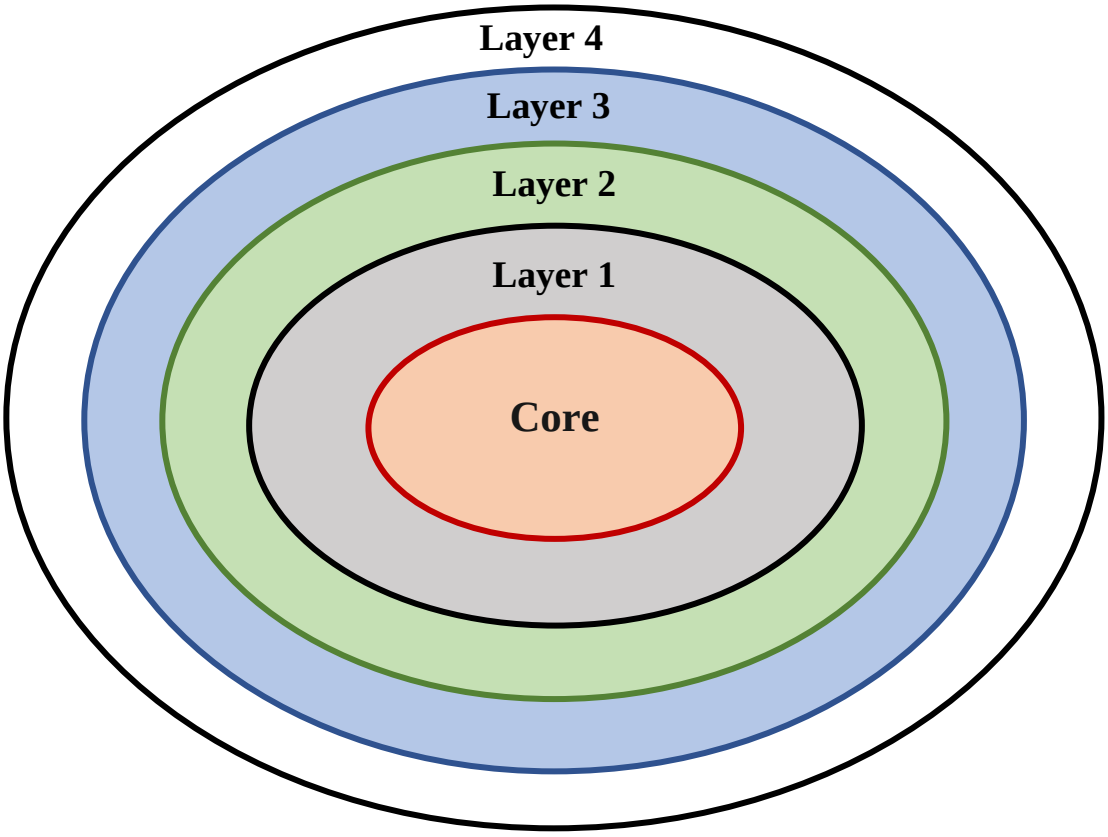

Figure 4.1 Illustration of conflict layers used in all variable instantiation

The core argument chosen to be attacked was TotalMins. The path of all the argument attacks planned are listed in Table 4-13. Each additional attack adds an extra layer.

Full Attack Path
AvgTimeSameArea attacks TotalMins
DistTravPerMin attacks DistTrav attacks TotalMins
NumKeysPerMin attacks NumKeys attacks TotalMins
FirstVdClk attacks VdClks attacks TotalMins
AvgTimeBetVdClks attacks VdClksPerMin attacks VdClks attacks TotalMins
FirstClick attacks NumClks attacks TotalMins
AvgTimeBetClks attacks NumClksPerMin attacks NumClks attacks TotalMins
FixationsPerMin attacks AvgFixTime attacks Fixations attacks TotalMins
FirstScroll attacks NumScrolls attacks TotalMins
AvgTimeBetScrls attacks AvgScrollDur attacks NumScrollsPerMin attacks NumScrolls attacks TotalMins

Table 4-13 The argument attack paths used in the All Variables instantiation

4.5 Implementation of the instantiations of the 5-layer schema

This shows the implemented argumentation system structures used for the instantiations of the 5-layer schema and provides some basic characterisation details for each of them.

4.5.1 Single variable graph

The graph structure created for the single variable instantiation of the 5-layer schema can be seen in Figure 4.2.

Figure 4.2 Argumentation system structure created for the single variable instantiation of the 5-layer schema

The graph has one set of user activity indicator arguments representing:

- DistTravPerMin which represents the distance the mouse travelled per minute during an online web task

There are no argument attacks in this graph.

4.5.2 Two-variable graph

The graph structure created for the two variable instantiation of the 5-layer schema can be seen in Figure 4.3.

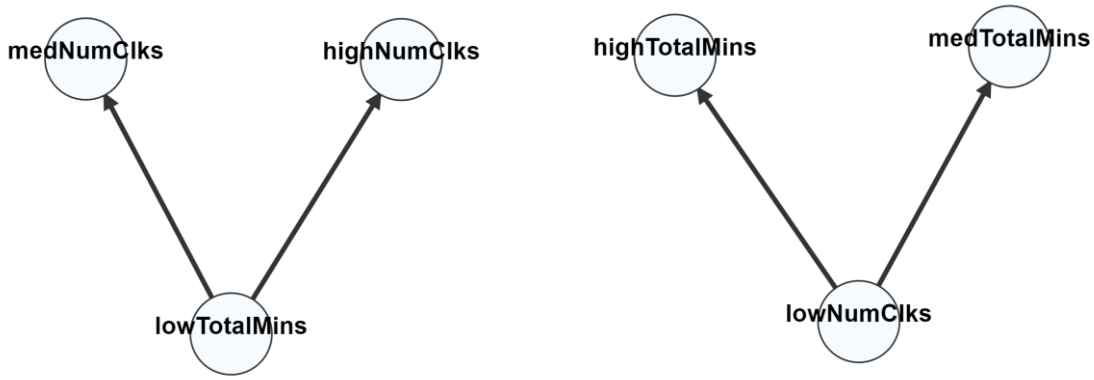

Figure 4.3 Argumentation system structure created for the two variable instantiation of the 5-layer schema

The graph has two sets of user activity indicator arguments representing:

- NumClks which represents the number of clicks a user made during an online web task
- TotalMins which represents the time it took a user to complete an online web task

The low-level arguments from each set is seen attacking the high and med level arguments from the other set.

4.5.3 Multi-variable graph

The graph structure created for the multi variable instantiation of the 5-layer schema can be seen in Figure 4.4.

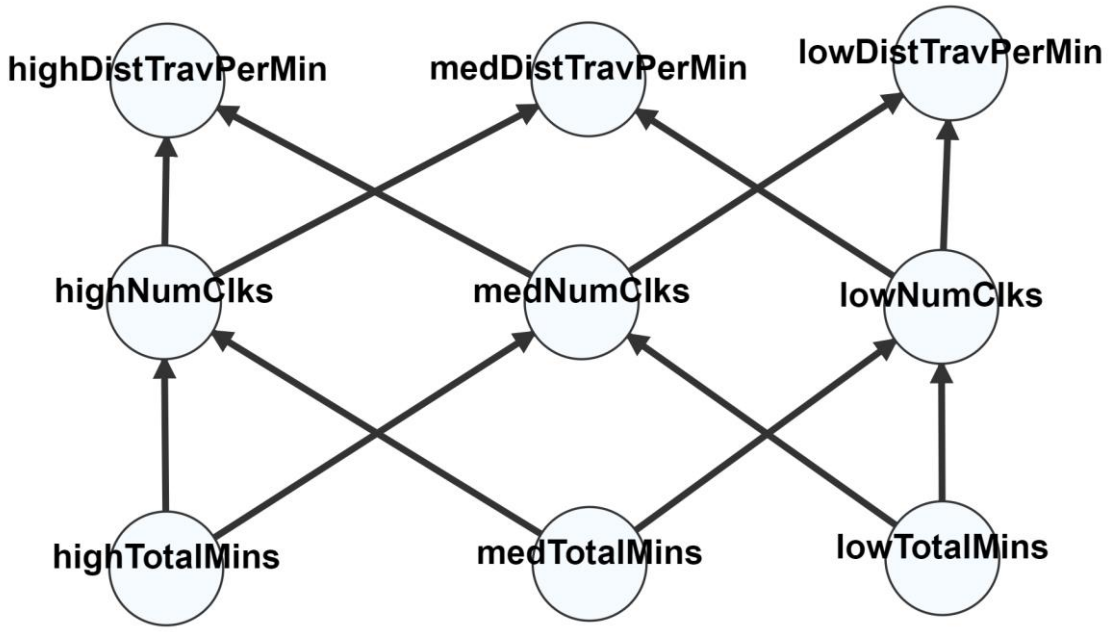

Figure 4.4 Argumentation system structure created for the multi variable instantiation of the 5-layer schema

The graph has three sets of user activity indicator arguments representing:

- The DistTravPerMin argument detailed for the Single variable graph
- The NumClks and TotalMins arguments detailed for the Two-variable graph

The arguments can be seen attacking as follows:

- TotalMins arguments are attacking NumClks arguments
- NumClks arguments are attacking DistTravPerMin arguments
- low level arguments are attacking low and med level arguments
- med level arguments are attacking low and high-level arguments
- high level arguments are attacking high and med level arguments

The attacks are based on the underlying MWL associations within the arguments as defined in the section on the Multi-variable instantiation of the 5-layer schema.

4.5.4 Multi-variable graph with rebuttal attacks

The graph structure created for the multi variable instantiation of the 5-layer schema with rebuttal attacks can be seen in Figure 4.5.

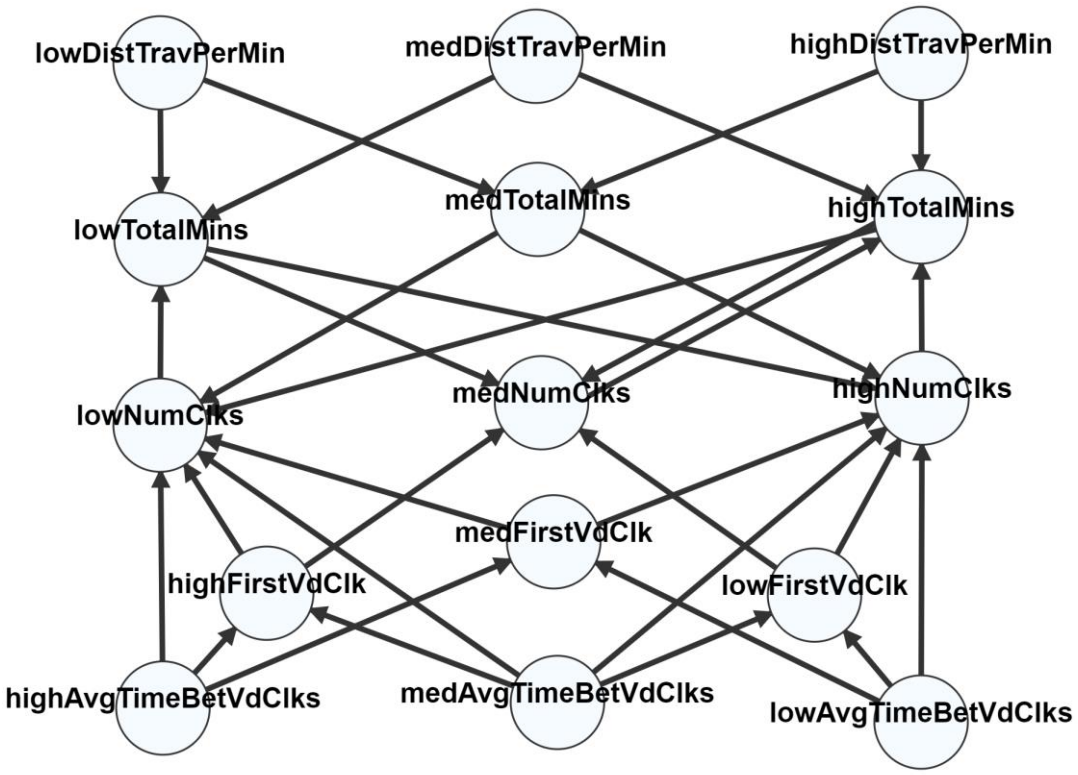

Figure 4.5 Argumentation system structure created for the multi variable with rebuttal attacks instantiation of the 5-layer schema

The graph has three sets of user activity indicator arguments representing:

- The DistTravPerMin argument detailed for the Single variable graph
- The NumClks and TotalMins arguments detailed for the Two-variable graph
- FirstVdClk which represents the first time a void click (a click not associated with a button or a link) was made during an online web task
- AvgTimeBetVdClks which represents the average time between void clicks

The attack arguments can be understood similar to those seen in the Two-variable graph and the Multi-variable graph. Rebuttal attacks can be seen between medNumClks argument and highTotalMins argument.

4.5.5 All-variable graphs

The graph structures created for the NASA-TLX all variable instantiation of the 5-layer schema can be seen in in Figure 4.6.

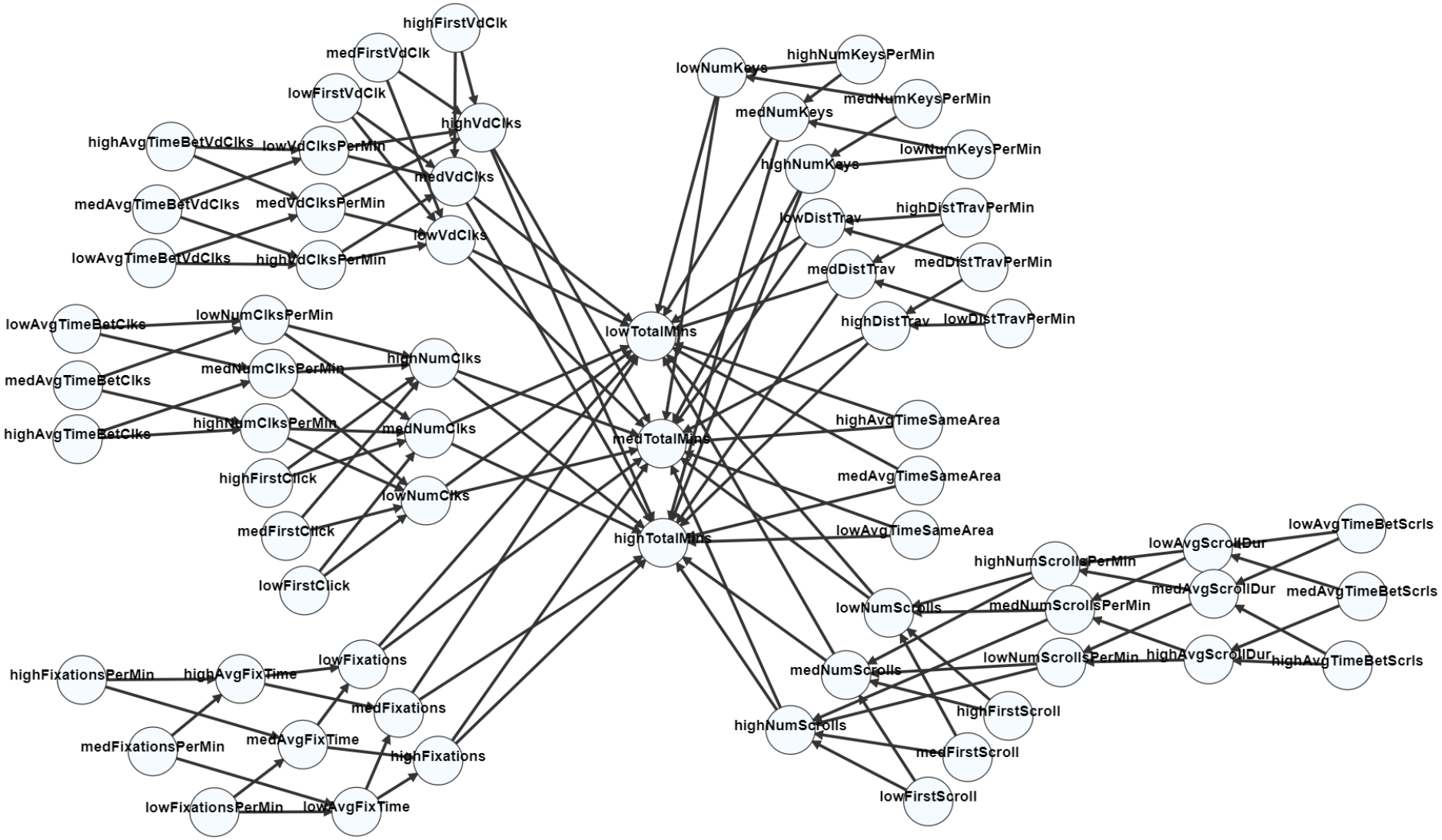

Figure 4.6 Argumentation system structure created for the NASA-TLX all variable experiment

The graph has 22 sets of user activity indicator arguments representing:

- The DistTravPerMin argument detailed for the Single variable graph
- The NumClks and TotalMins arguments detailed for the Two-variable graph
- FirstVdClk and AvgTimeBetVdClks arguments detailed for the Multi-variable instantiation of the 5-layer schema with rebuttal attacks
- Arguments representing all the remaining user activity indicators. The argument names used to represent user activity indicators can be seen in Table 4-12

The attack arguments can be understood similar to those seen in the Two-variable graph and the Multi-variable graph. They are defined in Table 4-10

The graph structures created for the WP all variable instantiation of the 5-layer schema can be seen in Figure 4.7.

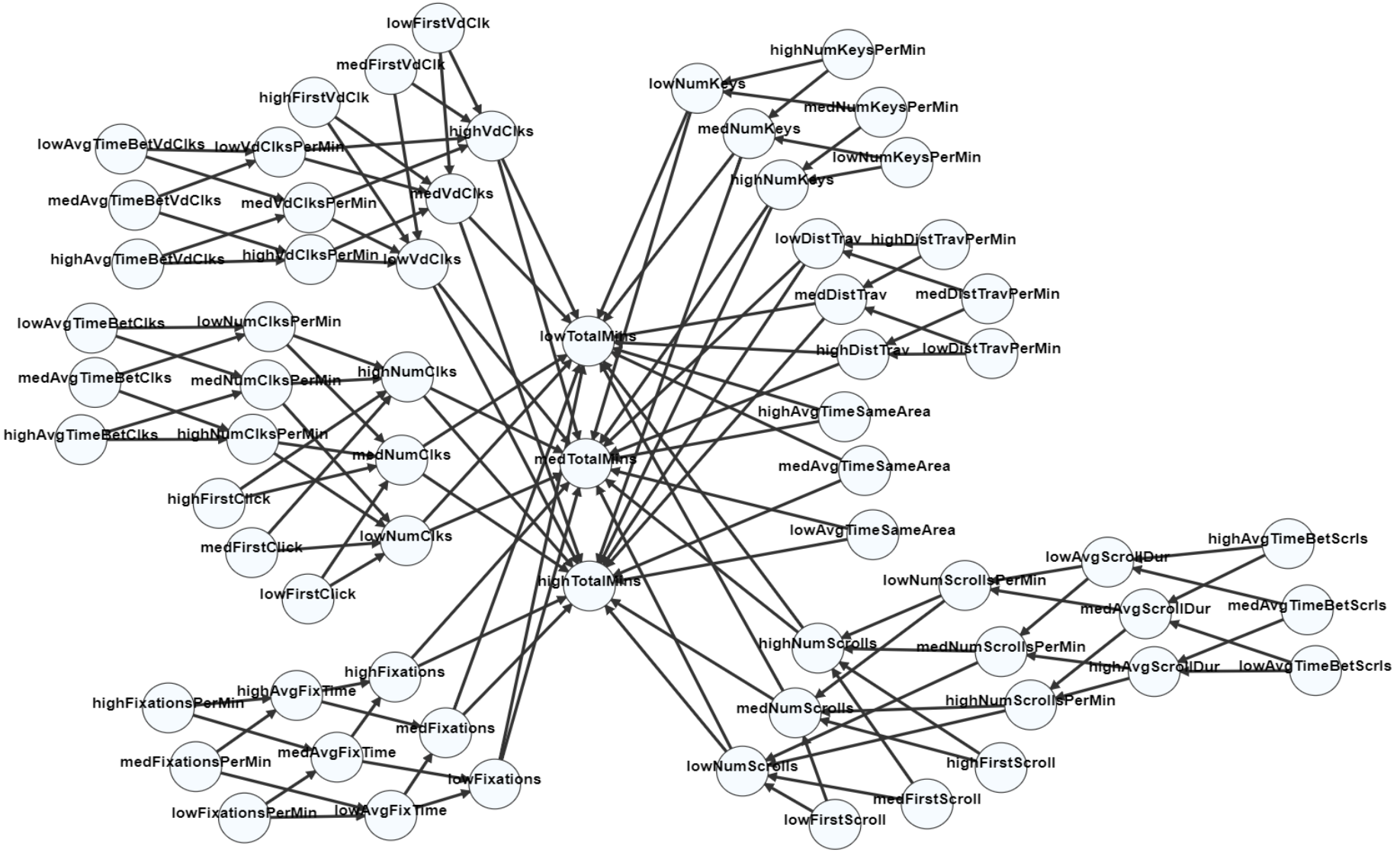

Figure 4.7 Argumentation system structure created for the WP all variable experiment

The graph has 22 sets of user activity indicator arguments representing:

- The DistTravPerMin argument detailed for the Single variable graph
- The NumClks and TotalMins arguments detailed for the Two-variable graph
- FirstVdClk and AvgTimeBetVdClks arguments detailed for the Multi-variable instantiation of the 5-layer schema with rebuttal attacks
- Arguments representing all the remaining user activity indicators. The argument names used to represent user activity indicators can be seen in Table 4-12

The attack arguments can be understood similar to those seen in the Two-variable graph and the Multi-variable graph. They are defined in.Table 4-11

4.6 Experimentation Results

This section shows the results of the face validity and convergent validity tests. The data produced using the argumentation system along with the NASA-TLX and WP MWL measures which were used as input to analyse can be seen in APPENDIX A.

4.6.1 Single variable instantiation of the 5-layer schema

A correlation matrix plotted for the single variable instantiation of the 5-layer schema can be seen in Figure 4.8.

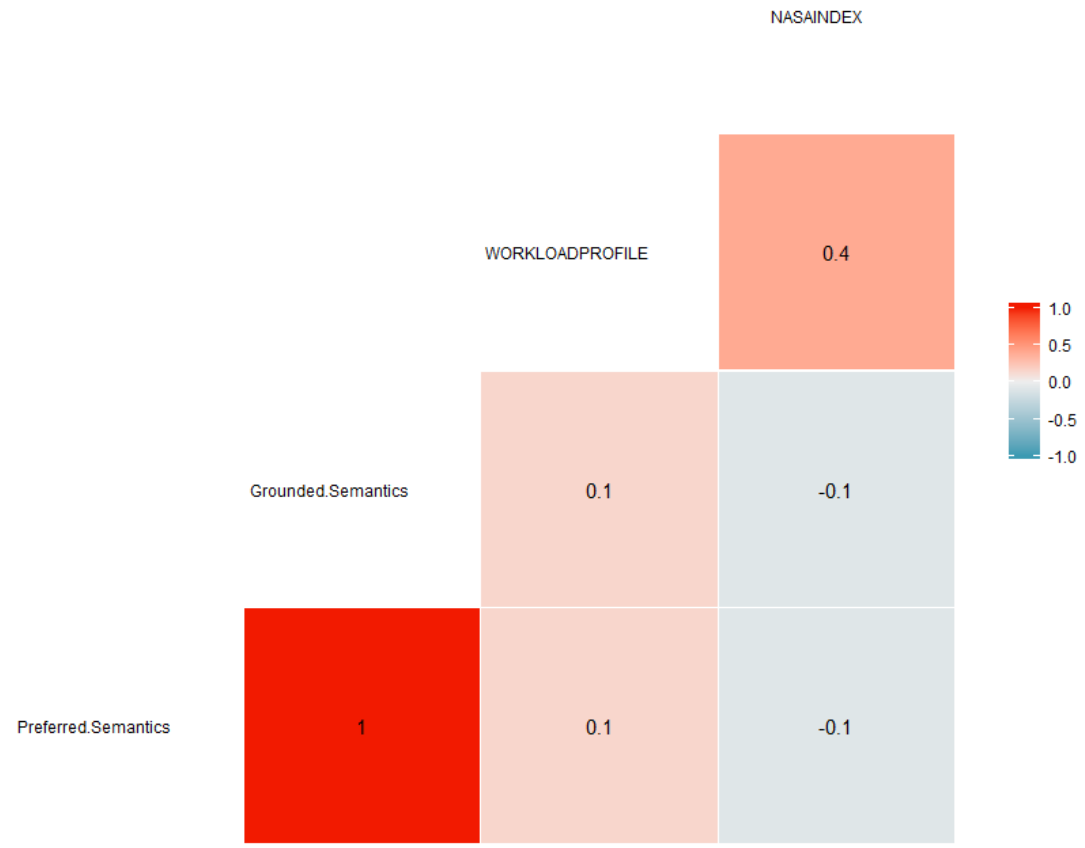

Figure 4.8 Correlation matrix plotted for the single variable instantiation of the 5-layer schema

- The correlation coefficients between the Grounded and Preferred measures with the NASA-TLX MWL measures are both -0.1.
- The correlation coefficients between the Grounded and Preferred measures with the WP MWL measures are both 0.1.

4.6.2 Two-variable instantiation of the 5-layer schema

A correlation matrix plotted for the two variable instantiation of the 5-layer schema can be seen in Figure 4.9.

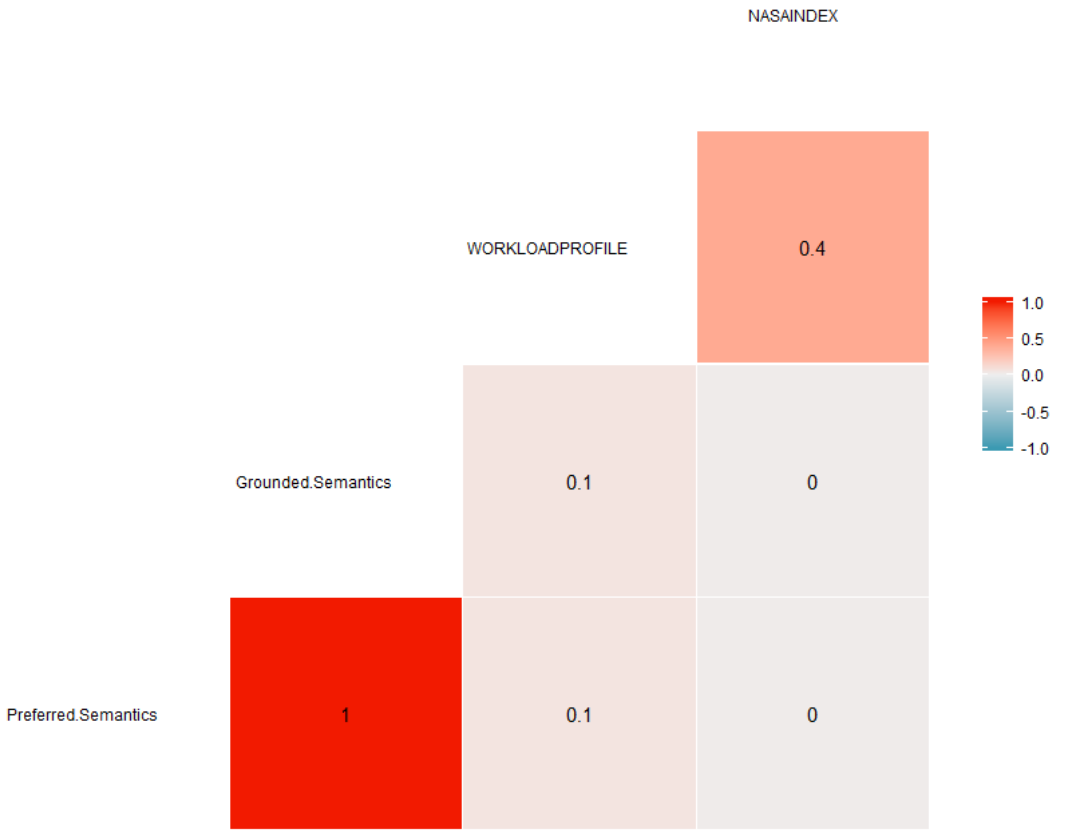

Figure 4.9 Correlation matrix plotted for the two variable instantiation of the 5-layer schema

- The correlation coefficients between the Grounded and Preferred measures with the NASA-TLX MWL measures are both 0.
- The correlation coefficients between the Grounded and Preferred measures with the WP MWL measures are both 0.1.

4.6.3 Multi-variable instantiation of the 5-layer schema

A correlation matrix plotted for the multi variable instantiation of the 5-layer schema can be seen in Figure 4.10.

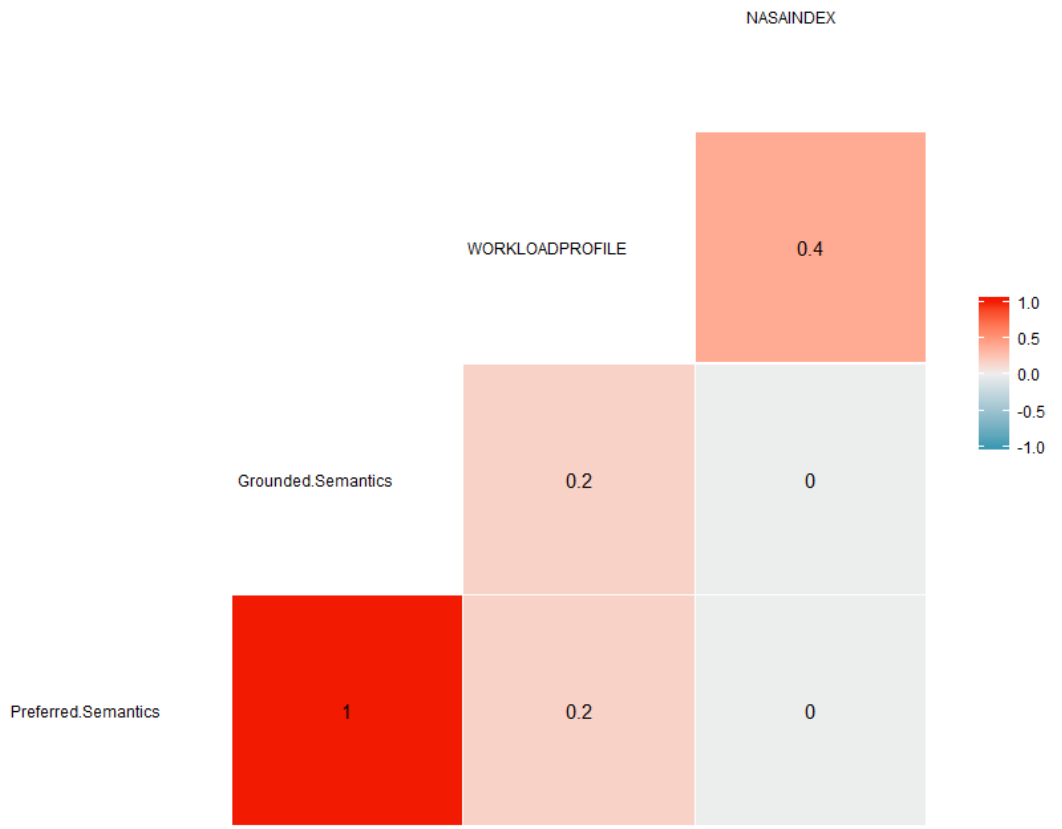

Figure 4.10 Correlation matrix plotted for the multi-variable instantiation of the 5-layer schema

- The correlation coefficients between the Grounded and Preferred measures with the NASA-TLX MWL measures are both 0.
- The correlation coefficients between the Grounded and Preferred measures with the WP MWL measures are both 0.2.

4.6.4 Multi-variable instantiation of the 5-layer schema with rebuttal attacks

A correlation matrix plotted for the multi variable instantiation of the 5-layer schema with rebuttal attacks can be seen in Figure 4.11.

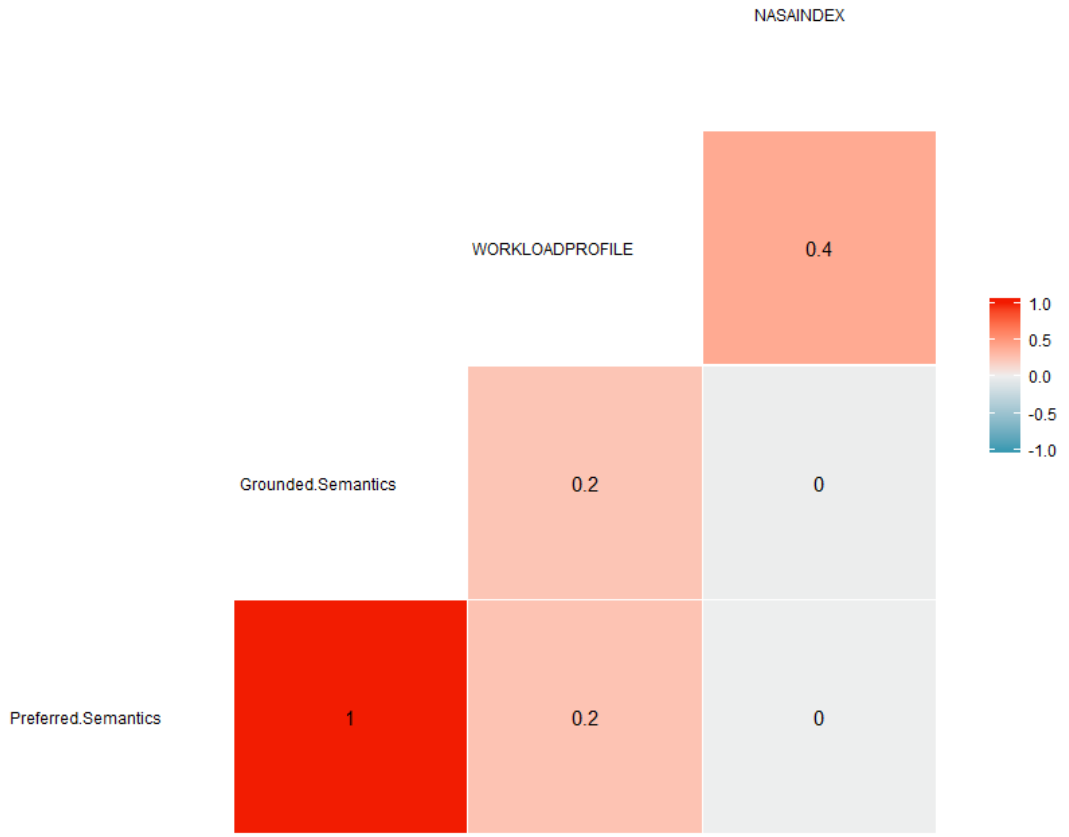

Figure 4.11 Correlation matrix plotted for the multi-variable with rebuttal attacks instantiation of the 5-layer schema

- The correlation coefficients between the Grounded and Preferred measures with the NASA-TLX MWL measures are both 0.
- The correlation coefficients between the Grounded and Preferred measures with the WP MWL measures are both 0.2.

4.6.5 All-variable instantiations of the 5-layer schema

Correlation matrixes were plotted for both sets of results of the all variable instantiations of the 5-layer schema.

The first for NASA-TLX can be seen in Figure 4.12. The correlation coefficients are either 0.1 or 0.3.

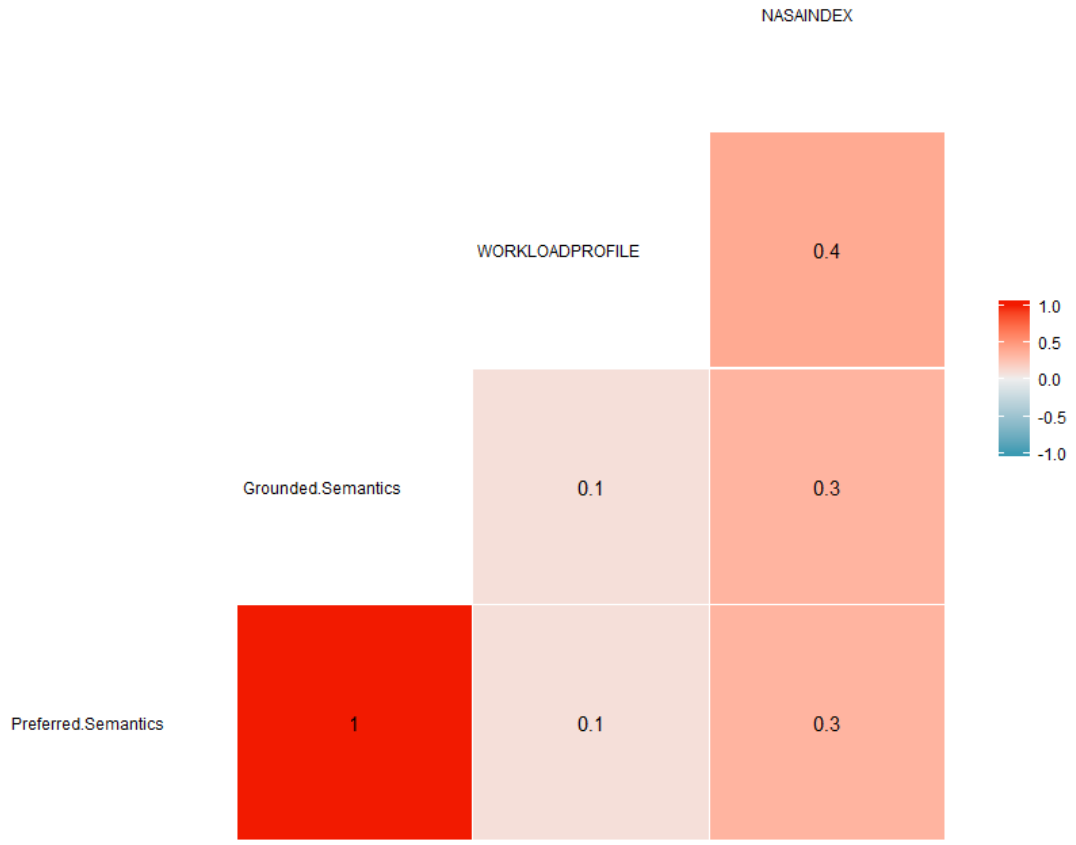

Figure 4.12 Correlation matrix plotted for the NASA-TLX all variable instantiation of the 5-layer schema

- The correlation coefficients between the Grounded and Preferred measures with the NASA-TLX MWL measures are both 0.3.
- The correlation coefficients between the Grounded and Preferred measures with the WP MWL measures are both 0.1.

As the correlation between the Grounded and Preferred measures with the NASA-TLX MWL measures was 0.3 and the correlation between the NASA-TLX MWL measures and WP MWL measures was 0.4, further inspection of the results was carried out. The output of the R function to generate a Spearman correlation matrix is seen Figure 4.13.

	Preferred.Semantics	Grounded.Semantics	WORKLOADPROFILE	NASAINDEX
Preferred.Semantics	1.00000000	1.00000000	0.08515955	0.3314765
Grounded.Semantics	1.00000000	1.00000000	0.08515955	0.3314765
WORKLOADPROFILE	0.08515955	0.08515955	1.00000000	0.3897744
NASAINDEX	0.33147647	0.33147647	0.38977442	1.0000000

Figure 4.13 Spearman correlation matrix for the NASA-TLX all variable instantiation of the 5-layer schema

As the Grounded and Preferred semantics correlation have the same result, a single further correlation test was carried out. The output of the R function to test the Spearman correlation between the NASA-TLX MWL measures and the Preferred measures of is seen in Figure 4.14.

```
        Spearman's rank correlation rho

data:  exp05NASA_redo$NASAINDEX and exp05NASA_redo$Preferred.Semantics
S = 12317, p-value = 0.02136
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3314765
```

Figure 4.14 Correlation between the NASA-TLX MWL measures and the Preferred Semantics measures in the NASA-TLX all variable experiment

For reference, the output of the R function to test the Spearman correlation between the NASA-TLX MWL measures and the WP MWL measures is seen in Figure 4.15.

```
        Spearman's rank correlation rho

data:  exp05NASA_redo$WORKLOADPROFILE and exp05NASA_redo$NASAINDEX
S = 11243, p-value = 0.006172
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3897744
```

Figure 4.15 Correlation between the NASA-TLX MWL measures and the WP MWL measures in the NASA-TLX all variable instantiation of the 5-layer schema

The second WP A correlation matrix plotted for the multi variable instantiation of the 5-layer schema with rebuttal attacks can be seen in Figure 4.16.

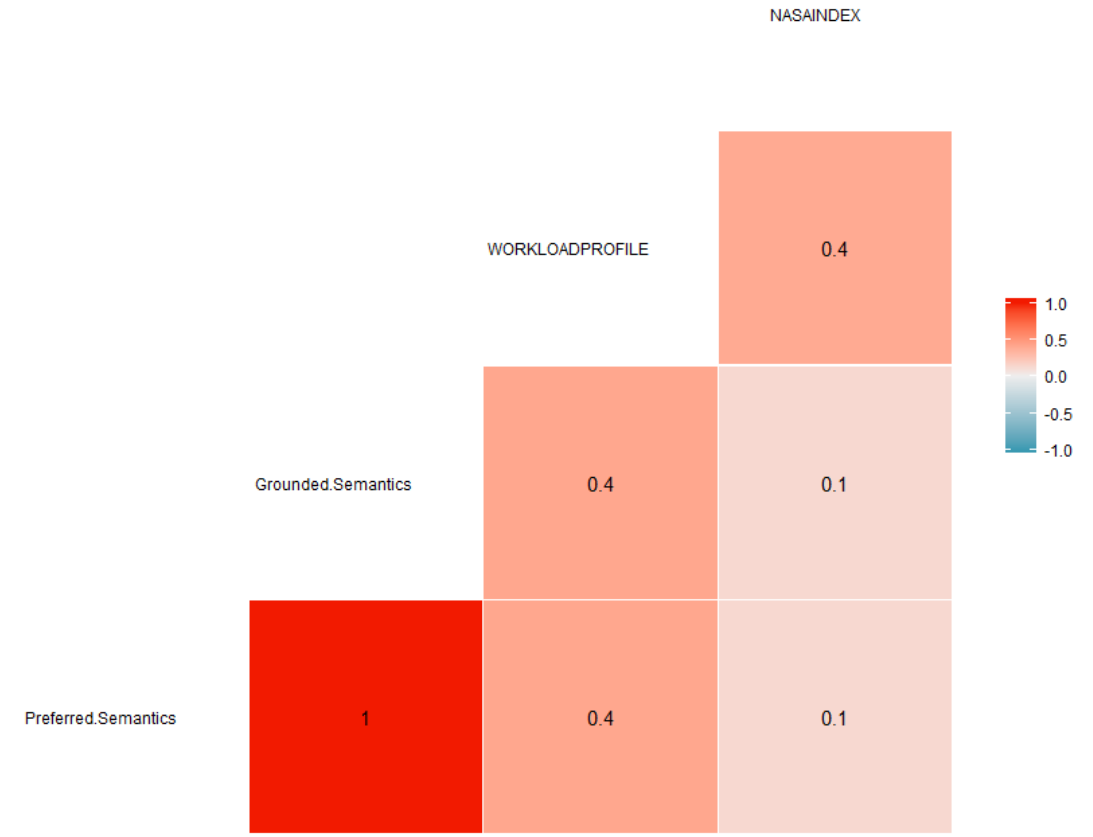

Figure 4.16 Correlation matrix plotted for the WP all variable instantiation of the 5-layer schema

- The correlation coefficients between the Grounded and Preferred measures with the NASA-TLX MWL measures are both 0.1.
- The correlation coefficients between the Grounded and Preferred measures with the WP MWL measures are both 0.4.

As the correlation between the Grounded and Preferred measures with the WP MWL measures was 0.4 and the correlation between the NASA-TLX MWL measures and WP MWL measures was 0.4, further inspection of the results was carried out. The output of the R function to generate a Spearman correlation matrix is seen in Figure 4.17

	Preferred.Semantics	Grounded.Semantics	WORKLOADPROFILE	NASAINDEX
Preferred.Semantics	1.0000000	1.0000000	0.4062458	0.1260024
Grounded.Semantics	1.0000000	1.0000000	0.4062458	0.1260024
WORKLOADPROFILE	0.4062458	0.4062458	1.0000000	0.3897744
NASAINDEX	0.1260024	0.1260024	0.3897744	1.0000000

Figure 4.17 Spearman correlation matrix for the WP all variable instantiation of the 5-layer schema

As the Grounded and Preferred semantics correlation have the same result, a single further correlation test was carried out. The output of the R function to test the Spearman correlation between the WP MWL measures and the Preferred measures of is seen in Figure 4.18.

```
        Spearman's rank correlation rho

data:  exp05WP_redo$WORKLOADPROFILE and exp05WP_redo$Preferred.Semantics
S = 10939, p-value = 0.004169
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4062458
```

Figure 4.18 Correlation between the WP MWL measures and the Preferred Semantics measures in the WP all variable instantiation of the 5-layer schema

For reference, the output of the R function to test the Spearman correlation between the NASA-TLX MWL measures and the WP MWL measures is seen in Figure 4.19.

```
        Spearman's rank correlation rho

data:  exp05WP_redo$WORKLOADPROFILE and exp05WP_redo$NASAINDEX
S = 11243, p-value = 0.006172
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3897744
```

Figure 4.19 Correlation between the NASA-TLX MWL measures and the WP MWL measures in the WP all variable instantiation of the 5-layer schema

4.7 Evaluation

This section summarises the outcomes of the results in each instantiation of the 5-layer schema. First if the outcome of the face validity tests warranted further investigation, second if the convergent validity correlation tests produced a significant result and finally if the results enabled acceptance of the research hypothesis.

4.7.1 Hypothesis testing

This section looks at the results and evaluates them in the context of the research hypotheses which were created.

The null hypothesis (H0) is: the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly unrelated to baseline NASA-TLX or WP mental workload measures for the task.
The alternative hypothesis (H1): the mental workload represented and inferred from a defeasible reasoning knowledge base created using indicators of user interactions from a web-based task is linearly correlated to baseline NASA-TLX or WP mental workload measures for the task.

4.7.2 Non-All-variable instantiations of the 5-layer schema

Two sets of results

- **For the Single variable instantiation of the 5-layer schema**:Face validity results were only either 0.1 or -0.1 and did not warrant further investigation.

- **For the Two variable instantiation of the 5-layer schema**:Face validity results were only either 0 or 0.1 and did not warrant further investigation.

- **For the Multi-variable instantiation of the 5-layer schema**: Face validity results were only either 0 or 0.2 and did not warrant further investigation

- **For the Multi-variable instantiation of the 5-layer schema with rebuttal attacks**: Face validity results were only either 0 or 0.2 and did not warrant further investigation

4.7.3 All-variable instantiations of the 5-layer schema

Two sets of results, one for NASA-TLX and one for WP, were tested.

- **NASA-TLX Result Set**

  Face validity results were 0.3 for NASA-TLX measures but only 0.1 for WP measures and further investigation of the NASA-TLX measures was warranted.

  A convergent validity correlation test showed a correlation coefficient of 0.3314765 with a p-value of 0.02136 which is less than a significance level of

0.05. This was compared to the NASA-TLX MWL measures and the WP MWL measures reference a correlation coefficient of 0.3897744 with a p-value of 0.006172.

As the correlation coefficient was less than 0.40, this test did not result in acceptance of the research hypothesis.

- **WP Result Set**

  Face validity results were only 0.1 for NASA-TLX measures but were 0.4 for WP measures and further investigation of the WP measures was warranted.

  A convergent validity correlation test showed a correlation coefficient of 0.4062458 with a p-value of 0.004169 which is less than a significance level of 0.05. This was compared to the NASA-TLX MWL measures and the WP MWL measures reference a correlation coefficient of 0.3897744 with a p-value of 0.006172.

  At 0.4062458, the correlation coefficient was less than both the moderate correlation coefficient criteria level of 0.5 but it was greater than the. 0.3897744 correlation coefficient for the NASA-TLX MWL measures and the WP MWL measures.

  As this was the greatest correlation within the results call for rejection of the research hypothesis.

4.7.4 Accepting or rejecting H1

Following analysis of the evaluation of the results, particularly that the greatest correlation within the results was less than a moderate to high correlation, then the .null hypothesis is accepted and the alternative hypothesis is rejected.

4.8 Conclusion: Strength and Limitations of Findings

In summary, the main finding was that acceptance of the research hypothesis was possible as the correlation coefficient of 0.4062458 between the Grounded and Preferred measures with the WP MWL measures was statistically significant and:

- not contained within the range of moderate (greater than 0.5) to high (greater than 0.7) correlation coefficient acceptance criteria set out.
- Greater than the. Reference 0.3897744 correlation coefficient for the NASA-TLX MWL measures and the WP MWL measures.

Another noteworthy finding was that, using the all variable NASA-TLX result set, the correlation coefficient between the Grounded and Preferred measures with the NASA-TLX MWL measures of 0.3314765 was both statistically significant and close to the reference correlation coefficient of 0.3897744 calculated for that set.

None of the other instantiations had statistically significant results. The single and two variable instantiations did not have a positive correlation coefficient above 0.1 or negative correlation coefficient below -0.1. Both multi variable instantiations did not have a positive correlation coefficient above 0.2 and only the WP MWL measures showed promise.

Finally, it should be noted that dataset was refined from the original 109 rows to 48 rows. This in turn reduced the reference correlation coefficient between NASA-TLX and WP from around 0.6 to around 0.4.

5 CONCLUSION

This chapter provides a review of the research outlined in this document, identifies how the results contribute to the scientific body of knowledge and proposes potential future directions for research in this area.

5.1 Introduction

This research investigated the potential of using defeasible reasoning to infer the MWL associated with performing online web tasks.

MWL can be a useful measure to assist in optimising UX designs and understanding how user performance can be affected. However, measuring MWL has proved difficult as solutions had issues such as difficulties scaling in size, being costly or not being performed in real time. Recent work looked at measuring MWL using indicators of user activity recorded during the performance of web-based tasks. Correlation was found within subsets of results and study of the interaction of multiple user activity indicators was identified as potential future direction of the work.

Defeasible reasoning has been used to investigate the representation of MWL as a defeasible concept where one reason can defeat another. This had favourable results and used an argumentation system to describe MWL as a defeasible concept through interactions of defeasible arguments along with MWL survey data.

For this research a series of instantiations of the 5-layer schema were used to explore the use of user activity indicators with the argumentation system to infer a value of MWL.

5.2 Research Overview

This section gives a brief overview of the Literature Review, Design and Implementation and Results and Evaluation and Conclusion chapters in this research document:

- **Chapter 2** gave details from the **Literature Review** performed in the fields of Human-Computer Interaction, Design and Usability, Mental Workload, User Interaction Tracking, User Interaction Indicators and Defeasible Reasoning and Argumentation Theory which are relevant to this research.

- **Chapter 3** gave details the **Design and Implementation** of a solution and series of experiments to test the research hypothesis.

- **Chapter 4** discussed the **Results and Evaluation** of the experiment to test the research hypothesis including the results and associated results data with the aim of analysing them and check for correlations between inferred MWL and subjective measures already in the dataset.

- **Chapter 5** concluded the document by presenting the findings of the research, the contribution made to the body of knowledge, potential future areas of research and details the conclusions found from the research.

5.3 Experimentation, Evaluation and Results

An original 109 row dataset was obtained but refined to 48 rows to remove those with blank data. It was noted that the reference correlation coefficient between WP MWL values and NASA-TLX values reduced from around 0.6 to around 0.4 after the dataset was refined.

Six instantiations of the 5-layer schema, and their associated rules (knowledge-bases), were created out using one, two, multiple and all variables representing the user activity indicators available in the dataset. Six structures (two in the all variable instantiations) were created in the argumentation system and the refined dataset used with them. The results were correlated to reference NASA-TLX and WP measures of MWL.

The correlations found were significant using the two all variable instantiations structures. The first was orientated to NASA-TLX MWL values and the second to WP MWL values. The NASA-TLX orientated structure correlation was not above the criteria to accept the research hypothesis. Similarly, the WP orientated structure correlation was not above the criteria to accept the research hypothesis. The values it generated were stronger than a reference correlation between the NASA-TLX MWL and WP MWL values which was used.

The correlations for the other experimental results were not statistically significant. For those instantiations, it was noted that the correlations of the WP MWL values were stronger than those of the NASA-TLX values.

5.4 Contributions to the Body of Knowledge

This research has not demonstrated that defeasible reasoning can be used to infer MWL of web-based tasks. The following contributions to the body of knowledge have been made:

- Greater amounts of user activity indicators were required to produce a statistically significant result.
- Arguments which represented the indicators and attacks between them were used in an online argumentation system in five instantiations of the 5-layer schema. A layered hierarchical approach to creating and attacking arguments was used for creating the structures within the experiment which produced statistically significant results.
- WP MWL values appear to correlate stronger with results than those of the NASA-TLX values

5.5 Future Work & Research

There are multiple potential areas to explore in the future:

- A step forward for this research would be to perform an online web-based user experiment where MWL inferred using defeasible reasoning is used to change the design of user interfaces to make them more or less complex depending on which outcome would be optimal for each individual participant.
- Another step forward for this research would be further investigation of defeasible reasoning in order to create more complex argumentation system structures to produce stronger correlations to reference MWL values.
- One limitation of this research arose with the dataset and how it was refined from 109 rows to 48. Future work could reproduce the original experiment or perform alternative experiments to create larger datasets which ensure to record all the user activity indicators. These datasets could be used to replicate the findings of this research.

5.6 Conclusion

No moderate or high strength correlation of reference MWL values to values inferred using the argumentation system was found. However, the results showed promise as correlation level did appear to be rising and becoming significant as more indicators of user activity were included.

Altogether this indicates that the research hypothesis should be rejected as the mental workload inferred from a defeasible reasoning knowledge base using indicators of user interactions from a web-based task was less than moderately correlated to baseline NASA-TLX and WP mental workload measures for the task.

This indicates that the mental workload of web-based tasks has not yet been shown to be inferred implicitly from user activity using defeasible reasoning in a multi-layer argument-based framework, built upon argumentation theory and that the research question cannot yet be answered affirmatively.

BIBLIOGRAPHY

Albers, M. J. (2011). Tapping as a measure of cognitive load and website usability. In *Proceedings of the 29th ACM international conference on Design of communication* (pp. 25–32). ACM.

Arapakis, I., & Leiva, L. A. (2016). Predicting User Engagement with Direct Displays Using Mouse Cursor Information. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 599–608). New York, NY, USA: ACM. https://doi.org/10.1145/2911451.2911505

Atterer, R., Wnuk, M., & Schmidt, A. (2006). Knowing the User's Every Move: User Activity Tracking for Website Usability Evaluation and Implicit Interaction. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 203–212). New York, NY, USA: ACM. https://doi.org/10.1145/1135777.1135811

Balogh, J., Cohen, M., & Giangola, J. P. (2004). Voice user interface design: Minimizing cognitive load. *InformIT. May*, *13*, 2004.

Baroni, P., Caminada, M., & Giacomin, M. (2011). An introduction to argumentation semantics. *Knowledge Eng. Review*, *26*, 365–410. https://doi.org/10.1017/S0269888911000166

Baroni, P., Guida, G., & Mussi, S. (1997). Full nonmonotonicity: a new perspective in defeasible reasoning. In *ESIT 97* (pp. 58–62). Bari, Italy.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Cain, B. (2007). *A review of the mental workload literature*. Defence Research And Development Toronto (Canada). Retrieved from http://www.dtic.mil/docs/citations/ADA474193

Chao, G. (2009). Human-Computer Interaction: Process and Principles of Human-Computer Interface Design. In *2009 International Conference on Computer and Automation Engineering* (pp. 230–233). https://doi.org/10.1109/ICCAE.2009.23

Contreras, L. (2018). An Application of Natural Language Processing for Triangulation of Cognitive Load Assessments in Third Level Education. *Dissertations*. Retrieved from https://arrow.dit.ie/scschcomdis/112

Dawe, M. (2006). Desperately Seeking Simplicity: How Young Adults with Cognitive Disabilities and Their Families Adopt Assistive Technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1143–1152). New York, NY, USA: ACM. https://doi.org/10.1145/1124772.1124943

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/10.3758/s13428-015-0567-2

de Santana, V. F., & Baranauskas, M. C. C. (2015). WELFIT: A remote evaluation tool for identifying Web usage patterns through client-side logging. *International Journal of Human-Computer Studies*, *76*, 40–49. https://doi.org/10.1016/j.ijhcs.2014.12.005

Ford, M., & Billington, D. (2000). Strategies in Human Nonmonotonic Reasoning. *Computational Intelligence*, *16*(3), 446–468. https://doi.org/10.1111/0824-7935.00119

Guo, Q., & Agichtein, E. (2012). Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and Other Post-click Searcher Behavior. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 569–578). New York, NY, USA: ACM. https://doi.org/10.1145/2187836.2187914

Guo, Q., Lagun, D., & Agichtein, E. (2012). Predicting Web Search Success with Fine-grained Interaction Data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2050–2054). New York, NY, USA: ACM. https://doi.org/10.1145/2396761.2398570

Gwizdka, J. (2010). Distribution of Cognitive Load in Web Search. *J. Am. Soc. Inf. Sci. Technol.*, *61*(11), 2167–2187. https://doi.org/10.1002/asi.v61:11

Hancock, P. A., & Caird, J. K. (1993). Experimental Evaluation of a Model of Mental Workload. *Human Factors*, *35*(3), 413–429. https://doi.org/10.1177/001872089303500303

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hartmann, J., Sutcliffe, A., & Angeli, A. D. (2008). Towards a Theory of User Judgment of Aesthetics and User Interface Quality. *ACM Trans. Comput.-Hum. Interact.*, *15*(4), 15:1–15:30. https://doi.org/10.1145/1460355.1460357

Hornbæk, K., & Hertzum, M. (2017). Technology Acceptance and User Experience: A Review of the Experiential Component in HCI. *ACM Trans. Comput.-Hum. Interact.*, *24*(5), 33:1–33:30. https://doi.org/10.1145/3127358

Huang, J., White, R. W., & Dumais, S. (2011). No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1225–1234). ACM.

Karray, F., Alemzadeh, M., Abou Saleh, J., & Nours Arab, M. (2008). Human-Computer Interaction: Overview on State of the Art. *International Journal on Smart Sensing and Intelligent Systems*, *1*(1), 137–159. https://doi.org/10.21307/ijssis-2017-283

Loft, S., Sanderson, P., Neal, A., & Mooij, M. (2007). Modeling and Predicting Mental Workload in En Route Air Traffic Control: Critical Review and Broader Implications. *Human Factors*, *49*(3), 376–399. https://doi.org/10.1518/001872007X197017

Longo, L. (2015). Designing Medical Interactive Systems Via Assessment of Human Mental Workload. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems* (pp. 364–365). https://doi.org/10.1109/CBMS.2015.67

Longo, L., & Dondio, P. (2014). Defeasible Reasoning and Argument-Based Systems in Medical Fields: An Informal Overview. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems* (pp. 376–381). https://doi.org/10.1109/CBMS.2014.126

Longo, L., & Dondio, P. (2015). On the Relationship between Perception of Usability and Subjective Mental Workload of Web Interfaces. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Vol. 1, pp. 345–352). https://doi.org/10.1109/WI-IAT.2015.157

Longo, L., Kane, B., & Hederman, L. (2012). Argumentation theory in health care. In *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 1–6). https://doi.org/10.1109/CBMS.2012.6266323

Longo, Luca. (2011). Human-computer Interaction and Human Mental Workload: Assessing Cognitive Engagement in the World Wide Web. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part IV* (pp. 402–405). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=2042283.2042335

Longo, Luca. (2014). *Formalising Human Mental Workload as a Defeasible Computational Concept*.

Longo, Luca. (2015a). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology*, *34*(8), 758–786. https://doi.org/10.1080/0144929X.2015.1015166

Longo, Luca. (2015b). Designing Medical Interactive Systems Via Assessment of Human Mental Workload. In *Proceedings of the 2015 IEEE 28th International Symposium on Computer-Based Medical Systems* (pp. 364–365). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/CBMS.2015.67

Longo, Luca. (2016a). Argumentation for Knowledge Representation, Conflict Resolution, Defeasible Inference and Its Integration with Machine Learning. In *Machine Learning for Health Informatics* (pp. 183–208). Springer, Cham. https://doi.org/10.1007/978-3-319-50478-0_9

Longo, Luca. (2016b). Mental workload in medicine: Foundations, applications, open problems, challenges and future perspectives. In *Computer-Based Medical Systems (CBMS), 2016 IEEE 29th International Symposium on* (pp. 106–111). IEEE.

Longo, Luca. (2017). Subjective Usability, Mental Workload Assessments and Their Impact on Objective Human Performance. In R. Bernhaupt, G. Dalvi, A. Joshi,

D. K. Balkrishan, J. O'Neill, & M. Winckler (Eds.), *Human-Computer Interaction - INTERACT 2017* (pp. 202–223). Springer International Publishing.

Longo, Luca, & Hederman, L. (2013). Argumentation Theory for Decision Support in Health-Care: A Comparison with Machine Learning. In K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, & N. Zhong (Eds.), *Brain and Health Informatics* (pp. 168–180). Springer International Publishing.

Longo, Luca, Rusconi, F., Noce, L., & Barrett, S. (2012). The Importance of Human Mental Workload in Web Design. In *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies* (pp. 403–409).

Moustafa, K., Luz, S., & Longo, L. (2017). Assessment of Mental Workload: A Comparison of Machine Learning Methods and Subjective Assessment Techniques. In Luca Longo & M. C. Leva (Eds.), *Human Mental Workload: Models and Applications* (pp. 30–50). Springer International Publishing.

Pantic, M., Nijholt, A., Pentland, A., & Huanag, T. S. (2008). Human-Centred Intelligent Human Computer Interaction (HCI^2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, *1*(2), 168. https://doi.org/10.1504/IJAACS.2008.019799

Phillips, B., & Zhao, H. (1993). Predictors of assistive technology abandonment. *Assistive Technology: The Official Journal of RESNA*, *5*(1), 36–45. https://doi.org/10.1080/10400435.1993.10132205

Pollock, J. L. (1987). Defeasible Reasoning. *Cognitive Science*, *11*(4), 481–518. https://doi.org/10.1207/s15516709cog1104_4

Prakken, H., & Vreeswijk, G. (2001). Logics for Defeasible Argumentation. In *Handbook of Philosophical Logic* (pp. 219–318). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-0456-4_3

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rizzo, L., Dondio, P., Delany, S. J., & Longo, L. (2016). Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile. In *Artificial Intelligence Applications and Innovations* (pp. 215–229). Springer, Cham. https://doi.org/10.1007/978-3-319-44944-9_19

Rizzo, L., & Longo, L. (2017). Representing and inferring mental workload via defeasible reasoning: a comparison with the NASA Task Load Index and the Workload Profile. In *1st Workshop on Advances In Argumentation In Artificial Intelligence*. Bari, Italy. Retrieved from http://aiia2017.di.uniba.it/ai3-2017/papers/paper9.pdf

Romero, J. F. (2017). An Investigation of the Correlation Between Mental Workload and Web User's Interaction. *Dissertation*. Retrieved from http://arrow.dit.ie/scschcomdis/108

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology*, *53*(1), 61–86. https://doi.org/10.1111/j.1464-0597.2004.00161.x

Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., … Larmarange, J. (2018). GGally: Extension to 'ggplot2' (Version 1.4.0) [R Package]. Retrieved from https://CRAN.R-project.org/package=GGally

Shapira, B., Taieb-Maimon, M., & Moskowitz, A. (2006). Study of the Usefulness of Known and New Implicit Indicators and Their Optimal Combination for Accurate Inference of Users Interests. In *Proceedings of the 2006 ACM Symposium on Applied Computing* (pp. 1118–1119). New York, NY, USA: ACM. https://doi.org/10.1145/1141277.1141542

Silva, F. P. da. (2014). Mental Workload, Task Demand and Driving Performance: What Relation? *Procedia - Social and Behavioral Sciences*, *162*(Supplement C), 310–319. https://doi.org/10.1016/j.sbspro.2014.12.212

Tracy, J. P., & Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. In *Annual Conference-society for technical communication* (Vol. 53, p. 256). Retrieved from https://www.researchgate.net/profile/Michael_Albers/publication/253713707_Measuring_Cognitive_Load_to_Test_the_Usability_of_Web_Sites/links/55ef096608ae199d47bff6cd.pdf

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381. https://doi.org/10.1080/00140139608964470

van Steenbergen, H., & Bocanegra, B. R. (2016). Promises and pitfalls of Web-based experimentation in the advance of replicable psychological science: A reply to Plant (2015). *Behavior Research Methods*, *48*(4), 1713–1717. https://doi.org/10.3758/s13428-015-0677-x

Verheij, B. (2009). The Toulmin Argument Model in Artificial Intelligence. In *Argumentation in Artificial Intelligence* (pp. 219–238). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-98197-0_11

Walton, D. (2009). Argumentation Theory: A Very Short Introduction. In *Argumentation in Artificial Intelligence*.

Wästlund, E., Norlander, T., & Archer, T. (2008). The effect of page layout on mental workload: A dual-task experiment. *Computers in Human Behavior*, *24*(3), 1229–1245. https://doi.org/10.1016/j.chb.2007.05.001

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: a tutorial review. *PeerJ*, *3*, e1058. https://doi.org/10.7717/peerj.1058

Wu, Y., & Caminada, M. (2010). A Labelling-Based Justification Status of Arguments, *3*(4), 18.

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & Stress*, *14*(1), 74–99.

Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., & Bailey, P. (2014). Relevance and Effort: An Analysis of Document Utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 91–100). New York, NY, USA: ACM. https://doi.org/10.1145/2661829.2661953

Zhu, H., & Hou, M. (2009). Restrain mental workload with roles in HCI. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)* (pp. 387–392). https://doi.org/10.1109/TIC-STH.2009.5444469

APPENDIX A – ARGUMENTATION SYSTEM RESULTS AND THE NASA-TLX AND WP MWL MEASURES

Single Variable Experiment

Preferred Semantics	Grounded Semantics	WORKLOADPROFILE	NASAINDEX
17.69	17.69	2.7	58
16.71	16.71	3	54
13.74	13.74	4.6	68
29.04	29.04	3.25	55
15.56	15.56	3.25	50
63.07	63.07	2.25	40
9.02	9.02	3.15	43
21.64	21.64	3.05	38
28.39	28.39	2.55	46
30.04	30.04	1.75	17
21.44	21.44	3.8	53
18.47	18.47	2.55	58
36.24	36.24	2.8	71
42.61	42.61	3	50
19.58	19.58	2.95	48
17.03	17.03	3.5	50
16.82	16.82	2.6	49
28.94	28.94	2.25	40
39.41	39.41	4.3	61
38.9	38.9	3.4	33
51.06	51.06	3	53
26.73	26.73	3.65	48
53.28	53.28	1.75	47
100	100	1.35	42
18.25	18.25	2.4	41
80.87	80.87	3.9	37
49.84	49.84	1.2	36
34.86	34.86	2.75	59
32.41	32.41	5.5	75
17.29	17.29	2.5	59
73.45	73.45	4.9	42
70.43	70.43	4.85	59
16.71	16.71	3.05	50
8.77	8.77	1.25	30
43.21	43.21	4.25	45
12	12	2.35	43
42.6	42.6	3.4	57
67.87	67.87	3.5	55
60.75	60.75	4.15	46
49.23	49.23	3	50
50.82	50.82	4.25	47
24.76	24.76	3.6	50
26.78	26.78	2.25	35

42.39	42.39	0.95	45
38.79	38.79	3.6	37
64.22	64.22	1.85	43
12.38	12.38	2.75	41
21.06	21.06	2.1	49

Two Variable Experiment

Preferred Semantics	Grounded Semantics	WORKLOADPROFILE	NASAINDEX
34.27	34.27	2.7	58
85.34	85.34	3	54
79.76	79.76	4.6	68
53.09	53.09	3.25	55
90.93	90.93	3.25	50
27.32	27.32	2.25	40
50.57	50.57	3.15	43
57.89	57.89	3.05	38
45.5	45.5	2.55	46
29.88	29.88	1.75	17
21.07	21.07	3.8	53
60.64	60.64	2.55	58
43.48	43.48	2.8	71
51.86	51.86	3	50
64.24	64.24	2.95	48
42.58	42.58	3.5	50
56.01	56.01	2.6	49
19.78	19.78	2.25	40
48.01	48.01	4.3	61
51	51	3.4	33
9.73	9.73	3	53
49.77	49.77	3.65	48
21.63	21.63	1.75	47
17.51	17.51	1.35	42
58.15	58.15	2.4	41
14.03	14.03	3.9	37
31.37	31.37	1.2	36
44.31	44.31	2.75	59
59.6	59.6	5.5	75
23.05	23.05	2.5	59
40.49	40.49	4.9	42
26.35	26.35	4.85	59
70.38	70.38	3.05	50
55.59	55.59	1.25	30
48.06	48.06	4.25	45
87.73	87.73	2.35	43
23.95	23.95	3.4	57
52.75	52.75	3.5	55
63.18	63.18	4.15	46
44.55	44.55	3	50

17.64	17.64	4.25	47
62.68	62.68	3.6	50
99.7	99.7	2.25	35
57.55	57.55	0.95	45
90.13	90.13	3.6	37
19.88	19.88	1.85	43
52.78	52.78	2.75	41
86.14	86.14	2.1	49

Multi-variable Experiment

Preferred Semantics	Grounded Semantics	WORKLOADPROFILE	NASAINDEX
40.19	40.19	2.7	58
33.35	33.35	3	54
75.91	75.91	4.6	68
26.53	26.53	3.25	55
87.29	87.29	3.25	50
53.87	53.87	2.25	40
19.19	19.19	3.15	43
24.45	24.45	3.05	38
45.5	45.5	2.55	46
21.24	21.24	1.75	17
21.26	21.26	3.8	53
22.82	22.82	2.55	58
21.72	21.72	2.8	71
28.9	28.9	3	50
25.38	25.38	2.95	48
34.6	34.6	3.5	50
20.15	20.15	2.6	49
24.36	24.36	2.25	40
45.14	45.14	4.3	61
25.78	25.78	3.4	33
30.39	30.39	3	53
21.26	21.26	3.65	48
37.46	37.46	1.75	47
58.75	58.75	1.35	42
23.81	23.81	2.4	41
47.45	47.45	3.9	37
40.61	40.61	1.2	36
23.06	23.06	2.75	59
29.95	29.95	5.5	75
20.17	20.17	2.5	59
40.49	40.49	4.9	42
48.39	48.39	4.85	59
58.35	58.35	3.05	50
18.1	18.1	1.25	30
46.45	46.45	4.25	45
77.95	77.95	2.35	43
33.28	33.28	3.4	57

48	48	3.5	55
66.54	66.54	4.15	46
30.89	30.89	3	50
34.23	34.23	4.25	47
28	28	3.6	50
91.45	91.45	2.25	35
32.89	32.89	0.95	45
39.27	39.27	3.6	37
42.05	42.05	1.85	43
15.5	15.5	2.75	41
34.82	34.82	2.1	49

Multi-variable Experiment with rebuttal attacks

Preferred Semantics	Grounded Semantics	WORKLOADPROFILE	NASAINDEX
25.93	25.93	2.7	58
24.85	24.85	3	54
58.04	58.04	4.6	68
63.9	63.9	3.25	55
72.91	72.91	3.25	50
48.08	48.08	2.25	40
49.98	49.98	3.15	43
64.12	64.12	3.05	38
50.29	50.29	2.55	46
43.47	43.47	1.75	17
37.63	37.63	3.8	53
33.5	33.5	2.55	58
61.41	61.41	2.8	71
76.16	76.16	3	50
64.3	64.3	2.95	48
43.04	43.04	3.5	50
26.63	26.63	2.6	49
42.83	42.83	2.25	40
58.26	58.26	4.3	61
67.26	67.26	3.4	33
51.54	51.54	3	53
66.95	66.95	3.65	48
54.12	54.12	1.75	47
63.55	63.55	1.35	42
39.11	39.11	2.4	41
55.2	55.2	3.9	37
64.28	64.28	1.2	36
64.94	64.94	2.75	59
67.67	67.67	5.5	75
42.67	42.67	2.5	59
65.01	65.01	4.9	42
56.78	56.78	4.85	59
42.925	33.78	3.05	50
59.32	59.32	1.25	30

45.34	45.34	4.25	45
38.33	38.33	2.35	43
52.43	52.43	3.4	57
60.03	60.03	3.5	55
65.64	65.64	4.15	46
65.5	65.5	3	50
59.98	59.98	4.25	47
32.57	32.57	3.6	50
36.06	36.06	2.25	35
43.83	43.83	0.95	45
38.66	38.66	3.6	37
68.27	68.27	1.85	43
58.72	58.72	2.75	41
12.73	12.73	2.1	49

All-variable experiment – NASA-TLX

Preferred Semantics	Grounded Semantics	WORKLOADPROFILE	NASAINDEX
59.36	59.36	2.7	58
69.98	69.98	3	54
66.13	66.13	4.6	68
50.21	50.21	3.25	55
78.53	78.53	3.25	50
27.78	27.78	2.25	40
65.1	65.1	3.15	43
59.03	59.03	3.05	38
43.03	43.03	2.55	46
43.32	43.32	1.75	17
51.19	51.19	3.8	53
63.8	63.8	2.55	58
45.49	45.49	2.8	71
56.87	56.87	3	50
59.7	59.7	2.95	48
48.57	48.57	3.5	50
50.58	50.58	2.6	49
45.18	45.18	2.25	40
53.3	53.3	4.3	61
39.1	39.1	3.4	33
41.14	41.14	3	53
55.65	55.65	3.65	48
45.38	45.38	1.75	47
38.84	38.84	1.35	42
60.68	60.68	2.4	41
35.93	35.93	3.9	37
40.57	40.57	1.2	36
43.13	43.13	2.75	59
52.66	52.66	5.5	75
51.63	51.63	2.5	59
35.99	35.99	4.9	42

46.53	46.53	4.85	59
69.45	69.45	3.05	50
61.29	61.29	1.25	30
39.93	39.93	4.25	45
64.91	64.91	2.35	43
45.94	45.94	3.4	57
46.38	46.38	3.5	55
65.95	65.95	4.15	46
53.08	53.08	3	50
38.16	38.16	4.25	47
61.89	61.89	3.6	50
48.75	48.75	2.25	35
41.95	41.95	0.95	45
32.61	32.61	3.6	37
44.03	44.03	1.85	43
58.51	58.51	2.75	41
49.26	49.26	2.1	49

All-variable experiment - Workload Profile

Preferred Semantics	Grounded Semantics	WORKLOADPROFILE	NASAINDEX
46.12	46.12	2.7	58
52.68	52.68	3	54
54.7	54.7	4.6	68
51.87	51.87	3.25	55
69.53	69.53	3.25	50
42.2	42.2	2.25	40
60.48	60.48	3.15	43
55.4	55.4	3.05	38
43.03	43.03	2.55	46
39.83	39.83	1.75	17
48.26	48.26	3.8	53
41.84	41.84	2.55	58
37.34	37.34	2.8	71
63.09	63.09	3	50
54.54	54.54	2.95	48
41.97	41.97	3.5	50
45.85	45.85	2.6	49
40.94	40.94	2.25	40
62.77	62.77	4.3	61
50.66	50.66	3.4	33
41.01	41.01	3	53
51.07	51.07	3.65	48
55.64	55.64	1.75	47
58.72	58.72	1.35	42
42.94	42.94	2.4	41
52.33	52.33	3.9	37
45.36	45.36	1.2	36
40.98	40.98	2.75	59

51.32	51.32	5.5	75
44.39	44.39	2.5	59
56.52	56.52	4.9	42
52.96	52.96	4.85	59
69.72	69.72	3.05	50
48.91	48.91	1.25	30
46.26	46.26	4.25	45
44.09	44.09	2.35	43
50.95	50.95	3.4	57
60.18	60.18	3.5	55
69.47	69.47	4.15	46
56.49	56.49	3	50
51.06	51.06	4.25	47
40.4	40.4	3.6	50
48.21	48.21	2.25	35
45.83	45.83	0.95	45
35.33	35.33	3.6	37
28.44	28.44	1.85	43
52.12	52.12	2.75	41
39.78	39.78	2.1	49