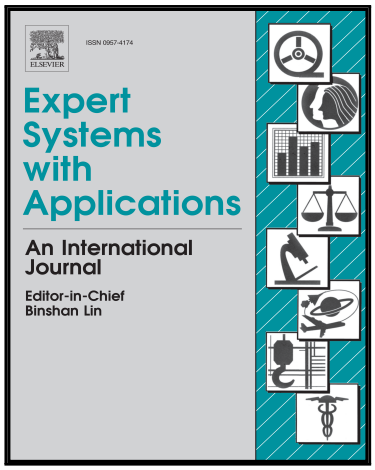



# An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems

Lucas Rizzo, Luca Longo

PII: S0957-4174(20)30046-4
DOI: https://doi.org/10.1016/j.eswa.2020.113220
Reference: ESWA 113220

To appear in: *Expert Systems With Applications*

Received date: 1 August 2019
Revised date: 10 December 2019
Accepted date: 17 January 2020

Please cite this article as: Lucas Rizzo, Luca Longo, An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems, *Expert Systems With Applications* (2020), doi: https://doi.org/10.1016/j.eswa.2020.113220

**Highlights**

- Replicable comparison of inferences produced by non-monotonic reasoning approaches.
- Assessment of the ill-defined construct of mental workload using real-world data.
- Defeasible argumentation presented a superior inferential capacity of mental workload.
- Use of defeasible argumentation in practical fields seldom reported in the literature.
- Robust results analysed in two real-world contexts with three knowledge-bases.

# An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems

Lucas Rizzo*, Luca Longo

*Technological University Dublin, School of Computer Science, Kevin Street, Dublin, Ireland*

**Abstract**

Several non-monotonic formalisms exist in the field of Artificial Intelligence for reasoning under uncertainty. Many of these are deductive and knowledge-driven, and also employ procedural and semi-declarative techniques for inferential purposes. Nonetheless, limited work exist for the comparison across distinct techniques and in particular the examination of their inferential capacity. Thus, this paper focuses on a comparison of three knowledge-driven approaches employed for non-monotonic reasoning, namely expert systems, fuzzy reasoning and defeasible argumentation. A knowledge-representation and reasoning problem has been selected: modelling and assessing mental workload. This is an ill-defined construct, and its formalisation can be seen as a reasoning activity under uncertainty. An experimental work was performed by exploiting three deductive knowledge bases produced with the aid of experts in the field. These were coded into models by employing the selected techniques and were subsequently elicited with data gathered from humans. The inferences produced by these models were in turn analysed according to common metrics of evaluation in the field of mental workload, in specific validity and sensitivity. Findings suggest that the variance of the inferences of expert systems and fuzzy reasoning models was higher, highlighting poor stability. Contrarily, that of argument-based

*Corresponding author

*Email addresses:* lucas.rizzo@tudublin.ie (Lucas Rizzo), luca.longo@tudublin.ie (Luca Longo)

models was lower, showing a superior stability of its inferences across knowledge bases and under different system configurations. The originality of this research lies in the quantification of the impact of defeasible argumentation. It contributes to the field of logic and non-monotonic reasoning by situating defeasible argumentation among similar approaches of non-monotonic reasoning under uncertainty through a novel empirical comparison.

*Keywords:* Defeasible Argumentation, Argumentation Theory, Explainable Artificial Intelligence, Non-monotonic Reasoning, Fuzzy Logic, Expert Systems, Mental Workload

## 1. Introduction

Uncertainty associated to incomplete, imprecise or unreliable knowledge is inevitable in daily reasoning and in many real-world contexts. Within Artificial Intelligence (AI), many approaches have been proposed for the development of inferential models capable of addressing such uncertainty. Among them, non-monotonic reasoning emerged from the area of logical AI as an alternative to deductive inferences in logical systems. These were perceived as inadequate for decision making in realistic situations (Bochman, 2007). Hence, reasoning is non-monotonic, or defeasible, when a conclusion can be withdrawn in the light of new information (Reiter, 1988; McCarthy, 1980; Kowalski & Sadri, 1991; Longo, 2015; Brewka, 1991). A number of approaches for dealing with quantitative reasoning under uncertainty exist (Parsons & Hunter, 1998), including computational argumentation (also referred to as defeasible argumentation) (Prakken & Vreeswijk, 2001), fuzzy reasoning (Zadeh et al., 1965) and expert systems (Durkin & Durkin, 1998). These approaches have led to the development of non-monotonic reasoning models based upon knowledge bases often provided by human experts. Intuitively, since these models have been developed with a human-in-the loop intervention, their reasoning processes and their inferences have an intrinsic higher degree of interpretability and transparency when compared to data-driven approaches for inference. Moreover, they assist on the

creation of models that can be verified, replicated and expanded, thus enhancing the trustworthiness of domain experts towards automated inferences and the understanding of the application under investigation. Nonetheless, these approaches have unique features that differentiate them. For instance, previous studies (Rizzo et al., 2018b,a) suggest that defeasible argumentation offers more powerful conflict resolution strategies; fuzzy reasoning is suitable for robust representation of linguistic information through the application of fuzzy membership functions; and expert systems focus on imitating the problem-solving ability of an expert. These approaches have all been extensively used in practical domains such as medicine, pharmaceutical industry and engineering (Longo, 2016; Glasspool et al., 2006; Mardani et al., 2015; Liao, 2005). However, scholars have predominantly focused on their individual application for non-monotonic reasoning, but barely attempted to empirically investigate their differences in terms of inferential capacity.

The aim of this study is to empirically evaluate the inferential capacity of defeasible argumentation models when compared to other models produced by other well established reasoning approaches, in this case non-monotonic fuzzy reasoning and expert systems. This evaluation can clarify the predictive accuracy of the investigated reasoning models, allowing defeasible argumentation to be better situated among similar reasoning approaches and enabling different applications and experiments to be carried out. To achieve this goal, the problem of representing the construct of Mental Workload (MWL) has been chosen. MWL is an ill-defined construct with no clear and widely accepted definition. In a nutshell, it can be seen as the amount of mental activity devoted to a certain task over time (Cain, 2007). A number of knowledge bases – developed by experts in MWL – were employed as the basis of the modelling and assessment done by the selected approaches. Resulted models are used to infer mental workload scalars employed for achieving the envisioned comparison. In particular, the inferential capacity is compared and quantified in terms of the validity and sensitivity (O'Donnell & Eggemeier, 1986) of the produced inferences. Fig. 1 depicts a streamlined design of the study. With the above elements, a pre-

cise research question can be set: *"To what extent does the inferential capacity of defeasible argumentation differ from non-monotonic fuzzy reasoning and expert systems in terms of validity and sensitivity when applied to the problem of mental workload modelling?"*

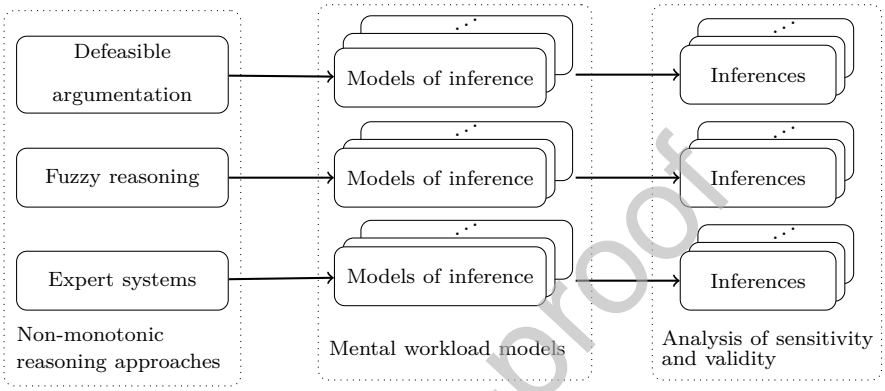



Figure 1: Streamlined design of the study using three non-monotonic reasoning approaches for mental workload modelling, compared according to their inferential capacity.

The remainder of this paper continues with Section 2 providing the related work on non-monotonic reasoning, knowledge-based techniques for dealing with non-monotonic problems and a precise description of the construct of MWL. Section 3 presents the design of the empirical experiment aimed at answering the above research question and the tasks performed by participants of the study in order to collect information for inference of MWL. The results, the analysis and the discussion of this experiment are provided in Section 4. Eventually, Section 5 concludes the study and provides recommendations for future research.

## 2. Literature and related work

Inconsistent and conflicting pieces of information are often involved in real-world argumentative activities. To solve these, classical propositional logic has demonstrated to be inadequate due to its monotonicity property (Reiter, 1980). In monotonic reasoning, a knowledge base of reasons supporting certain conclu-

sions, usually provided by domain experts, may only grow monotonically with new reasons, not allowing the retraction of the previous conclusions. Therefore, defeasible reasoning has emerged as a potential solution to this problem, since it is aimed at formalising non-monotonic reasoning activities (Dung, 1995; Rahwan & Simari, 2009; Chesñevar et al., 2000). This section introduces some of the main non-monotonic formalisms and a few works that have attempted to make a comparison among them. Subsequently, knowledge-base approaches, in particular expert systems, non-monotonic fuzzy reasoning and defeasible argumentation, are explained in depth. The theories in which these approaches are grounded are used as the building blocks for development of non-monotonic reasoning models of inference employed in the context of human mental workload. To the best of the authors' knowledge, there is a lack of comparisons among knowledge-based systems adopted for quantitative reasoning under uncertainty. Hence, the main goal is to provide the reader with the intuitions and the required knowledge for comparing defeasible argumentation with similar reasoning approaches.

### *2.1. Non-monotonic reasoning*

In non-monotonic reasoning, conclusions can be retracted in the light of new reasons. In other words, non-monotonic reasoning relies on the idea that a claim can be defeasibly derived from premises partially specified, but in the case of an exception arising the claim can be withdrawn (Kowalski & Sadri, 1991). Many non-monotonic reasoning formalisms exist in Artificial Intelligence (Brewka, 1991). For instance inheritance networks with exception (Horty et al., 1990) or semantic networks using Demptster's rule (Ginsberg, 1984). Other examples include non-monotonic logics like circumscription (McCarthy, 1980), autoepistemic (Moore, 1985) and default logic (Reiter, 1980). Brewka et al. (1997) provide a nice overview of non-monotonic logics categorized by modal-preference logics, fixed point logics and abductive methods. The recent work of Hlobil (2018) presents a guideline for selection of non-monotonic logics based on principles they reject, such as the Deduction-Detachment Theorem and Cumu-

lative Transitivity (Czelakowski, 1985; Gabbay & Guenthner, 1984), resulting in 17 different types of logics. A few works have proposed the extension of rule-based approaches, such as expert systems and fuzzy reasoning systems, to incorporate a non-monotonic layer (El-Azhary et al., 2002; Nute et al., 1990; Siler & Buckley, 2005; Castro et al., 1998; Morgenstern & Singh, 1997). An alternative approach for performing non-monotonic reasoning is given by argumentation systems as proposed in early studies (Birnbaum et al., 1980; Lin & Shoham, 1989) and other thorough surveys (Atkinson et al., 2017; Chesñevar et al., 2000). This type of systems formalize non-monotonic reasoning by the construction of arguments that can support or be against certain conclusions. Nonetheless, only a few works have proposed a comparison among these formalisms. For instance, Delladio et al. (2006) investigate the relations between a normal default logic and a variant of a defeasible logic programming. Dutilh Novaes & Veluwenkamp (2017) make an empirical test of the accuracy of two formal non-monotonic reasoning models: preferential logic and screened belief revision. Yang et al. (2004) compare first order predicate logic, fuzzy logic and non-monotonic logic implemented through negation as failure. Despite highlighting interesting connections among these formalisms, the focus of the studies is usually theoretical or limited by a narrow scope. In this study, three knowledge-based systems are investigated: expert systems, non-monotonic fuzzy reasoning and defeasible argumentation. Knowledge-based systems are better suited for capturing the intuitions of a specific problem when compared to non-monotonic logics or other proof-theoretic formalisms. Since rules or arguments have to be predefined, only relevant non-monotonic contexts are modelled, living little, if any, place for confusion. The next subsections provide readers with further specific information on these.

### *2.2. Expert systems*

First developed by the AI community in the 1960s, expert systems are computer programs created to emulate a human in a given field (Durkin & Durkin, 1998). In a nutshell, they try to transfer a vast body of specific knowledge

from a human to a computer. In turn, the computer can make inferences and reach a justifiable conclusion. In respect to expert system methodologies, some examples include rule-based systems, knowledge-based systems and fuzzy expert systems (Liao, 2005). Respectively, rule-based systems are based on rules typically of the form "*IF (antecedent) THEN (consequent)*"; knowledge-based systems are human-centred, focusing on the users, their needs and requirements; and fuzzy expert systems employ fuzzy logic for dealing with uncertainty and linguistic terms. Nonetheless, regardless of the methodology, expert systems are usually built upon two internal components: a *knowledge base* and an *inference engine* (Durkin & Durkin, 1998). The former is provided by a human expert and generally translated into a set of logical rules. The latter is aimed at eliciting, firing and aggregating such rules towards a conclusive inference. Moreover, engines might employ common strategies for producing inferences, such as backward-chaining inferencing and forward-chaining inferencing. In both cases, reasoning is exploited in a multi-step process in order to prove some goal or hypothesis. For instance, in a backward-chaining inference process, rules that contain a goal in their consequent part are collected and fired if their premises (same as antecedent) evaluate true. In turn, such premises might be supported by other rules, causing the system to define sub-goals and to work in a recursive fashion. Reflecting that behaviour, a forward-chaining inference process starts by firing rules whose premises match the information initially available. In turn, fired rules might trigger the firing of new rules, leading to a continuation of the process until the goal is reached or no other rule is fired. If multiple rules are fired, both forward-chaining and backward-chaining engines might employ some conflict resolution strategy. Common methods include choosing the first rule located, deciding a priority for each rule or firing all possible lines of reasoning. Other types of expert systems can also be found in the literature, such as frame-based expert systems or probabilistic expert systems (Durkin & Durkin, 1998; Spiegelhalter et al., 1993).

Concerning areas of application, expert systems have been prominently used in fields like medicine and robotics (Nohria, 2015; Singholi & Agarwal, 2018).

For instance, medicine presents strong motivators for the development of medical expert systems, like the lack of specialists and lack of health facilities. Most often they also require interpretable systems. Medical professionals need to have the possibility to understand the reasoning behind a machine and the causes that led it to make a decision. Therefore, in medical area, diagnosis and treatment of diseases are the main goal, with expert systems built for the treatment of influenza, risk of hypertension, memory loss, liver disorders and others (Nohria, 2015). In turn, robotics presents systems developed for fault detection and fault tolerance, path and trajectory planning, vision control, mobile robot control, obstacle detection in industrial robot and so on. The integration of expert systems and robotics is a step forward factory automation still active and researched by the AI community (Singholi & Agarwal, 2018). A wide range of other applications can be found in the expert system literature. Liao (Liao, 2005) provides a decade review, with a considerable amount of specific applications by system methodologies, such as: teaching, agriculture, financial analysis, knowledge management, climate forecasting, decision making, urban design, psychiatric treatment, sensor control, waste water treatment and others. In addition, due to its precondition of encoding human knowledge bases, expert systems have naturally made use of different approaches for knowledge representation, as presented in Hvam et al. (Hvam et al., 2008). These might include graphical notations, logic, scientific formulas and rules. On more specific cases: Mitra and Basu (Mitra & Basu, 1997) implement an expert system which contains distinct knowledge representation schemes for designing microprocessor based systems, while Hatzilygeroudis and Prentzas (Hatzilygeroudis & Prentzas, 2004) propose the integration of symbolic rules, neural networks and cases for the enhancement of knowledge representation and reasoning in expert systems.

Ultimately, non-monotonic techniques have been employed in expert systems in different ways (Gabbay, 1985) and used in industry with certain difficulty (Morgenstern, 1998). A few examples include non-monotonic techniques modelled through inheritance methods (Morgenstern & Singh, 1997), defeasible

logic (Nute et al., 1990) and default reasoning (El-Azhary et al., 2002). Here, the notions of "contradictions" or "exceptions" are employed. These are defined by domain experts, and describe special cases in which a rule is no longer valid and has to be retracted from the reasoning process.

### *2.3. Non-monotonic fuzzy reasoning*

Fuzzy set theory, as proposed by Zadeh (Zadeh et al., 1965), uses the notion of membership function, a special function that assigns to each object or linguistic term a grade of membership in the range [0,1] $\in \mathbb{R}$. Fuzzy sets are formed by fuzzy objects and include similar notions to classical set theory such as inclusion, union and intersection. A fuzzy control system or fuzzy expert system is a control system based on fuzzy reasoning. It is usually formed by a set of inputs defined as a fuzzy set, a rule set and a defuzzification module (Passino et al., 1998). In this case, this process is characterised as a *Mamdani* fuzzy inference (Mamdani, 1974) (Fig. 2) and is the approach employed in this study. Moreover, two other types of fuzzy inference methods are commonly found in the literature. The first, the *Takagi-Sugeno* fuzzy inference (Takagi & Sugeno, 1993), presents the same fuzzification process, however, the output membership functions are always linear or constant, producing in either case a single number. On the one hand, there is no defuzzification process and on the other hand, it is necessary to define weighting mechanisms or parameters for the linear output functions to compute a final crisp value. The second, the *Tsukamoto* fuzzy inference (Tsukamoto, 1979), also differs from the other types only by its output membership functions. In this case, consequents of each rule are crisp values defined by a monotonical membership function and the real input of the associated rule. Intuitively, it is a combination of the Mamdani and the Takagi-Sugeno fuzzy inference methods.

Since the original development of fuzzy set theory by Zadeh (Zadeh et al., 1965), the range of its applications has been vast. Examples of application domains include pattern recognition, decision making, signal processing, control engineering, medicine, finance and many others. Precup and Hallendoorn

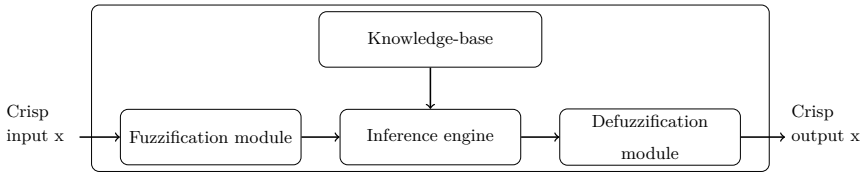



Figure 2: General structure of a Mamdani fuzzy inference process (Cordón, 2011).

(Precup & Hellendoorn, 2011) present an extensive survey paper on industrial applications of fuzzy control. Particularly, numerous applications of Mamdani fuzzy control systems have been reported in the fields of robotics, automotive industry and process industry. Due to the concern on the accuracy of such applications, learning techniques have also been incorporated into fuzzy control systems in order to deal with the interpretability-accuracy trade-off (Cordón, 2011), leading to the fields of neuro-fuzzy systems (Nauck et al., 1997) and genetic fuzzy systems (Cordón et al., 2004). Learning techniques might cover structural changes ranging from the parameters optimization to the learning of the rule set. Other works have also suggested additional extensions of fuzzy inference systems in order to support non-monotonicity of rules. Unfortunately, these extensions are not well established. For example, in (Castro et al., 1998) conflicting rules have their conclusions aggregated by an averaging function, while in (Gegov et al., 2014) a rule-base compression method is proposed for the reduction of non-monotonic rules. A third approach can be seen in (Siler & Buckley, 2005), whereby Possibility Theory (Dubois & Prade, 1998) is included into the fuzzy reasoning system to tackle conflicting instructions. In Possibility Theory, contrarily to traditional fuzzy systems, propositions have two truth values: *possibility* and *necessity*. The first indicates the extent to which data fails to refute its truth while the second indicates the extent to which data supports its truth. This theory is adopted in this study for the development of a non-monotonic fuzzy reasoning system (detailed in Section 3.2).

### 2.4. Defeasible argumentation

Argumentation, with origins grounded in philosophy, deals with the study of assertion and definition of arguments usually emerged from divergent opinions. In the field of Artificial Intelligence, argumentation, also referred to as defeasible argumentation (Bryant & Krause, 2008), is aimed at developing computational models of arguments. Such models have become increasingly significant within AI (Bench-Capon & Dunne, 2007), making defeasible argumentation widely employed for modelling non-monotonic reasoning (Chesñevar et al., 2000). Many studies also described its potential for practical applications, such as dialogue and negotiation (Bench-Capon & Dunne, 2007; Black & Hunter, 2009; Kraus et al., 1998; Amgoud et al., 2000), knowledge representation (Longo, 2015; Dondio & Longo, 2014) and decision making in health-care (Glasspool et al., 2006; Longo & Dondio, 2014; Patkar et al., 2006). Some of the appealing properties of argument-based models include the lack of statistics or probability for inference and capability to deal with partial and inconsistent pieces of evidence. Thus, being closer to the way humans reason under uncertainty and leading to a higher explanatory capacity (Longo, 2016). This can be exemplified by its attempted use for the development of argumentation-based approaches to explainable AI (Zeng et al., 2018). Moreover, their conflict resolution strategy is strengthened by the large body of literature on acceptability semantics (Dung, 1995; Amgoud et al., 2017; Baroni et al., 2011; Baroni & Giacomin, 2009; Dondio, 2018). Acceptability semantics provide solid mechanisms for the selection of acceptable arguments within a set of conflicting arguments. This set is usually represented by a graph in which arguments are depicted as nodes and attacks (conflicts) between arguments are depicted as arrows. The set of acceptable arguments is usually referred to as an *extension*. Acceptability semantics can provide a unique extension or multiple extensions for the same set of conflicting arguments. For instance, the common Dung's *grounded* semantics (Dung, 1995) always returns a single extension while the Dung's *preferred* semantics might return a single or multiple ones (detailed in Section 3.3.4).

Several approaches also exist for quantitative argumentation, or argumen-

tation that deals with numerical measurable arguments, such as Bipolar Argumentation, Probabilistic Argumentation, Multi-valued Argumentation and Weighted Argumentation (Rahwan & Simari, 2009). Despite this number of approaches, computational argumentation systems are usually structured around layers specialised on the the definition of internal structure of arguments, the definition of arguments interactions, the resolution of conflicts between arguments and the possible resolution strategies for reaching a justifiable conclusion (Prakken & Vreeswijk, 2001). Still, the boundaries of such layers might not be accurately defined. For that reason a few layered structures have been proposed for the development of computational models of argument. Prakken & Sartor (2002) suggest a four-layered view applied to legal argumentation that contains: a logical layer, which defines the arguments themselves; a dialectical layer, focused on the definition of notions such as attack and defeat; a procedural layer, which regulates how parties can challenge and introduce new arguments; and a strategic or heuristic layer, which defines how a dispute should be conducted within the bounds of the procedural layers. Differently, Atkinson et al. (2017) consider five main layers as the basic building blocks of an argumentation model: structural layer, relational layer, dialogical layer, assessment layer and rhetorical layer. Another example of multi-layered structure can be found in (Longo, 2016) and is depicted in Fig. 3.

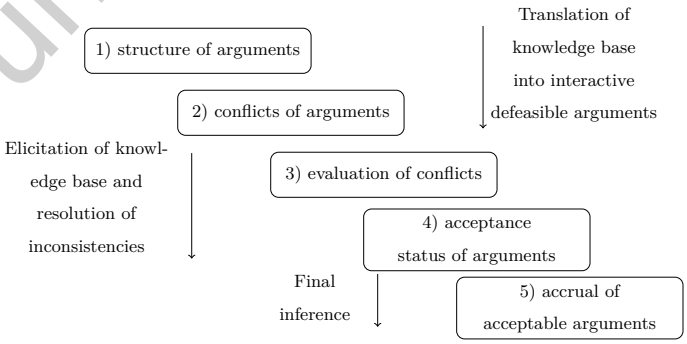



Figure 3: Five layers structure (Longo, 2016) adopted for the development of argument-based models.

This research study adopts this structured due to the nature of the application selected for evaluation – modelling and assessment of human mental workload. In this case, each knowledge base employed is the result of the reasoning of a single agent and do not require a rhetorial layer. The objective is to reason with arguments neutrally built from domain experts so as to achieve a numerical inference representing the imposed mental workload by a specific task. Each layer in this structure is supported by theoretical works in the field of defeasible argumentation. For example, in Layer 1, Toulmin (Toulmin, 1958) provides one the first conceptual models of arguments aimed at contributing with a more articulated structure for arguments. Another example is given by Walton (Walton, 2013), who identifies and evaluates a variety of argumentation structures in everyday discourse, such as argument from consequence, appeal to expert opinion, argument from analogy and argument by example. Other models of argument are also described in (Bentahar et al., 2010). In Layer 2 the focus is on the relationship between arguments and management of their conflicts. Prakken (Prakken, 2010) proposes a conflict classification with three different classes: *undermining attack* when an argument is attacked on one of its premises, *rebutting attack* when an argument negates the conclusion of another argument and *undercutting attack*, when an argument is attacked at one of its defeasible inference rules. Following to Layer 3, the focus is now on the ability to characterize the success of an attack. Commonly, attacks have a form of a binary relation. In a binary attack relation all attacks are successful if they have a target (argument being attacked) and source (argument attacking) defined. However, other approaches are presented in the literature, such as: strength of arguments, preferentiality and strength of attack relations (Dunne et al., 2011; Modgil, 2009; Martınez et al., 2008). The first one presents the inequality of the strength of arguments that has to be accounted for in a decision-making process. Preferentiality assumes the information necessary to decide whether an attack between two arguments is successful is pre-specified. The last approach, strength of attack relations, tries to associate weights to attack relations instead of arguments. Given an evaluation of attacks, acceptability semantics,

placed in Layer 4, can be employed for the definition of the acceptability status of arguments. Dung semantics (Dung, 1995) and its variations (Caminada, 2007; Caminada et al., 2012) are the most well known. Other types include SCC-recursive semantics (Baroni et al., 2005) focused on solving cyclic attack relations of odd-length and ranking-based semantics (Bonzon et al., 2016) which rank arguments from most acceptable to weakest one(s). Finally, the selection of extensions and the accrual of acceptable arguments is done in Layer 5. A few strategies (Coste-Marquis et al., 2012; Konieczny et al., 2015) can be found in the literature for selection of extensions, such as the employment of the strength of arguments from Layer 3 or the selection of the extension(s) with higher cardinality. Nonetheless, this layer is not always required and is seemingly the less developed in the literature, requiring further investigation.

Some works tackle all these 5 layers (Chang et al., 2009; Hunter & Williams, 2010; Craven et al., 2012) while others do not (Patkar et al., 2006; Glasspool et al., 2006; Grando et al., 2013). This structure has also been reproduced in past studies (Rizzo & Longo, 2017; Rizzo et al., 2018a; Longo, 2015; Rizzo & Longo, 2018) demonstrating structural effectiveness in different domains of application. Unfortunately, despite the increasing application of argumentation in various theoretical fields, the use of defeasible argumentation in practical fields is one of the challenges in respect to the general deployment of argumentation technology as suggested by Bench-Capon et al. (Bench-Capon & Dunne, 2007). This challenge represents the main motivation behind the research question outlined in the introductory section.

### *2.5. Mental workload*

To tackle the research question, a precise knowledge representation and reasoning problem has been selected: mental workload (MWL) modelling. Note that this problem is not the focus of this research study, but only an application that allows the proposed comparison among the non-monotonic reasoning approaches to be performed. Thus, only a brief introduction of its concept, methods of measurement and evaluation metrics are provided here. The inter-

ested reader can refer to the citations along this section for further information.

Although no single definition has been developed so far (Young et al., 2015; Hart, 2006), MWL can be intuitively described as the total cognitive cost needed to accomplish a specific task over time (Cain, 2007). According to Cain (2007), the main reason for measuring MWL is to quantify the mental cost of performing a certain task in order to predict operator and system performance. It is mainly used in the areas of psychology and ergonomics, with applications in aviation and auto-mobile industries (Paxion et al., 2014) and in interface and web design (Tracy & Albers, 2006).

Since no correct measure of MWL exists, there are different methods that have been proposed for measuring it (Eggemeier, 1988). These can be categorised into subjective measures, task performance measures and physiological measures. Task performance measures try to infer MWL from objective notions of performance, like number of errors, completion time and time to respond to a secondary task. Physiological measures try to infer a MWL scalar from physiological responses, like pupillary reflex or muscle activity. In this work we adopt the class of subjective measures. This class leans on the analysis of the subjective feedback (such as questionnaires) provided by humans engaging with an underlying task. Among well known methods, the NASA-Task Load Index (NASA-TLX) (Hart & Staveland, 1988) has been largely employed in the last decades (Rizzo et al., 2016; Longo, 2014, 2015) and it is adopted in this research study for comparison purposes. It is a combination of six factors believed to influence mental workload: temporal demand, physical demand, mental demand, frustration, effort and performance (Hart & Staveland, 1988). Each factor $d$ is quantified with a subjective judgement coupled with a weight $w$ computed via a pairwise comparison procedure. The set of questionnaires employed for measurement of each factor can be seen in Table A.11 (page 74). The final MWL scalar is the weighted average of these six factors $d_i$ and weights $w_i$ provided by the operator (equation 1). The pairwise comparison procedure is made through a set of questions, for example "which contributed more for the MWL: mental demand or effort?", "performance or frustration?", giving a total of 15 prefer-

ences. The number of times each feature is chosen defines its weight. A few modified versions of the NASA-TLX have also been proposed. Among them, the most common is referred to as Raw TLX (RTLX) (Hart, 2006). It removes the pairwise comparison procedure of NASA-TLX and instead averages the features (equation 2). According to (Hart, 2006), comparisons between the NASA-TLX and the RTLX seem inconclusive, being both more or less sensitive than the other to changes in task difficulty.

$$\mathrm{TLX}_{\mathrm{MWL}} = \Big(\sum_{i=1}^{6} d_i \times w_i\Big)\frac{1}{15} \qquad (1) \qquad\qquad \mathrm{RTLX}_{\mathrm{MWL}} = \Big(\sum_{i=1}^{6} d_i\Big)\frac{1}{6} \qquad (2)$$

Another MWL assessment technique is the Workload Profile (WP) which is based on the Multiple Resource Theory (MRT) (Wickens, 1991). Contrarily to the NASA-TLX, it is built upon 8 dimensions: solving and deciding, selection of response, task and space, verbal material, visual resources, auditory resources, manual response and speech response (Table A.17, questions 6-13). The user is required to rate each feature in the range 0 to 1. The final scalar is given then by their sum (eq. 3).

$$\mathrm{WP}_{\mathrm{MWL}} = \sum_{i=1}^{8} d_i \qquad (3)$$

Several criteria have been proposed for the selection and development of inferential models of MWL (O'Donnell & Eggemeier, 1986), such as: diagnosticity, reliability, sensitivity and validity among others. Since the goal of this research study is to evaluate the ability of non-monotonic reasoning techniques to represent and assess MWL, the focus is on three different forms of validity and sensitivity:

- *convergent validity*: it demonstrates the extent to which different MWL techniques correlate to each other(Tsang & Velazquez, 1996).
- *concurrent validity*: it determines to what extent a technique can explain measures of objective performance, such as task execution time (Rubio et al., 2004).
- *face validity*: it determines the extent to which a technique is relevant to

the persons answering the questions. Or if the workload reported seems to be valid to participants of the experiment (Spielberger et al., 2010).

- *sensitivity*: it determines the capability of a technique to discriminate significant variations in MWL and changes in resource demand or task difficulty (O'Donnell & Eggemeier, 1986).

Validity and its particular sub-forms have normally been assessed through the analysis of correlation coefficients (Rubio et al., 2004) between produced MWL scalars, while sensitivity has been formally evaluated by analysis of variance coupled with post hoc analysis (Rubio et al., 2004; Longo, 2015).

In summary, MWL is a complex construct built over a network of pieces of evidence; accounting and understanding the relationships of these pieces of evidence as well as resolving the inconsistencies arising from their interaction is essential in modelling MWL (Longo, 2014). In formal logics, these activities are the key components of a defeasible argumentative process, where a set of interactive pieces of evidence, called arguments, can be defeated by additional arguments (Longo, 2014). To the best of our knowledge, Longo (2012) was the first to attempt to model MWL as a non-monotonic concept. Thus, in spite of MWL not being the focus of this research, it is important to highlight that no other authors have followed this modelling approach. Previous works have investigated the use of expert systems for MWL modelling (Rizzo et al., 2016) and the comparison of defeasible argumentation and non-monotonic fuzzy reasoning (Rizzo & Longo, 2019, 2017). Nonetheless, these are not comprehensive studies, employing small sets of data and limited sets of inference models. Here, a thorough investigation has been proposed, extending preceding studies and fine tuning designed inference models. In particular, this research is secondary in terms of data employed. It employs information of studies proposed in (Longo, 2018b; Longo & Orru, 2019; Longo, 2018a, 2017; Longo & Dondio, 2015) for the evaluation of MWL imposed on participants who performed two types of tasks: information seeking web-based tasks and attendance to third-level classes delivered at the Technological University Dublin (a detailed description of these

tasks if given in Section 3.4). The answers provided by these participants led to the creation of three different datasets evaluated simultaneously in this study. In specific, they were used to elicit the non-monotonic reasoning models introduced in the next section.

## 3. Design and methodology

In order to answer the research question a primary quantitative research was designed as depicted in Fig. 4. Empirical evidence was employed with two objectives in mind:

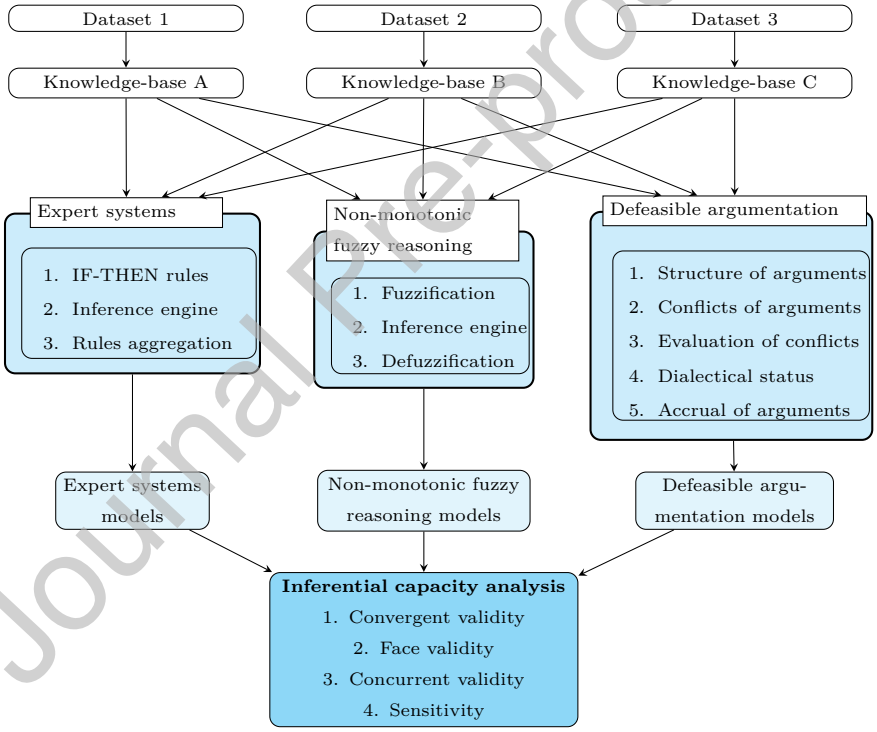



Figure 4: Evaluation strategy schema and full inferential process applied to three distinct knowledge-bases instantiated by three distinct datasets.

1. To investigate the capacity of non-monotonic reasoning models to assess the construct of MWL according to state-of-the-art MWL measurement

techniques (NASA-TLX, Raw TLX and WP).

2. To investigate the quality of inferences produced by non-monotonic reasoning models.

The hypothesis for objective 1 is that non-monotonic reasoning models will demonstrate high convergent validity with baseline instruments, thus being able to assess MWL. The hypothesis for objective 2 is that defeasible argumentation models will demonstrate higher sensitivity, higher concurrent validity and higher face validity than fuzzy reasoning and expert system models, thus showing that defeasible argumentation has a better inferential capacity than the other non-monotonic reasoning approaches. Table 1 lists the hypotheses and methods associated to each objective of this research study.

Table 1: Objectives and hypotheses of the research study.

| | |
|---|---|
| **Objective 1** | Evaluation of the capacity to assess the construct of MWL. |
| **Method** | Evaluation of convergent validity. |
| **Hypothesis 1** | Non-monotonic reasoning models will demonstrate moderate to high convergent validity with baseline instruments. |
| **Objective 2** | Investigate the quality of produced inferences. |
| **Method** | Evaluation of face validity, concurrent validity and sensitivity. |
| **Hypothesis 2** | Defeasible argumentation models will demonstrate higher sensitivity, higher concurrent validity and higher face validity than fuzzy reasoning and expert system models. |

Three knowledge bases (Appendix A), designed by two interviewed experts, were employed for the construction of models capable of inferring a mental workload scalar (value in the range $[0, 100] \in \mathbb{R}$). Each knowledge base was built with rules constructed by only considering the information gathered with well known self-reporting mental workload instruments. Each rule was subsequently elicited with the data associated to its premises. The construction of datasets, knowledge bases and description of performed tasks designed to as-

sess MWL are detailed in the following subsections. As summarised in Fig. 4, non-monotonic reasoning models are firstly built upon an expert knowledge base and a reasoning approach. Secondly, these models are instantiated with the data associated to the selected knowledge base and the respective inferences are produced (MWL scalars). This process is repeated for each knowledge base. Finally, the inferences produced using all knowledge bases are compared against each other to test the research hypotheses.

### *3.1. Expert systems*

Focused on imitating the problem-solving ability of a human expert, expert systems are one of the most well known reasoning approaches in the literature. A step-by-step description of their inferential process is provided along with a running example (Fig. 5) for the problem chosen in this paper: mental workload modelling and assessment. This example is referred throughout this section and is aimed at providing a complete overview of the expert system procedure for inferring a MWL scalar with real-world data.

#### *3.1.1. IF-THEN rules and contradictions*

The first step of an expert system is to model a knowledge base usually gathered from an expert with rules of the form *"IF (antecedent) THEN (consequent)"*. In this research study, the antecedent is one or a set of premises associated to a number of MWL features, believed by the expert, to influence MWL, while the consequent is associated to a possible MWL level that can be deductively derived from the premises. Examples of hypothetical rules are described below:

- Rule 1: **IF** *low* `mental demand` **THEN** *underload* MWL
- Rule 2: **IF** *low* `effort` **THEN** *fitting load* MWL

Each level of a premise in the antecedent, as well as each level of the consequent, are mapped to a numerical range by the domain expert. The input values then determine the activated rules and contradictions. A rule can also

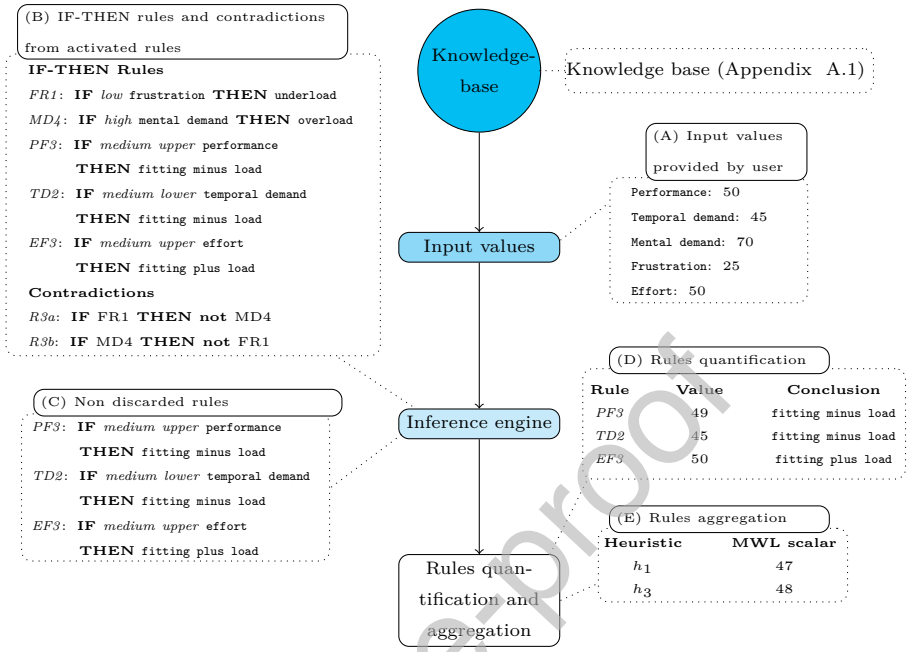



Figure 5: An illustration of a reasoning process of an expert system. The order of operations is from step (A) to step (E).

be contradicted by other rules which intend to bring forward and support contradictory information. An example of a hypothetical contradiction is:

- Contradiction 1: **IF** *high* `effort` **THEN not** Rule 1

The set of IF-THEN rules and the set of contradictions is now ready to be elicited. In detail, the second step of the expert system is to define the inference engine aimed at firing rules and solving contradictions among them.

### *3.1.2. Inference engine*

The inference engine starts with the activation of IF-THEN rules and contradictions with real-world data. This means that input data will be used to evaluate antecedents of rules and contradictions, firing a sub-set whose evaluation returns true. If both a IF-THEN rule and at least one contradiction challenging the rule have been activated, then the inference engine discards the

rule. This mechanism will eventually form a set of surviving rules. Fig. 5.A, 5.B and 5.C respectively depict the input values in the running example, the set of activated rules and the set of surviving rules. Note that these rules and arguments come from a real knowledge base that can be seen in Appendix A. They may not be the same as hypothetical rules and contradictions, such as Rule 2 and Contraction 1. Experts can have different opinions and the fact that a set of premises infers a conclusion in one knowledge base does not mean it has to infer the same conclusion in another knowledge base.

*3.1.3. Rules quantification and aggregation*

The rules in the set of surviving rules might have distinct consequents. For example, in this research study, there might be rules inferring different MWL levels. Since the goal is to aggregate them and extract an unique scalar, most representative of the imposed mental workload, an aggregation strategy is needed. In this situation, a usual expert system would have a typical set of choices for selection of rules, for example: deciding a priority for each rule, returning multiple outcomes or choosing the first rule activated. However, none of these strategies is applicable in this research study. The knowledge bases do not explicit preferences among rules, order of activation or possibility to compute more than one output. Because of that, rules have to be quantified and aggregated[1] to infer a MWL scalar in the range $[0, 100] \in \mathbb{R}$.

In the quantification step, a value has to be attributed for each surviving IF-THEN rule. In this study, this value is defined according to the numerical range of the consequent of the rule, the numerical range of its premises and the input values provided for the rule activation. In the basic scenario of an IF-THEN rule with only one premise, it will be quantified as the minimum (resp. maximum) value of the numerical range of its consequent if its premise is activated with its

[1]A third step, after the definition of rules and inference engine, is provided here for the design of expert system models. Commonly, the final inference of usual expert systems is given by the inference engine. However, in the interest of clarity, quantification and aggregation of rules are defined in a third step, which could theoretically still be part of the inference engine.

minimum (resp. maximum) value. For instance, consider Rule 2 rewritten with hypothetical numerical ranges:

- Rule 2 rewritten: **IF** `effort` $\in$ [0, 33] **THEN** MWL $\in$ [33, 66]

In this case, if the input value for `effort` is 0, then Rule 2 value will be 33. Analogously, if the input value for `effort` is 33, Rule 2 value will be 66. Activation values in between 0 and 33 are evaluated according to a linear relationship. To formalize the generic case, IF-THEN rules are precisely defined, followed by the definition of the function $f$ that returns their value:

**Definition 1 (Generic IF-THEN rule).** *A generic IF-THEN rule is defined, without loss of generalisability, as:*

$$\textit{IF } (i_1 \in [l_1, u_1] \text{ AND } i_2 \in [l_2, u_2]\ ) \text{ OR } (i_3 \in [l_3, u_3] \text{ AND } i_4 \in [l_4, u_4]) \textit{ THEN } MWL \in [l_c, u_c]$$

*Where $i_n \in \mathbb{R}$ is the input value of the feature n with numerical range [$l_n \in \mathbb{R}$, $u_n \in \mathbb{R}$]; [$l_c \in \mathbb{R}$, $u_c \in \mathbb{R}$] is the numerical range for the MWL level being inferred; and* AND *and* OR *are boolean logical operators.*

**Definition 2 (Generic rule value).** *The value of a generic IF-THEN rule r is given by the function:*

$$f(r) = \frac{(u_c - l_c)}{R_{max} - R_{min}} \times (v - R_{max}) + u_c, \textit{ where}$$

$$v = min[max(i_1, i_2), max(i_3, i_4)],$$
$$R_{max} = min[max(u_1, u_2), max(u_3, u_4)],$$
$$R_{min} = min[max(l_1, l_2), max(l_3, l_4)]$$

Note that the value of a rule will always lies between the numerical range [$l_c$, $u_c$] of the MWL level being inferred. In a nutshell, Def. 2 provides a normalization formula for rules that employ logical operators AND/OR, replacing them for *max* and *min* operators[2]. Fig. 5.D provides a numerical example.

[2]Different operators could have been employed if defined by the knowledge base designer.

Finally, four heuristics are defined to accomplish the aggregation of surviving IF-THEN rules inferring some MWL level. The strategies are developed in order to extract different points of view from the remaining rules and accommodate the use of rule weights. No preference or weight among rules is provided in the employed knowledge bases, still the pairwise comparison procedure of the NASA-TLX is adapted here as a form of rule weight. The aim is to investigate the impact of adding this extra information on the inferential capacity of the expert system models. In the pairwise comparison procedure, the number of times a feature has been chosen over another is its respective weight, which in turn will also represent the weight of the IF-THEN rules whose antecedents contain such feature. Observe that instead of general rule weights, rules will have different weights on a case by case basis.

- $\boldsymbol{h_1}$: definition of the sets of surviving rules grouped by their MWL level. Extraction of the largest set. Average of the values of the rules in the largest set. In case two or more largest sets exist, the above process is repeated for each of them and their average is returned. The idea is to give importance to the largest set of surviving rules supporting the same MWL level.

- $\boldsymbol{h_2}$: same as $h_1$ but applying the weighted average instead of the average. The goal here is to add the information from the pairwise comparison procedure provided by the NASA-TLX questionnaire.

- $\boldsymbol{h_3}$: average value of all surviving IF-THEN rules. This is to give equal importance to all surviving IF-THEN rules, regardless of which level of MWL they were supporting.

- $\boldsymbol{h_4}$: same as $h_3$ but applying the weighted average instead of the average. Again, the goal is to employ the information of the pairwise comparison procedure of the NASA-TLX.

Fig. 5.E depicts the output for two heuristics.

### *3.2. Non-monotonic fuzzy reasoning*

For comparison purposes, fuzzy reasoning is the second reasoning approach selected in this research study. It provides a robust representation of linguistic information by using fuzzy membership functions. In addition, it considers Possibility Theory (Dubois & Prade, 1998) in the reasoning process to tackle non-monotonicity. Similarly to expert systems, a running example of a single inference with real-world data is depicted in Fig. 6 and referred throughout this subsection.

#### *3.2.1. Fuzzification module*

The first step, the fuzzification module, starts with the definition of *fuzzy* IF-THEN rules and *fuzzy* contradictions. Hypothetical examples of these are:

- Fuzzy Rule 1: **IF** *low* `mental demand` **THEN** *underload* MWL
- Fuzzy Rule 2: **IF** *low* `effort` **THEN** *fitting load* MWL
- Fuzzy Contradiction 1: **IF** *high* `effort` **THEN not** Fuzzy Rule 1.

Fig. 6.A and Fig. 6.B depict the representation of the knowledge base of an expert with fuzzy IF-THEN rules and fuzzy contradictions.

Afterwards, each linguistic term associated to a feature level or MWL level, such as *low* or *underload*, is described by a fuzzy membership function (FMF) that is also provided by the knowledge base designers. Appendix A.4 depicts the three options provided, using linear, trapezoidal and Gaussian shapes. In the running example, membership functions for MWL levels and feature levels and can be seen in Fig. 6.C and Fig. 6.D respectively.

#### *3.2.2. Inference engine*

Once the fuzzification step has been completed and the knowledge base of the expert translated into fuzzy rules and fuzzy contradictions, the next step is to solve such contractions. Possibility Theory is used here as a possible approach, as implemented in (Siler & Buckley, 2005) for fuzzy reasoning with

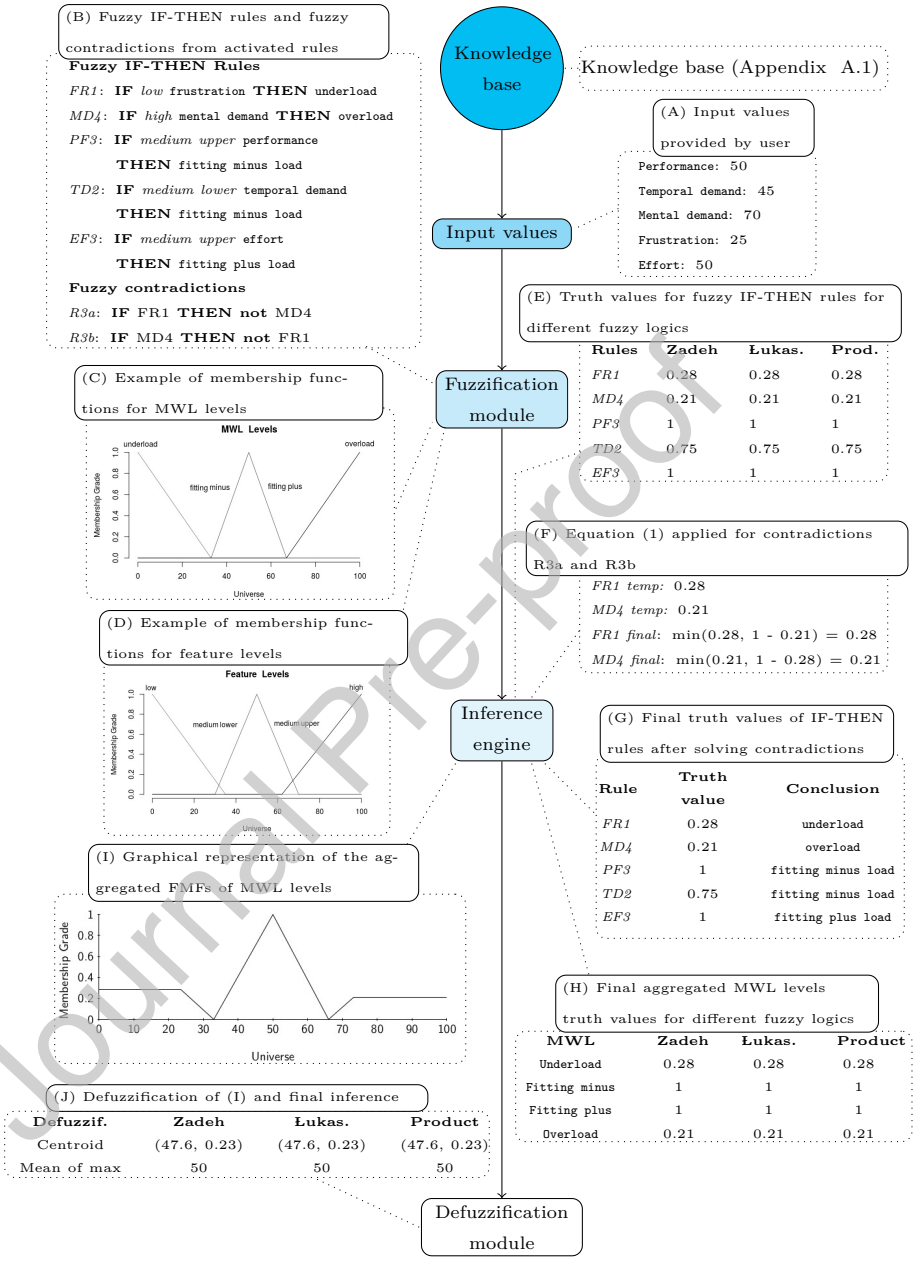



Figure 6: An illustration of a reasoning process of a fuzzy reasoning system with the property of non-monotonicity. The order of operations is from step (A) to step (J).

rule based systems. According to this approach, truth values can be represented by *possibility* (Pos) and *necessity* (Nec) as defined in Section 2.3. Both are values between [0, 1] $\in \mathbb{R}$. Possibility of a proposition can also be seen as the upper bound of its respective necessity (Pos $\geq$ Nec). In this study, necessity represents the membership grade of a proposition and possibility is always 1 for all propositions. Under these circumstances, the effect on the necessity of a proposition $A$ by a set of propositions $Q$ which contradicts $A$ is derivable as:

$$Nec(A) = min(Nec(A), \neg Nec(Q_1), \ldots, \neg Nec(Q_n)) \tag{4}$$

where $\neg Nec(Q) = 1 - Nec(Q)$. Sidnce there is no addition of supporting information but only attempts to contradict or refute information, equation (4) can deal with the contradictions in the knowledge bases of this study. For instance, the truth value of the Fuzzy Rule 1, assuming that it is contradicted only by the Fuzzy Contradiction 1, is given by:

- Truth value of Fuzzy Rule 1 =

  min (Nec(*low* `mental demand`), 1 - Nec(*high* `effort`))

Nec(*low* `mental demand`) is the membership grade of the linguistic variable *low* of the feature `mental demand`. For instance, if `mental demand` = 1, then Nec(*low* `mental demand`) = 1, according to the membership function *low* of Fig. A.26b (p. 82). Also, for instance, if $Nec$(*high* `effort`) = 0 then it must be noted that the Fuzzy Contradiction 1 has no impact on the Fuzzy Rule 1 and if Nec(*high* `effort`) = 1 the new truth value of the Fuzzy Rule 1 is 0. Values between 1 and 0 indicates that the Fuzzy Rule 1 is partially refuted. The truth value of the Fuzzy Rule 1 represents the truth value of `underload` in this particular rule.

It is important to highlight that the approach developed in (Siler & Buckley, 2005) has been inspired by a multi-step forward-chaining reasoning system. In this research study, reasoning is done in a single step, in the sense that data is imported and all rules are fired at once. However, it is possible to define a precedence order of fuzzy contradictions. More exactly, it is possible to define a

tree structure in which the consequent of a fuzzy contradiction is the antecedent of the next fuzzy contradiction. In this way, equation (4) can be applied from the root or roots to the leaves. This approach is sufficient for knowledge bases that do not contain cyclic exceptions, but according to the knowledge bases employed in this study, that is not the case. For instance consider the following hypothetical fuzzy IF-THEN rules and their fuzzy contradictions:

- Fuzzy Rule 3: **IF** *low* `temporal demand` **THEN** *underload*
- Fuzzy Rule 4: **IF** *high* `frustration` **THEN** *overload*
- Fuzzy Contradic. 2: **IF** *low* `temporal demand` *THEN not* Fuzzy Rule 4
- Fuzzy Contradic. 3: **IF** *high* `frustration` *THEN not* Fuzzy Rule 3

In this case it is not clear if Fuzzy Contradiction 2 or 3 should be solved first. Given that there is no information on the knowledge bases (accounted in this study as per Appendix A) to decide whether a fuzzy rule or a fuzzy contradiction is more important than another, here they are solved simultaneously. Firstly, the truth values of all fuzzy rules are stored before solving any cyclic fuzzy contradictions. Secondly, the final truth value of fuzzy rules is calculated according to equation (4) and the temporary values stored before as per example below:

- Temp1 = Nec(Fuzzy Rule 3) = Nec(*low* `temporal demand`)
- Temp2 = Nec(Fuzzy Rule 4) = Nec(*high* `frustration`)
- Truth value Fuzzy Rule 3 = min (Nec(*low* `temporal demand`), 1 - Temp2))
- Truth value Fuzzy Rule 4 = min (Nec(*high* `frustration`), 1 - Temp1))

Having a mechanism to solve fuzzy contradictions, fuzzy operators can be applied to the antecedents of fuzzy IF-THEN rules and for the aggregation of the consequents (MWL levels) across the rules. Three known operators are selected for investigation: the *Zadeh*, the *Product* and the *Łukasiewicz* operators. Table

2 lists the t-norms and t-conorms (fuzzy AND and fuzzy OR) respectively for each operator. Antecedents might employ OR or/and AND, while consequents (MWL levels) are aggregated only by the OR operator. For instance, the truth value of *underload* in a context where only Fuzzy Rule 1 and Fuzzy Rule 3 infer *underload* is "Nec(Fuzzy Rule 1) OR Nec(Fuzzy Rule 3)".

Table 2: T-Norms and t-Conorms employed for two propositions $a$ and $b$

| **Fuzzy operator** | **T-Norm** | **T-Conorm** |
|---|---|---|
| Zadeh | min($a$,$b$) | max($a$,$b$) |
| Łukasiewicz | max($a$ + $b$ - 1, 0) | min($a$ + $b$, 1) |
| Product | $a.b$ | $a$ + $b$ - $a.b$ |

Fig 6.E, 6.F and 6.G respectively depicts the truth values of fuzzy rules, the resolution of the contradictions and the updated truth values of fuzzy rules.

At this stage if rule weights are defined these should be applied to the current truth values of fuzzy IF-THEN rules. In this study, the approach proposed by (Ishibuchi & Nakashima, 2001) is selected. In this case, rule weights are normalized in the range $[0, 1] \in \mathbb{R}$ and multiplied by the current truth value of each rule. Weights are provided by the pairwise comparison procedure of the NASA-TLX questionnaire (Table A.13) and adapted as in the expert systems design (Section 3.1.3).

Eventually, the truth values of the final MWL levels are generated by aggregating the consequents of the fuzzy IF-THEN rules using the OR operator. Fig. 6.H depicts an example with no rule weights.

### *3.2.3. Defuzzification module*

The output of the inference engine is a graphic representation of the aggregation of the consequents (MWL levels) of the updated fuzzy IF-THEN rules (Fig. 6.I). Several methods can be used for calculating a single defuzzified scalar (Hellendoorn & Thomas, 1993). Two are selected here: *mean of max* and *centroid*. The first returns the average of all elements (MWL levels) with maximal

membership grade. The second returns the coordinates ($x$, $y$) of the centre of gravity of the geometric shape formed by the aggregation of the membership functions associated to each consequent (MWL level). The defuzzified scalar is represented then by the $x$ coordinate of the centroid (as per Fig. 6.J).

### *3.3. Defeasible argumentation*

The definition of argument based-models follows the 5 layer modelling approach proposed in (Longo, 2016) and depicted in Fig. 3 (Section 2.4). It starts with the definition of the internal structure of arguments, followed by the definition of conflicts among arguments, the definition of the acceptance status of each argument and the aggregation of the accepted arguments. A running example is depicted in Fig. 7 and referred throughout this subsection.

#### *3.3.1. Layer 1 - Definition of the internal structure of arguments*

Most commonly an argument is composed of one or more premises that provides reason or support a conclusion. Thus, the first step of an argumentation process usually focuses on the construction of *forecast arguments* defined as:

$$Forecast\ argument : premises \rightarrow conclusion$$

This structure includes a set of premises (believed to influence the conclusion being inferred) and a conclusion derivable by applying the inference rule $\rightarrow$. It is an uncertain implication which is used to represent a defeasible argument. In order to solve the application in hand (MWL), similarly to the rules of expert systems, premises and conclusions are strictly bounded in numerical ranges associated to natural language terms (for instance low and underload). An example of a hypothetical forecast argument is given below (it matches Rule 1 of Section 3.1.1):

– ARG 1: *low* `mental demand` $\rightarrow$ `underload`

In the running example, the selected knowledge base and input values (Fig. 7.* and 7.A) are the same employed in the expert systems and the non-monotonic fuzzy reasoning system (as per Fig. 5 and Fig. 6 respectively). The forecast arguments that are activated from these can be seen in Fig. 7.B.

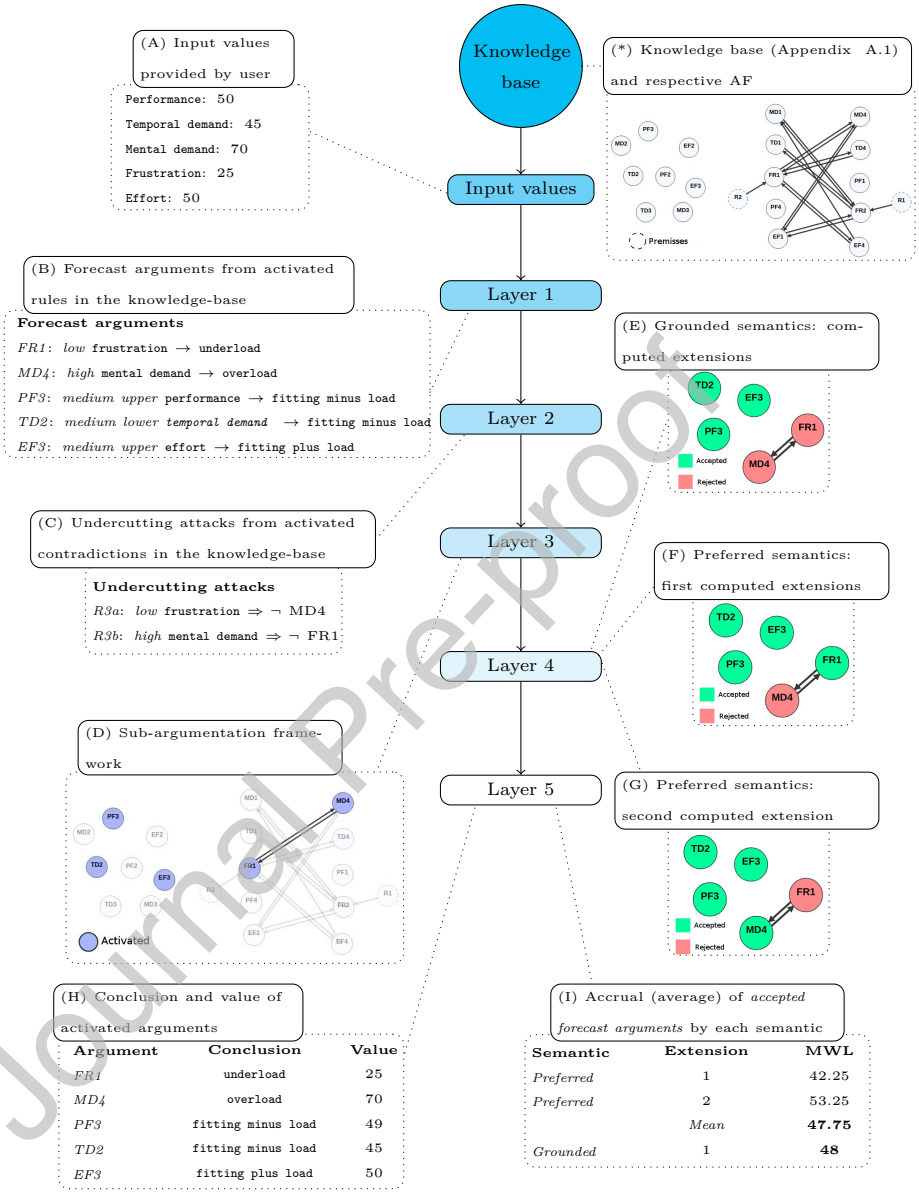



Figure 7: An illustration of a reasoning process of an argument-based defeasible reasoning system. The order of operations is from step (A) to step (I). The argumentation framework related to the knowledge base employed is depicted in step (*).

*3.3.2. Layer 2 - Definition of the conflicts of arguments*

In order to evaluate inconsistencies, the notion of *mitigating argument* (Matt et al., 2010) is introduced. This is formed by a set of premises and an undercutting inference $\Rightarrow$ to an argument B (forecast or mitigating):

$$Mitigating\ argument : premises \Rightarrow \neg B$$

Both forecast and mitigating arguments are special *defeasible rules*, as defined in (Prakken, 2010). Informally, if their premises hold then *presumably* (defeasibly) their conclusions also hold. Different types of mitigating arguments exist in the literature, such as rebuttal and undermining (Prakken, 2010). In this research, the notion of *undercutting attack* is employed for the construction of mitigating arguments and thus enabling the resolution of conflicts. An undercutting attack defines an exception, where some inference carried out in the attacked argument is no longer allowed. Contradictions, such as in Section 3.1.1, represent the information necessary for the construction of undercutting attacks. For example, the corresponding hypothetical mitigating argument that can be constructed from Contradiction 1 (Section 3.1.1) via an undercutting attack is:

– UA1: *high* `effort` $\Rightarrow \neg$ ARG 1

All forecast arguments and undercutting attacks form an *argumentation framework* (AF) (as in Fig. 7.*). Fig. 7.C lists the activated undercutting attacks for the input values (Fig. 7.A). In this example undercutting attacks originate from the contradiction "C3: FR1 AND MD4 *cannot coexist*", listed in Table A.15. It was defined by a domain expert and manually translated as two undercutting attacks.

*3.3.3. Layer 3 - Evaluation of the conflicts of arguments*

At this stage, the created AF can be elicited with data. Forecast and mitigating arguments can be activated or discarded, based on whether their premises evaluate true or false. Consequently, attacks between activated arguments will be evaluated before being activated as well. As mentioned in Section 2.4, attacks usually have a form of a binary relation. In a binary relation a successful (activated) attack occurs whenever both its source (attacking argument) and

its target (argument being attacked) are activated. Another approach that can be adapted in this study is the strength of arguments. In this case, similarly to the definition of rule weights in expert system and fuzzy reasoning, the strength of each argument is extracted from the pairwise comparison procedure of the NASA-TLX. The number of times a feature has been chosen in the pairwise comparison procedure will represent the feature strength, which in turn will also represent the strength of the arguments employing such feature. Consequently, an attack is considered successful only if the strength of its source is equal or greater to the strength of its target.

From the activated forecast/mitigating arguments and successful attacks, a *sub-argumentation framework* emerges (sub-AF), as in Fig. 7.D. This is equivalent to the Abstract Argumentation proposed in Dung (1995).

*3.3.4. Layer 4 - Definition of the acceptance status of arguments*

Given a sub-AF acceptability semantics (Baroni et al., 2011; Dung, 1995) are applied to compute the acceptance status of each argument, that means its acceptability. An argument A is *defeated* by B if there is a valid attack from A to B (Dung, 1995). Not only that, but it is also necessary to evaluate if the defeaters are defeated themselves. Hence, acceptability semantics are aimed at evaluating which arguments are ultimately defeated. A set of non defeated arguments is called *extension*, or a subset of arguments that can be mutually acceptable according to some rationale. Extensions are in turn used in the 5th layer of the reasoning structure of Fig. 3 (p. 12), to produce a final inference. The internal structure of arguments is not considered in this layer, that is why the definition of sub-AF here is equivalent to the notion of *abstract argumentation framework* (AAF) as proposed by Dung (Dung, 1995). An AAF is a pair $< Arg, attacks >$ where: $Arg$ is a finite set of abstract arguments, $attacks \subseteq Arg \times Arg$ is binary relation over $Arg$. Given sets $X, Y \subseteq Arg$, $X$ *attacks* $Y$ if and only if there exists $x \in X$ and $y \in Y$ such that $(x, y) \in attacks$. A set $X \subseteq Arg$ of argument is:

- *admissible* iff $X$ does not attack itself and $X$ attacks every set of arguments

$Y$ such that $Y$ attacks $X$;

- *complete* iff $X$ is admissible and $X$ contains all arguments it *defends*, where $X$ *defends* $x$ if and only if $X$ attacks all attackers of $x$;
- *grounded* iff $X$ is minimally complete (with respect to $\subseteq$);
- *preferred* iff $X$ is maximally admissible (with respect to $\subseteq$)

These represent a few argument-based semantics among others that have been proposed in the literature (Baroni et al., 2011). However, here the focus is on the grounded and preferred semantics. Fig. 7.E, 7.F and 7.G depict different extensions when employing the grounded and preferred semantics in the running example.

*3.3.5. Layer 5 - Accrual of acceptable arguments*

Eventually, in the last step of the reasoning process, a final inference has to be produced. In case multiple extensions are computed, one extension might be favoured over the others. In this study, the cardinality of an extension (number of accepted arguments) is used as a mechanism for selecting the favoured one. Intuitively, a larger extension of arguments might be seen as more relevant than smaller extensions. In case some of the computed extensions have the same highest cardinality, these are all brought forward in the reasoning process. After the selection of the larger extension/s, a single scalar is produced through the accrual of its/their arguments. This is defined by the set of accepted forecast arguments within an extension (those that support a MWL level). Mitigating arguments already completed their roles by contributing to the resolution of conflicting information (layer 4) and thus are not considered in this layer. For each forecast argument, a final scalar is generated for its representation. It follows from the same formula described in Def. 2 (Section 3.1.3). Fig. 7.H lists the values computed for the forecast arguments in the running example. The overall MWL level brought forward by an extension is computed by aggregating the scalars of its forecast arguments. This aggregation can be done in different

ways, for instance considering measures of central tendency. Here, similarly to expert systems, the average and the weighted average are accounted for, with arguments weights being defined the same way as their strengths are. Fig. 7.I concludes the running example by depicting the outcome of each semantics using the average operator. Note that since there are two preferred extensions with the same number of accepted forecast arguments, the outcome of the preferred semantics is the mean of its two extensions.

### *3.4. Participants and procedures*

Three distinct experiments were performed with human subjects. In the first and second, a number of third-level classes were delivered to students at the Technological University Dublin, School of Computer Science, Dublin, Ireland. In the third, nine information seeking web-based tasks of varying difficulty and demand were performed by volunteer participants over three popular web-sites: Google, Wikipedia and Youtube. Subjects were briefed about the study and they were requested to sign a consent form that included data protection and treatment. Privacy and anonymity of participants were in all respects protected by the authors. After each task, a self-reporting questionnaire aimed at assessing mental workload was given to subjects. These can be seen at Fig. A.11, A.13 and A.17 in the Appendixes. Besides completing the questionnaires, in some scenarios participants were required to fill in another scale providing an indication of their experienced mental workload (Fig. 8). This question was designed for triangulation purposes with the assumption that only the person executing a task can precisely self-assess its own experienced mental workload (Moustafa et al., 2017). Table 3 summarises the three experiments, the questionnaires employed and the number of participants. It also mentions the mental workload assessment instrument that will be employed as baseline for comparison purposes.

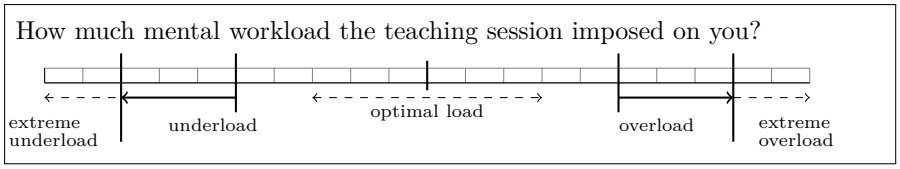



Figure 8: Baseline self-reporting measure of Mental Workload

Table 3: Set up of experiments under evaluation.

| **Label** | **Experimental setting** | **Questionnaire** (Appendixes) | **Features** | **Self Assess.** | **Baseline instruments** | **Records** |
|---|---|---|---|---|---|---|
| $E_a$ | Third-level classes | A.11+A.13 | NASA-TLX [3] | Fig. 8 | NASA-TLX[3] | 230 |
| $E_b$ | Third-level classes | A.17 | Longo [4] | Fig. 8 | Raw TLX [5] & WP [6] | 237 |
| $E_c$ | Seeking web-based information | A.17+A.13 | Longo[4] | None | NASA-TLX[3] & WP[5] | 405 |

### *3.4.1. Third-level classes at Technological University Dublin*

In the first two experiments ($E_a$ and $E_b$, Table 3) students attended third-level classes in the Technological University Dublin and filled either questionnaires A.11+A.13 or A.17 (Appendix A). The set of questionnaires were related to the features being analysed at each experiment. In experiment $E_a$ only features of the NASA-TLX measurement technique were being investigated, while in experiment $E_b$ a larger set of features was being considered for MWL modelling and assessment. Therefore, two distinct sets of data were generated. In total students were from 24 distinct countries (age 19-74, mean 30.9, std = 7.63) and attended four topics of the module 'Research Methods' in the Master of Science: science, scientific method, research planning and literature review. These topics were delivered in three different forms during the semesters of the

[3](Hart & Staveland, 1988).
[4](Longo, 2014).
[5](Hart, 2006).
[6](Tsang & Velazquez, 1996).

academic terms 2015-2018:

1. Traditional direct instruction, using slides projected to a white board;
2. Multimedia video of content. Transformation of the content of the slides of 1 into a multimedia video projected to a white board;
3. Constructivist collaborative activity added to 2.

Table 4 summarises the number of participants for each topic delivered in experiments $E_a$ and $E_b$, grouped by delivery method. It provides additional figures related to the experiments carried out. Further details of these activities are not necessary for this research study, but the reader can find specific information in (Longo, 2018b; Longo & Orru, 2019).

Table 4: Number of students across topics and delivery methods

| **Topic** | **Duration (Mins)** | **Delivery method** | | |
|---|---|---|---|---|
| | | **1** | **2** | **3** |
| Science | [18, 62] | 31 | 70 | 19 |
| Scientific method | [20, 46] | 39 | 36 | 41 |
| Research planning | [10, 68] | 43 | 45 | 41 |
| Literature review | [18, 57] | 41 | 43 | 18 |

*3.4.2. Information seeking web-based tasks*

In the third experiment, nine information seeking web-based tasks of varying difficulty and demand (Table B.19 in the Appendix), were performed by participants over three websites: Google, Wikipedia and Youtube. These websites were selected due to their popularity and assumption that participants were familiar with their interfaces. In this way, situations of underload MWL were expected to happen. If non-popular websites were chosen the chances of spotting underload MWL would be reduced. In addition, the original interface of each web-site was slightly manipulated in order to impose different MWL demands on participants interacting with them, leading to 9 tasks on the original websites and 9 tasks on the modified websites (18 in total). 46 volunteers performed all the tasks in a random order in different days, over 2 or 3 sessions

of approximately 45/70 minutes each. Afterwards, the questions of Table A.17 were answered using a paper-based scale in the range $[0..100] \in \aleph$, partitioned in 3 regions delimited at 33 and 66. 405 valid instances were generated. Despite not being necessary in this study, the reader can obtain more information on the construction of this dataset in (Longo, 2018a, 2017; Longo & Dondio, 2015).

### *3.5. Summary of models and comparative metrics*

Tables C.20, C.21 and C.22 in the Appendix list models built using the reasoning approaches detailed in Sections 3.1, 3.2 and 3.3. Each reasoning approach provides different configuration parameters that can impact results either positively or negatively. Thus, it is important to cover the highest possible number of configurations. Some examples of parameters are heuristics for expert systems, acceptability semantics for defeasible argumentation and fuzzy logic for fuzzy reasoning. Moreover, some types of data might require special configuration parameters, as it is the case in this study for the pairwise comparison procedure of the NASA-TLX. To adapt their use fuzzy reasoning and expert systems implement the notion of rule weights at different stages of their reasoning processes, while defeasible argumentation implements the notion of strength of arguments during the evaluation of conflicts between arguments. The inferential capacity of such models was evaluated by analysing the sensitivity and three forms of validity of their inferences (scalar values). As suggested in Section 2.5, the three forms of validity employed are *convergent*, *face* and *concurrent validity*. The first has been assessed through an analysis of the correlation coefficients of the inferences produced by the designed models and the scores produced by selected baseline instruments. The second has been assessed through an investigation of the mean squared error (MSE[7]) of the inference of each designed model against the mental workload scores reported by students using the scale of Fig. 8. The third has been assessed through an analysis of the correlation coefficients of the

[7] $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i\right)^2$, where $Y$ is the vector of inferences made by the designed models and $X$ the vector of self-reported values.

inferences produced by the designed models and an objective performance measure, in this case task completion time. Finally, sensitivity has been formally assessed by analysing the variance of the distributions generated by inferences of the designed non-monotonic reasoning models followed by a post hoc analysis. Table 5 summarises comparative metrics, the statistical test associated to them and in which experiment they were employed. Before presenting the results and the discussion of the study, Table 6 summarises experiments by reasoning models and statistical tests applied.

Table 5: Comparative metrics, associated statistical tests and experiments that contain information for their application.

| **Property** | **Definition** | **Statistical test** | **Experiment** (Table 3) |
|---|---|---|---|
| Convergent validity | It refers to the extent to which different MWL measures that should be theoretically related, are in fact related. | Correlation coefficient | $E_a$, $E_b$, $E_c$ |
| Face validity | It determines the extent to which a measure of MWL appears effective in terms of its stated aims (measuring mental workload). | Mean Squared Error (MSE)[3] | $E_a$, $E_b$ |
| Concurrent validity | It determines the extent to which a model correlates with an objective performance measure, in this case task completion time. | Correlation coefficient | $E_c$ |
| Sensitivity | It determines the capability of a technique to discriminate significant variations in MWL and changes in resource demand or task difficulty. | Analysis of variance plus post hoc analysis. | $E_a$, $E_b$, $E_c$ |

## 4. Results and discussion

Collected data was used to elicit models listed in Tables C.20, C.21 and C.22 (Appendix C). The evaluation metrics of Table 5 are analysed in the following sections.

Table 6: Streamlined design of experiments under evaluation. Additional details of experiments can be found in Table 3. Full list and detail of all the designed models can be seen in Appendix C. Additional details on statistical tests can be seen in Table 5.

| **Experiment $E_a$** | | |
|---|---|---|
| **Experimental settings** | **Models** | **Analysis** |
| Features: 6, Table A.11 | Expert systems: E{1-4} | Convergent validity |
| Task: Third level classes | Fuzzy reasoning: FL{1-12} and FC{1-12} | Face validity |
| Records: 230 | Defeasible argumentation: A{1-4} | Sensitive |
| **Experiment $E_b$** | | |
| **Experimental settings** | **Models** | **Analysis** |
| Features: 21, Table A.17 | Expert systems: E{5-6} | Convergent validity |
| Task: Third level classes | Fuzzy reasoning: FL{13-18} and FC{13-18} | Face validity |
| Records: 237 | Defeasible argumentation: A{5-6} | Sensitivity |
| **Experiment $E_c$** | | |
| **Experimental settings** | **Models** | **Analysis** |
| Features: 21, Table A.17 | Expert systems: E{7-8} | Convergent validity |
| Task: Seeking web-based | Fuzzy reasoning: FL{19-24} and FC{19-24} | Concurrent validity |
| Records: 405 | Defeasible argumentation: A{7-8} | Sensitive |

### *4.1. Convergent validity*

This property is aimed at determining whether, and to which extent, two MWL inference models are correlated. It is the metric employed to achieve objective 1 (Section 3) and test its research hypotheses. The expectation is a moderate to high correlation coefficient with state-of-the-art MWL measurement techniques, which demonstrates that the designed models are in fact representing and assessing the construct of MWL. Here, the Spearman correlation coefficient was selected because of the non-normality of most of the distributions of the inferences produced by the designed models. Formally, this was confirmed by the Shapiro-Wilk test, which was not greater than the alpha level set (alpha=0.05). Fig. D.27, p. 86, depicts the density plots of the inferences produced by all models, while Fig. 9, 10 and 11 depict the Spearman correlation coefficients of their inferences and those of the baseline instruments.

From Fig. 9 it is possible to observe that the models designed for experiment $E_a$ could all achieve a medium to high correlation coefficient with the

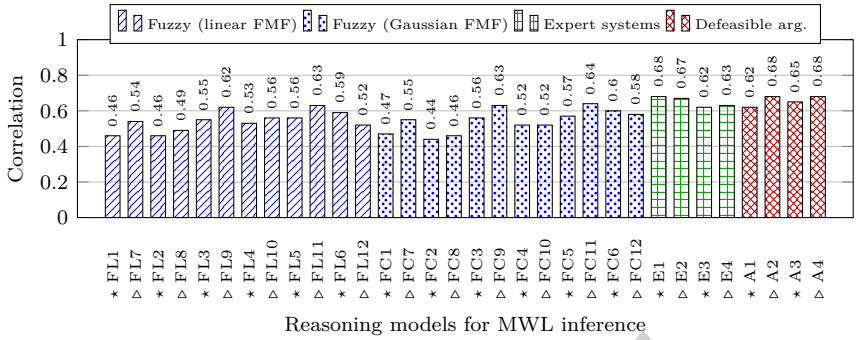
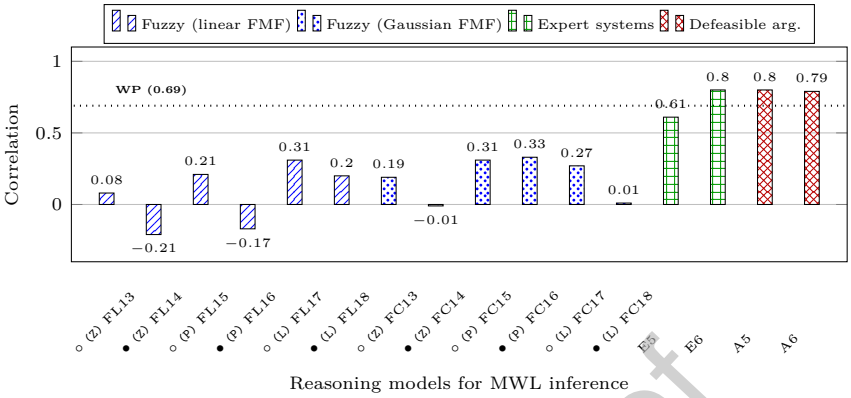


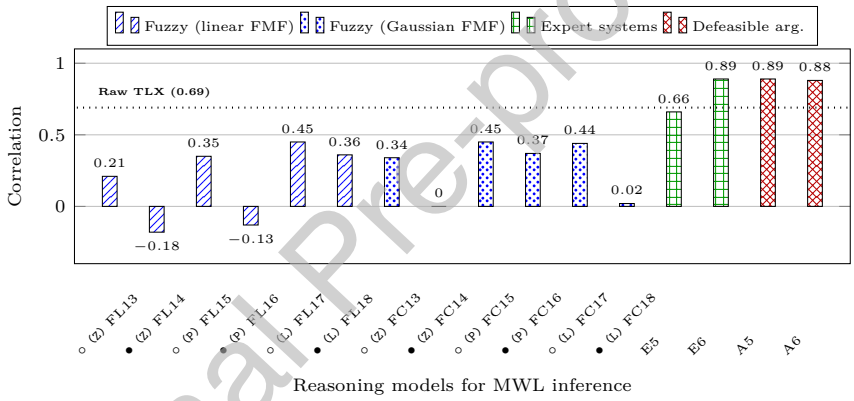

Figure 9: Spearmans correlation coefficients between NASA-TLX scores and inferences of designed models for experiment $E_a$ ($p < 0.05$). Models employing the pairwise comparison information of the NASA-TLX are labelled with an inferior $\triangleright$, while those not employing it are labelled with an inferior $\star$.

NASA-TLX baseline instrument (coefficients: 0.44 - 0.68). This demonstrates the capacity of the investigated reasoning approaches to allow the development of models to represent and assess the construct of MWL in experiment $E_a$, since they are in line with the baseline instrument. Models employing the pairwise comparison information of the NASA-TLX (labelled with an inferior $\triangleright$) had in general a slightly higher correlation coefficient than analogous models not employing this information (labelled with an inferior $\star$). Yet, a few exceptions can also be observed, such as: $FL6 \times FL12$, $FC6 \times FC12$ and $E1 \times E2$. This indicates that acceptable MWL inference models can be designed with less information than the original NASA-TLX instrument.

Fig. 10 depicts the correlation coefficients of the designed models and selected baseline instruments in experiment $E_b$: the Raw TLX in Fig. 10.a and the Workload Profile in Fig. 10.b. Contrarily to results of experiment $E_a$, not all models could achieve a moderate/high convergent validity. In detail, fuzzy models employing the mean of max defuzzification approach had the lowest correlation coefficients (labelled with an inferior $\bullet$) against both Raw TLX and Workload Profile. In addition, there is a stark contrast when these are compared to their counterparts employing the centroid defuzzification approach (labelled




(a) Correlations against RAW TLX




(b) Correlations against Workload Profile

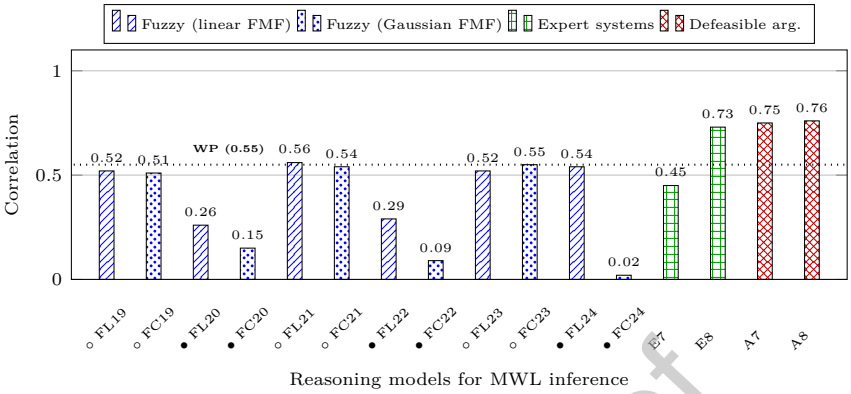
Figure 10: Spearmans correlation coefficients between Raw TLX scores (a), Workload Profile scores (b) and inferences of designed models for experiment $E_b$ ($p < 0.05$). Inferior symbols are used to represent: centroid defuzzification approach (○), mean of max defuzzification approach (●), fuzzy logic operator Zadeh `(Z)`, Product `(P)` and Łukasiewicz `(L)`.

with an inferior ○), be it among models of linear fuzzy membership functions or Gaussian fuzzy membership functions. This is a strong indication that the mean of max is not a suitable parameter within a model to assess MWL in experiment $E_b$, regardless of the fuzzy operator or shape of the fuzzy membership function employed. As for the FMFs, it is also possible to notice some differences when employing different fuzzy operators. For instance, models em-

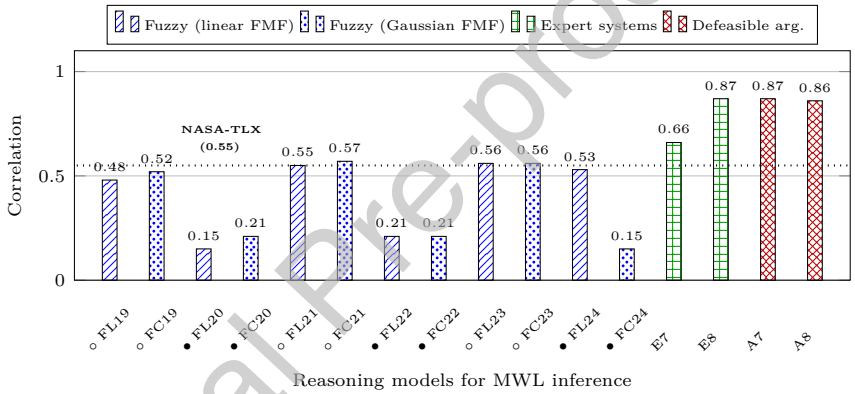
ploying the *Zadeh* and *Product* operator (labelled with an inferior (Z) and (P) respectively) tend to have a higher correlation coefficient when employing Gaussian FMFs ($FL13 \times FC13, FL14 \times FC14, FL15 \times FC15$ and $FL16 \times FC16$), while models employing the *Łukasiewicz* operator (labelled with an inferior (L)) present the inverse behaviour, with similar to lower correlation coefficient for models of Gaussian FMFs ($FL17 \times FC17$ and $FL18 \times FC18$). Among expert system models, also note a lower correlation coefficient for $E5$ whose heuristic is $h_1$ (the average of surviving rules inferring the MWL level supported by the greatest number of surviving rules) than $E6$ whose heuristic is $h_3$ (average of all surviving rules). This suggests that the process of filtering surviving rules ($h_1$) instead of taking all of them into account ($h_3$) for the final inference might not be a good strategy. In other words, it also suggests that all surviving rules might be of equal importance on the expert system reasoning process, regardless if their conclusions are the same or not of other surviving rules. Finally, defeasible argumentation models show very much alike correlation coefficients among them, suggesting no difference exists between preferred and grounded semantics in this experiment.

Fig. 11 depicts the results for experiment $E_c$. It is possible to observe some similar results to the convergent validity in $E_b$: the same correlation trend between the designed models and the distinct baseline instruments (NASA-TLX and WP), better correlation for expert systems employing heuristic $h_1$ ($E_7$) instead of $h_3$ ($E_8$), no significant difference between defeasible argumentation models and worse performance in general for fuzzy models employing the mean of max defuzzification approach (labelled with an inferior $\bullet$). However, the impact of the FMFs shape is not analogous as that of previous findings, in fact it is not possible to observe a significant difference in their correlation coefficients except for models FL24 and FC24.

In summary, it is worth highlighting some common findings and differences related to the convergent validity of models across reasoning approaches. For instance, the expert system and defeasible argumentation reasoning approaches appear to be more robust for modelling the construct of MWL across the dif-




(a) Correlations against NASA-TLX




(b) Correlations against Workload Profile

Figure 11: Spearmans correlation coefficients between NASA-TLX scores (a), Workload Profile scores (b) and inferences of designed models for experiment $E_c$ ($p < 0.05$). Fuzzy models employing the centroid defuzzification approach are labelled with an inferior $\circ$, while those employing the mean of max are labelled with an inferior $\bullet$.

ferent internal configurations of models and the different knowledge bases employed. This is demonstrated by the overall higher Spearman correlation coefficient between such models and baseline instruments across experiments (in the range 0.62 - 0.89 for defeasible argumentation and 0.45 - 0.89 for expert systems). Contrarily, parameters of the fuzzy reasoning models seem to lead to the development of models that are more sensitive to the knowledge bases employed.

Even when selecting the same fuzzy operator, the same defuzzification method and the same fuzzy membership functions, fuzzy models can behave in stark contrast when compared to baseline instruments. For instance, while model FC6 presents a high correlation coefficient (0.6) with NASA-TLX in experiment $E_a$, the analogous model FC24 with same parameters, except for knowledge base input, presents a low (0.15) correlation coefficient with NASA-TLX in experiment $E_c$. This suggests that there is no fuzzy logic, defuzzification method or fuzzy membership functions better than others, having these to be selected in a case by case analysis with the knowledge base. This can also be observed by the similar correlation coefficients of fuzzy models in experiment $E_a$ (overall coefficients: 0.44 - 0.64) and contrasting correlation coefficients in experiments $E_b$ and $E_b$ (respectively in ranges -0.21 - 0.45 and 0.02 - 0.57).

### *4.2. Face validity*

This property is aimed at determining the extent to which a measure of MWL appears effective. It is one of the metrics employed to achieve objective 2 (Section 3) and test its research hypotheses. It was analysed according to the mean square error (MSE) of produced inferences and self-reported MWL values (Fig. 8, p. 34). Fig. 12 and 13 depict the results for experiments $E_a$ and $E_b$ respectively. Experiment $E_c$ does not present information about self-reported MWL values. Overall, the majority of models across reasoning approaches could achieve similar or better MSE than baseline instruments. The higher discrepancy, and worst performance (higher MSE), is given by fuzzy models employing the mean of max defuzzification approach (labelled with an inferior $\bullet$). Similarly to convergent validity, defeasible argumentation models demonstrated robustness across the three experiments and expert system models performed better when employing heuristic $h_2/h_4$ (the average/weighted average of *all* surviving rules, labelled with an inferior $+$).

As for experiment $E_a$, a significant difference has been found between models employing the pairwise comparison information of the NASA-TLX and those not employing it. Among fuzzy models with linear FMF there is an average decrease

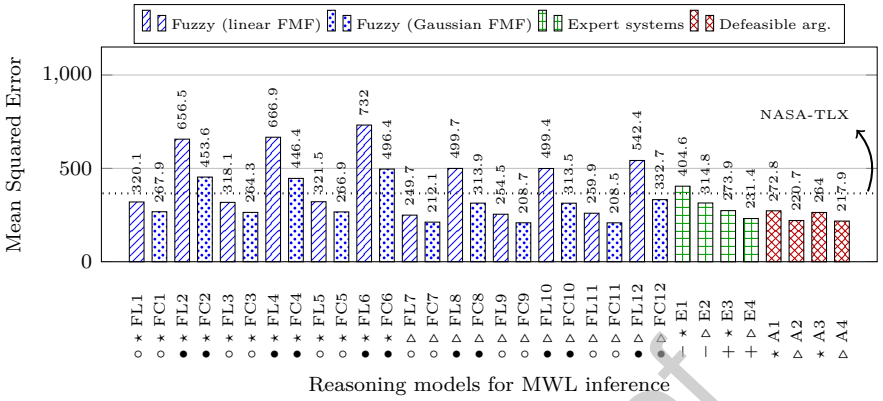



Figure 12: Mean squared error of each designed model for experiment $E_a$ and baseline instrument NASA-TLX. Inferior symbols are used to represent: centroid defuzzification approach ($\circ$), mean of max defuzzification approach ($\bullet$), heuristics $h_1$ ($-$) and $h_3$ ($+$), use (respectively no use) of the the pairwise comparison information of the NASA-TLX ($\triangleright$, respectively $\star$).

of 24% MSE when employing the pairwise comparison information ($FL\{1-6\} \times FL\{7-12\}$), while fuzzy models with Gaussian FMFs present a decrease of 27.6% ($FC\{1-6\} \times FC\{7-12\}$). A similar trend is observable in expert system models, with a decrease of 19.5% ($E2, E4 \times E1, E3$), and defeasible argumentation models, with a decrease of 18.4% ($A2, A4 \times A1, A3$). In contrast to convergent validity, the use of the information from the pairwise comparison procedure demonstrated to have a stronger impact in face validity, even when used in distinct ways by the investigated reasoning approaches. In other words, despite not being essential to achieve high convergent validity with baseline instruments, the information from the pairwise comparison procedure seems to have a positive impact on the quality of the produced inferences according to the analysis of face validity.

### *4.3. Concurrent validity*

Aimed at determining the extent to which a model correlate with an objective performance measure, in this case task completion time, concurrent validity was also assessed through an analysis of correlation coefficients between the designed models and baseline instruments in experiment $E_c$. A reminder that

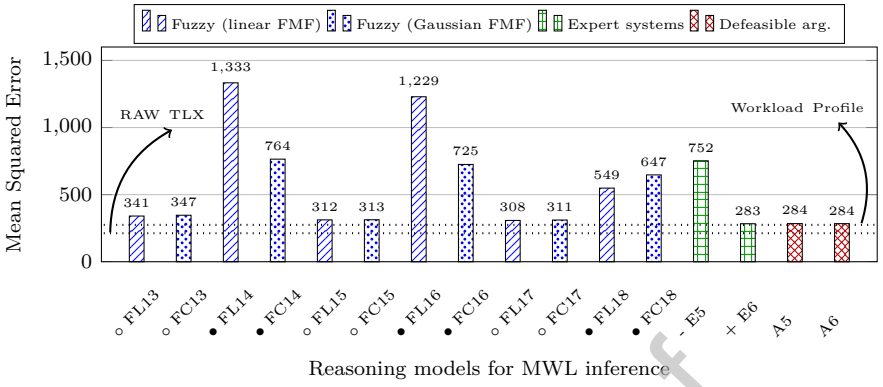



Figure 13: Mean squared error of each designed model for experiment $E_b$ and baseline instruments RAW TLX (lower dotted line at 212.57) and Workload Profile (upper dotted line at 274.91). Inferior symbols are used to represent: centroid ($\circ$) and mean of max defuzzification approach ($\bullet$) and heuristics $h_1$ ($-$) and $h_3$ ($+$).

in the experiments $E_a$ and $E_b$ an objective performance measure has not been gathered. From Fig. 14 it is possible to note that even the baseline instruments do not have a high Spearman correlation coefficient with task completion time (NASA-TLX: 0.28 and WP: 0.18), while most of the designed models present a coefficient between 0.2 and 0.26, lying between the two baseline instruments. This suggests that the investigated reasoning approaches, when set up with certain parameters, are as good as the baseline models. The exceptions presenting a lower correlation coefficient are the fuzzy models of Gaussian FMFs employing the mean of max defuzzification approach ($FC20$, $FC22$ and $FC24$) and the expert system $E7$ employing heuristic $h_1$. This trend is very similar to the one depicted for convergent validity in Fig. 11, suggesting that these combinations of parameters (Gaussian FMFs + mean of max for fuzzy models and heuristic $h_1$ for expert system models) do not help to create robust models of MWL. It is also worth noting that fuzzy models $FL20$ and $FL22$ could achieve a favourable correlation coefficient with task completion time, despite having low convergent validity. It suggests that models with low convergent validity might also produce acceptable inferences.

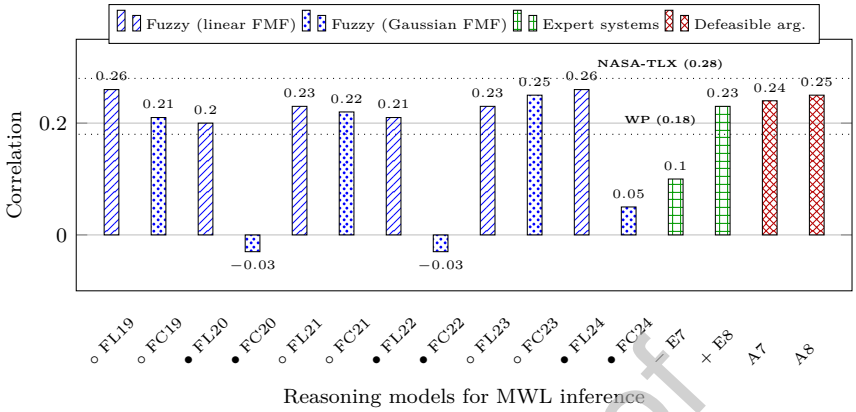
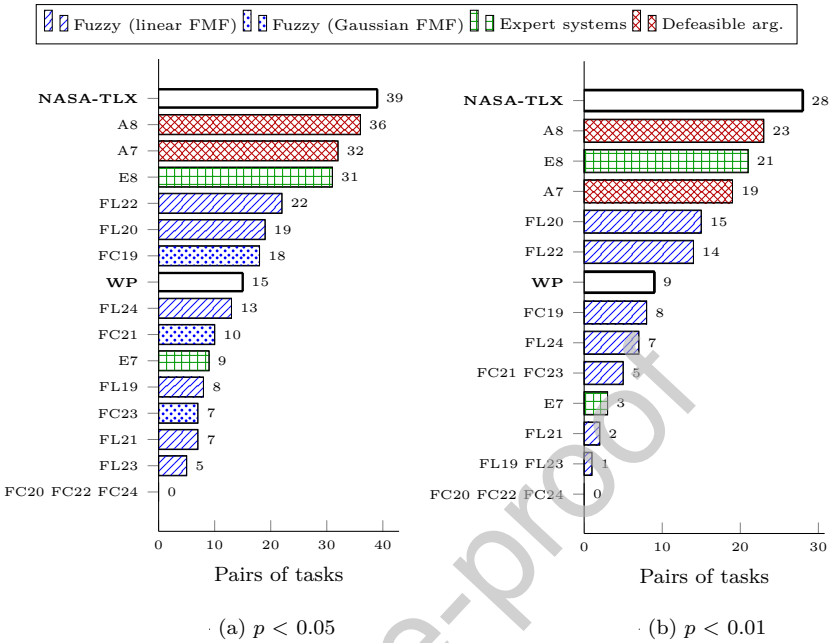



Figure 14: Spearmans correlation coefficients between task completion time and the inferences of designed models for experiment $E_c$ ($p < 0.05$). Only 288 instances (out of 405 of experiment $E_c$) have an associated time due to measurement errors. Inferior symbols are used to represent: centroid ($\circ$) and mean of max defuzzification approach ($\bullet$) and heuristics $h_1$ ($-$) and $h_3$ ($+$).

### *4.4. Sensitivity*

In line with other studies (Rubio et al., 2004; Longo, 2015), sensitivity was assessed by performing an analysis of variance over the MWL distributions generated by the designed models and the baseline instruments. The aim is to investigate the capability of a model to discriminate significant variations in MWL and changes in resource demand or task difficulty. In detail, the non-parametric Kruskal-Wallis H test was performed over the MWL distributions generated by each model. As mentioned before, normality of the distribution of most of the models was not found according to the Shapiro-Wilk test. Hence, the equivalent of one-way ANOVA could not be employed. Baseline instruments and designed models for experiments $E_a$ and $E_b$ were not capable of rejecting the null hypothesis of same distribution of MWL scalars across tasks ($p < 0.01$). In these experiments, it can be argued that the performed tasks are of pedagogical nature and are of similar complexity, since all classes are related to the same general topic: 'Research Methods'. Thus, it is difficult to create procedures that can statistically and significantly affect overall MWL (Longo, 2018c).

As for experiment $E_c$, the null hypothesis of same distribution of MWL




(a) $p < 0.05$ (b) $p < 0.01$

Figure 15: Sensitivity of MWL models designed for experiment $E_c$ with Games-Howell post hoc analysis. The maximum pairwise comparisons of 18 tasks is $\binom{18}{2} = 153$. Baseline instruments are depicted in bold.

scalars across tasks was rejected. That means that there exist models that lead to significantly different inferences when used to evaluate the MWL imposed by the web-based tasks. However, the Kruskal-Wallis H test does not tell exactly which pairs of tasks executed by participants are different from each other. Consequently, a post hoc analysis was performed and the Games-Howell test was chosen because of unequal variances of the distributions under analysis. Fig. 15 depicts how many pairs of tasks each model was capable of differentiating at two significance levels ($p < 0.05$ and $p < 0.01$). As it can be observed, similarly to convergent and concurrent validity, defeasible argumentation models and expert system $E8$ outperformed the other models. When compared to the baseline instruments, results for these models are in between the NASA-TLX and the WP for both significance levels. Despite the high sensitivity of defeasible argumentation models, it is possible to observe a slight difference between

them, with a better performance achieved by model A8 whose argumentation semantics is the *preferred* semantics. Among fuzzy models, it is worth noting that the best performance is given by $FL20$ and $FL22$. It strengthens the results of concurrent validity, suggesting that models of low convergent validity might produce satisfactory inferences. Another interesting observation comes from model $FC19$. In spite of presenting similar convergent and concurrent validity with its linear counterpart ($FL19$), in this case its sensitivity was superior, being close to or better than WP, while $FL19$ was always distant from the baseline instruments. It shows that Gaussian FMFs can provide more sensitive models when employed with certain fuzzy operators and defuzzification approaches (in this case *Zadeh* and centroid respectively). Other fuzzy models demonstrated to have poor sensitivity, underperforming the baseline models. In detail, as expected by convergent and face validity analysis of experiment $E_c$, fuzzy models of Gaussian FMFs employing the mean of max defuzzification approach led to the worst performance, not being able to statistically differentiate between any pair of tasks.

### *4.5. Internal configurations of models and interpretations*

Quantifications of the validity and sensitivity of the developed models suggest that, in general, the investigated reasoning approaches can be successfully employed for mental workload modelling and assessment. Nonetheless, the analysis across different experiments and evaluation metrics seems to indicate a contrasting performance when particular parameters of distinct reasoning techniques are employed. Table 7 summarises average results across experiments for the designed models grouped by internal parameters. Some results are in fact a single value, and so, have no standard deviation reported. For the other cases, Figures 16 - 22 depict the respective boxplots.

Most negative impacts seemed to be caused by the application of the mean of max defuzzification approach by fuzzy models and heuristics for the refinement of surviving rules by expert system models ($h_1/h_2$). These lead to the development of models that, in average, underperformed in all evaluation metrics

Table 7: Average and standard deviation of evaluation metrics in all experiments by specific parameters of each reasoning approach. Bold numbers are used to represent the best results among the pairwise comparisons inside the table.

| **Reasoning technique** | **Parameter** | **Average Validity ($\sigma$)** | | | **Avg. Sensitivity ($\sigma$)** |
|---|---|---|---|---|---|
| | | **Convergent** | **Face** | **Concurrent** | $p < 0.05$ / $p < 0.01$ |
| Fuzzy reasoning | Mean of Max | 0.27 (0.25) | 622.28 (277.81) | 0.11 (0.12) | 5.3 (6.0) / 2.16 (2.4) |
| | Centroid | **0.46** (0.15) | **282.07** (44.71) | **0.23** (0.01) | **8.3 (4.8) / 3.8 (1.6)** |
| | Linear | 0.37 (0.25) | 521.8 (316.44) | **0.23** (0.02) | **7.6 (3.8) / 3.8 (0.7)** |
| | Gaussian | **0.38** (0.21) | **382.89** (172.95) | 0.11 (0.12) | 6 (6.9) / 2.16 (2.7) |
| | Rule weight | 0.57 (0.06) | **324.58** (121.82) | - | - |
| | No rule weight | 0.57 (0.05) | 434.21 (170.65) | - | - |
| Expert systems | $h_1/h_2$ | 0.62 (0.09) | 490.53 (231.03) | 0.1 (-) | 9 (-) / 4 (-) |
| | $h_3/h_4$ | **0.75** (0.09) | **262.85** (27.62) | **0.23** (-) | **21 (-) / 14 (-)** |
| | $h_1/h_3$ | **0.69** (0.02) | 333.27 (92.44) | - | - |
| | $h_2/h_4$ | 0.67 (0.02) | **273.09** (58.92) | - | - |
| Defeasible argument. | Preferred | **0.77** (0.07) | **255.29** (33.90) | **0.25** (-) | **23 (-) / 16 (-)** |
| | Grounded | 0.76 (0.09) | 259.27 (33.86) | 0.24 (-) | 21 (-) / 15 (-) |
| | Strength of arg. | **0.7** (0.0) | **219.31** (2.02) | - | - |
| | Binary relation | 0.68 (0.02) | 268.4 (6.2) | - | - |

(validity, sensitivity) and, in the case of fuzzy models, also tend to have a much higher standard deviation when compared to their counterparts: the centroid approach and the heuristics $h_3/h_4$. The explanation for such discrepancy might lie in the role of the mean of max defuzzification approach and the role of the heuristics $h_1/h_2$ in their respective models. Note that despite being employed by distinct reasoning techniques these roles might in fact be related. While the mean of max defuzzification approach selects only the rules whose conclusion(s) have the highest degree of truth, the refinement of surviving rules by heuristics $h_1/h_2$ discards rules not inferring the MWL level supported by the greatest number of surviving rules. Thus, these can be seen as apparently unsuccessful attempts to resolve conflicts among rules by selecting some of them believed to be suitable for inferring a final MWL scalar.

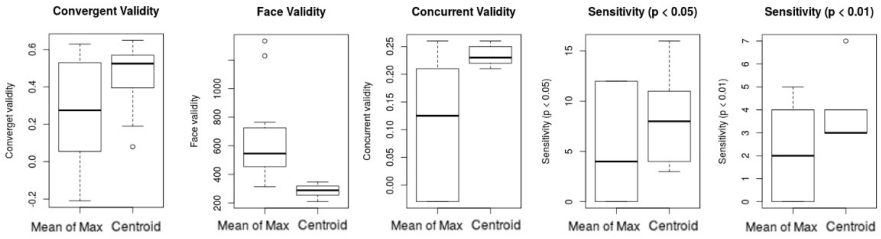



Figure 16: Boxplots of evaluation metrics by defuzzification approach of fuzzy reasoning models.

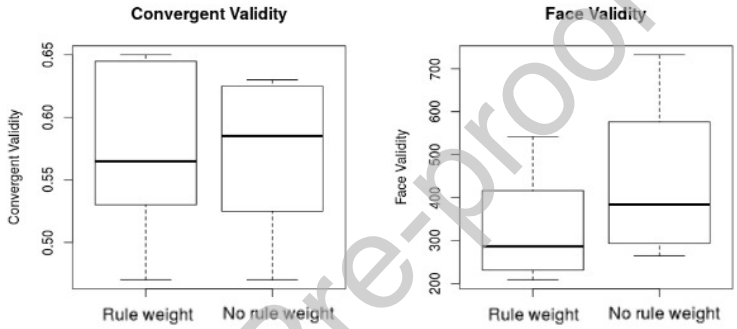



Figure 17: Boxplots of evaluation metrics by application of rule weight or not on fuzzy reasoning models.

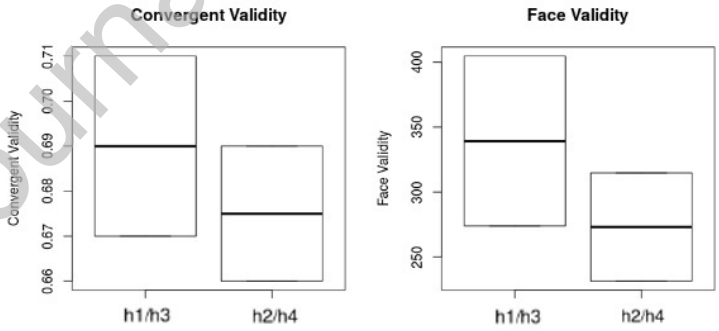



Figure 18: Boxplots of evaluation metrics by heuristics applying weighted average of arguments ($h_2/h_4$) and heuristics applying regular average of arguments ($h_1/h_3$) on expert system models.

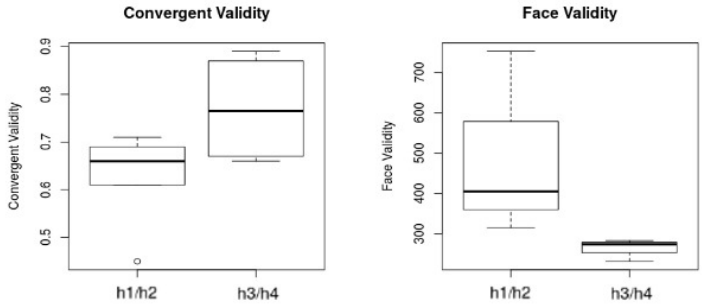



Figure 19: Boxplots of evaluation metrics by heuristics averaging all arguments ($h_1/h_2$) and heuristics averaging a subset of arguments ($h_3/h_4$) on expert system models.

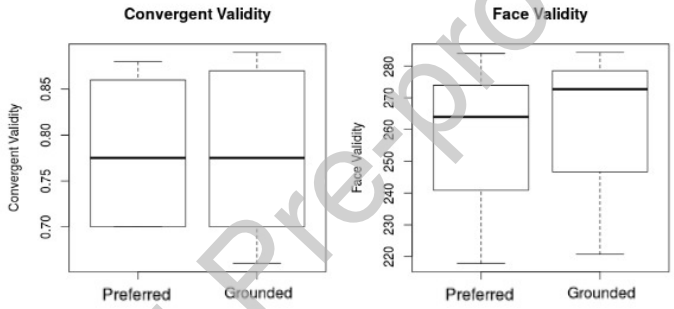



Figure 20: Boxplots of evaluation metrics by acceptability semantics on defeasible argumentation models.

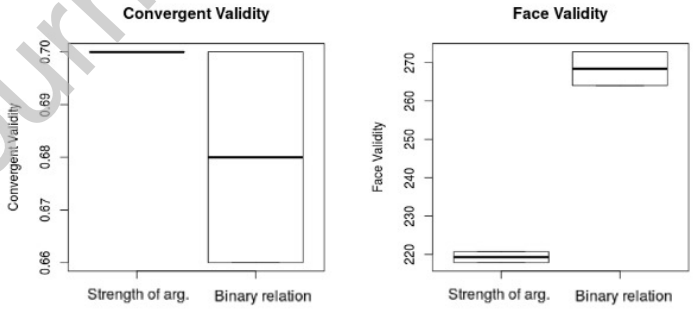



Figure 21: Boxplots of evaluation metrics by attack relation on defeasible argumentation models.

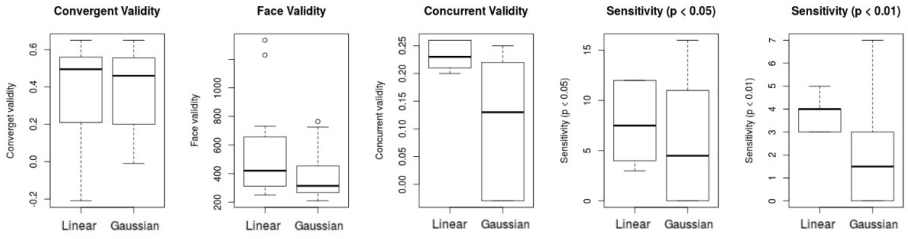



Figure 22: Boxplots of evaluation metrics by fuzzy membership function shape of fuzzy reasoning models.

In contrast, it is worth noting the robustness of defeasible argumentation, with only a slight performance variance among its models across distinct evaluation metrics and experiments. This suggests that defeasible argumentation has a greater capacity of resolving conflicts among rules, thus optimally handling non-monotonicity. It is also interesting to observe the small differences between results generated by models employing the preferred semantics and the grounded semantics. These semantics diverge when multiple extensions are generated by the preferred semantics, since the grounded semantics can only output a single extension. In case of multiple extensions the one(s) with the highest cardinality is (are) selected. Similarly to the heuristics $h_1/h_2$ and mean of max defuzzification approach, this selection of an extension among multiple ones is also an attempt of conflict resolution. However, in the case of defeasible argumentation, produced results are stronger, suggesting that the conflict resolution strategy of defeasible argumentation is likely stronger than the conflict resolution strategy of fuzzy reasoning and expert systems.

From Table 7 it is possible to inspect and further spot other differences between particular parameters employed by reasoning models. For instance, the impact of using the extra information from the pairwise comparison of the NASA-TLX is similar and, as expected, positive across all reasoning approaches. This use is made by fuzzy models employing rule weights, expert system models employing heuristics $h_2/h_4$ and defeasible argumentation models employing strength of arguments. In these cases the convergent validity is preserved and

the mean squared error between produced inferences and self-reported MWL values (face validity) is reduced. This can be observed in Fig. 17, 18 and 21 which compare models using and not using the information from the pairwise comparison. At last, the difference between linear and Gaussian FMFs on fuzzy models is not absolute. While models of Gaussian FMFs present analogous average results for convergent validity and better average results for face validity, linear models seem to have better average results for concurrent validity and sensitivity. This observation can also be supported by the boxplot comparison of Fig. 22. Such mixed results do not allow the drawing of conclusions in regards to the impact of the shape of FMFs on MWL modelling and assessment.

### *4.6. Discussion*

The overall medium to high degree of convergent validity of the investigated models indicated that their inferences can be considered valid, as per alternate hypothesis of objective 1 (Section 3). As a consequence, the findings from the analysis of the face validity, concurrent validity and sensitivity can be considered consistent, quantifying the extent by which the designed reasoning models can represent MWL. This analysis seems to also indicate a better inferential capacity of the defeasible argumentation models, or in this case, a better capacity of producing inferences with improved face validity, improved concurrent validity and improved sensitivity. This conclusion was further supported by the examination of average results of the designed models when grouped by their configuration parameters. Defeasible argumentation models presented the lowest standard deviations of such averages, demonstrating robustness across its internal configurations. This advantage was inspected over two other non-monotonic reasoning approaches namely fuzzy reasoning and expert systems. It also held despite the underlying knowledge bases employed. Comparable results were only achieved by expert systems employing one of the designed heuristics for conflict resolution. This similarity is likely due to the lower amount of conflictual rules employed within knowledge bases when elicited with real-world data. For example, the knowledge base solely built upon the NASA-TLX attributes

(Appendix A.1) can only have up to six arguments that can be activated given input data. Thus, the requirement of further comparisons for knowledge bases of higher topological complexity might be reasonable. Nonetheless, defeasible argumentation models consistently showed a higher correlation with baseline models, a significantly lower mean squared error against the subjective perception of mental workload rated by participants, an analogous concurrent validity to the baseline models and a sensitivity in-between the NASA-TLX and WP models. This suggests the potential of defeasible argumentation as a modelling tool for knowledge bases characterised by uncertainty, partiality and conflictual information. A summary of the comparison of defeasible argumentation against fuzzy reasoning and expert systems across experiments is listed in Table 8 for convergent validity and Table 9 for the other evaluation metrics. Based on these the acceptance statuses of Hypotheses 1 and 2 (Section 3) are listed in Table 10.

Table 8: Status of reasoning approaches according to convergent validity. A ✓ means medium to high convergent validity for all the designed models employing the reasoning approach.

| **Reasoning approach** | Convergent validity | | |
|---|---|---|---|
| | $\boldsymbol{E_a}$ | $\boldsymbol{E_b}$ | $\boldsymbol{E_c}$ |
| Expert systems | ✓ | ✓ | ✓ |
| Fuzzy reasoning | ✓ | Partially | Partially |
| Defeasible argumentation | ✓ | ✓ | ✓ |

Table 9: Status of defeasible argumentation (DA) compared to fuzzy reasoning (FR) and expert systems (ES) according to sensitivity, face validity and concurrent validity across the 3 experimental settings. Comparison symbols are used to represent equal (=), better (<) and considerably better (≪) results on *average* for models built upon defeasible argumentation. A (−) means not applicable. The reasoning approach employed by the best-performing model is listed in the last row.

| **Comparison approach** | **Sensitivity** | | | **Face validity** | | | **Concurrent validity** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\boldsymbol{E_a}$ | $\boldsymbol{E_b}$ | $\boldsymbol{E_c}$ | $\boldsymbol{E_a}$ | $\boldsymbol{E_b}$ | $\boldsymbol{E_c}$ | $\boldsymbol{E_a}$ | $\boldsymbol{E_b}$ | $\boldsymbol{E_c}$ |
| Expert systems | = | = | < | < | < | − | − | − | < |
| Fuzzy reasoning | = | = | ≪ | < | < | − | − | − | < |
| **Best model** | − | − | DA | DA | ES/DA | − | − | − | FR/DA |

Table 10: Acceptance status of the hypotheses of the research study.

| | |
|---|---|
| **Hypothesis 1** | Non-monotonic reasoning models will demonstrate moderate to high convergent validity with baseline instruments. |
| **Acceptance status** | *Accepted* by defeasible argumentation and expert systems. *Partially accepted* by fuzzy reasoning, with some models presenting low convergent validity. |
| **Hypothesis 2** | Defeasible argumentation models will demonstrate higher sensitivity, higher concurrent validity and higher face validity than fuzzy reasoning and expert system models. |
| **Acceptance status** | *Partially accepted.* On average sensitivity and validity are consistently better for defeasible argumentation. By individual models, defeasible argumentation has better results overall, but expert systems and fuzzy reasoning can produce results of equivalent face and concurrent validity on certain experiments. |



## 5. Conclusion and future work

This study presented an extensive comparison of non-monotonic rule-based reasoning techniques for the practical problem of mental workload modelling. These techniques are promising not only because they can approximate the inferential capacity of a knowledge representation and reasoning application, but they also offer a flexible approach for translating different knowledge bases and beliefs of domain experts into computational rules. Furthermore, they support the creation of models that can be falsified, replicated and extended, thus enhancing the understanding of the construct of mental workload itself and possibly other applications of interest. Such advantages, for instance, are not provived by data-driven techniques, even the ones able to produce interpretable solutions. Hence, if they are to be used in other domains of application and by other domain experts, it is necessary to perform a meticulous – and not performed before – examination of one of their crucial aspects namely inferential capacity. In particular, the inferential capacity of expert systems, non-monotonic

fuzzy reasoning and defeasible argumentation models was examined. A set of models, for each reasoning approach, was created following the structures employed in the literature. For instance, expert systems adopted the common two internal components: a knowledge base and an inference engine (Durkin & Durkin, 1998). Fuzzy reasoning models followed the structure of a typical Mamdani fuzzy inference process (Mamdani, 1974). Defeasible argumentation models were constructed based on a 5-layer schema upon which argumentation systems are typically built (Longo, 2016). Nonetheless, the implementation of the non-monotonicity property was not straightforward for expert systems and fuzzy reasoning. The former required different heuristics for aggregating rules and inferring MWL as a numerical index. Usual conflict resolution strategies of expert systems could not be employed due to the nature of the domain, which required all the reasoning to be made in a single step. The latter, fuzzy reasoning, had non-monotonicity implemented by using Possibility Theory, having truth values, named *possibility* and *necessity*, associated to each piece of information. Possibility allowed fuzzy reasoning models to determine the extent to which data fails to refute its truth, while necessity represented the usual truth values of fuzzy logic. Besides such adaptations, the investigation of configuration parameters was also performed for each reasoning technique for tuning purposes.

Findings indicated how models or a subset of models built upon the three reasoning techniques had a good convergent validity with three selected baseline models of mental workload: the NASA Task Load Index (Hart & Staveland, 1988), its RAW extension (Hart, 2006) and the Workload Profile (Wickens, 1991). The designed inferential models were elicited with three knowledge bases, three distinct sets of data and assessed according to common evaluation metrics of MWL, namely sensitivity and validity. Findings revealed a good convergent validity against baselines, suggesting how constructed reasoning models can actually model the underlying construct: mental workload. In detail, fuzzy reasoning presented varied results due to the higher number of available configuration parameters, providing greater flexibility but limiting its applicability and

use by domain experts. Equivalently, applicability and use by domain experts is also limited in expert systems due to varied results. Some of these are inferior to the results of defeasible argumentation when employing one set of heuristics, but similar when employing the complementary set of heuristics. Hence, the analysis of knowledge bases of topologies of higher complexity is a possible direction of future research. Finally, defeasible argumentation showed additional robustness compared to fuzzy reasoning and expert system models according to overall validity and sensitivity, holding despite the parameters being employed and underlying knowledge base. The originality of this research lies in the quantification of the impact of defeasible argumentation. It is a result of a thorough empirical research in two real-world experimental settings employing primary data gathered from humans, and three knowledge bases produced with the aid of human experts. All these elements provide some generalisability to the results and also help on identifying situations in which the non-monotonic reasoning approaches are likely better or worse to each other. It does not verify which of them is ultimately better. Other representations of fuzzy reasoning systems could give better outcomes, the same way other representations of defeasible argumentation and expert systems could also give better outcomes. However, this research has produced an extensive number of inferential models of different configurations. Hence, it contributes to the field of logic and non-monotonic reasoning by better situating defeasible argumentation among similar reasoning approaches and illustrating a replicable comparison process between them. This comparison has been performed using an application whose knowledge bases are formed by uncertain information. In spite of that, quantitative metrics of evaluation could still be employed since these were pre-defined in the literature of mental workload. Because of that, this study is even more significant to the field of non-monotonic reasoning, showing how a quantitative evaluation process can be performed in a uncertain context.

Future work will concentrate on investigating knowledge bases of different and increased topological complexities. In addition, this study is limited for being performed in a single domain of application. Comparisons performed in

other areas might enhance results and extend its generalisability. One possible adequate field of comparison in the domain of knowledge representation is computational trust modelling (Parsons et al., 2010; Dondio & Longo, 2014, 2011). In order to improve the acceptance of defeasible argumentation for non-monotonic activities, the investigation of its explanatory capacity is also suggested. Higher explanatory capacity might lead to higher levels of adoption not only in the field of knowledge representation and reasoning but also in areas such as health-care and autonomous vehicles. Previous work (Rizzo & Longo, 2018) have attempted to perform a preliminary qualitative analysis of defeasible argumentation and non-monotonic fuzzy reasoning in terms of a few properties for explainability analysis from explainable AI. However, explainability is a complex concept and additional examination should be performed so as to assess the usability and effectiveness of explanations provided. Another line of research may be pursued by increasing the explanatory capacity of models built upon defeasible argumentation through the addition of new explainable layers. For instance the argumentation semantics designed in (Fan & Toni, 2015) for giving explanations to arguments. Lastly, the application of hybrid reasoning techniques, such as neuro-fuzzy systems (Nauck et al., 1997), genetic fuzzy systems (Cordón et al., 2004) and fuzzy argumentation (Dondio, 2017) is recommended. Their investigation might lead to possible alternative solutions capable of presenting strong inferential and explanatory capacity for non-monotonic reasoning problems.

**Acknowledgements**

Lucas Middeldorf Rizzo would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for his Science Without Borders scholarship, proc n. 232822/2014-0. The authors also gratefully acknowledge the assistance of the anonymous reviewers for their many invaluable comments and suggestions.

**References**

**References**

Amgoud, L., Ben-Naim, J., Doder, D., & Vesic, S. (2017). Acceptability semantics for weighted argumentation frameworks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI* (pp. 56–62).

Amgoud, L., Parsons, S., & Maudet, N. (2000). Arguments, dialogue, and negotiation. In *Proceedings of the 14th European Conference on Artificial Intelligence* (pp. 338–342). Berlin, Germany.

Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., & Villata, S. (2017). Towards artificial argumentation. *AI Magazine*, *38*, 25–36.

Baroni, P., Caminada, M., & Giacomin, M. (2011). An introduction to argumentation semantics. *The Knowledge Engineering Review*, *26*, 365–410.

Baroni, P., & Giacomin, M. (2009). Semantics of abstract argument systems. In *Argumentation in artificial intelligence* (pp. 25–44). Springer.

Baroni, P., Giacomin, M., & Guida, G. (2005). Scc-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, *168*, 162–210.

Bench-Capon, T. J., & Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial intelligence*, *171*, 619–641.

Bentahar, J., Moulin, B., & Bélanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, *33*, 211–259.

Birnbaum, L., Flowers, M., & McGuire, R. (1980). Towards an ai model of argumentation. In *Proceedings of the First AAAI Conference on Artificial Intelligence* AAAI'80 (pp. 313–315). AAAI Press.

Black, E., & Hunter, A. (2009). An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems*, *19*, 173–209.

Bochman, A. (2007). Non-monotonic reasoning and belief change. In D. Gabbay, & J. Woods (Eds.), *Handbook of the History of Logic, Volume 8: The Many-Valued and Nonmonotonic Turn in Logic* (pp. 557–632). Amsterdam: Elsevier Science Publishers.

Bonzon, E., Delobelle, J., Konieczny, S., & Maudet, N. (2016). A comparative study of ranking-based semantics for abstract argumentation. In *AAAI* (pp. 914–920).

Brewka, G. (1991). *Nonmonotonic reasoning: logical foundations of commonsense* volume 12. Cambridge University Press.

Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning: an overview* volume 73. CSLI publications Stanford.

Bryant, D., & Krause, P. (2008). A review of current defeasible reasoning implementations. *The Knowledge Engineering Review*, *23*, 227–260.

Cain, B. (2007). *A review of the mental workload literature*. Technical Report Defence research and development Toronto (Canada).

Caminada, M. (2007). Comparing two unique extension semantics for formal argumentation: ideal and eager. In *Proceedings of the 19th Belgian-Dutch conference on artificial intelligence (BNAIC 2007)* (pp. 81–87). Utrecht University Press.

Caminada, M. W., Carnielli, W. A., & Dunne, P. E. (2012). Semi-stable semantics. *Journal of Logic and Computation*, *22*, 1207–1254.

Castro, J. L., Trillas, E., & Zurita, J. M. (1998). Non-monotonic fuzzy reasoning. *Fuzzy Sets and Systems*, *94*, 217–225.

Chang, C. F., Miller, A., & Ghose, A. (2009). Mixed-initiative argumentation: Group decision support in medicine. In *eHealth* (pp. 43–50). Springer.

Chesñevar, C. I., Maguitman, A. G., & Loui, R. P. (2000). Logical models of argument. *ACM Computing Surveys (CSUR)*, *32*, 337–383.

Cordón, O. (2011). A historical review of evolutionary learning methods for mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems. *International Journal of Approximate Reasoning*, *52*, 894–913.

Cordón, O., Gomide, F., Herrera, F., Hoffmann, F., & Magdalena, L. (2004). Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy sets and systems*, *141*, 5–31.

Coste-Marquis, S., Konieczny, S., Marquis, P., & Ouali, M. A. (2012). Selecting extensions in weighted argumentation frameworks. In *COMMA* (pp. 342–349).

Craven, R., Toni, F., Cadar, C., Hadad, A., & Williams, M. (2012). Efficient argumentation for medical decision-making. In *KR* (pp. 598–602).

Czelakowski, J. (1985). Algebraic aspects of deduction theorems. *Studia Logica*, *44*, 369–387.

Delladio, T., Rotstein, N. D., & Simari, G. R. (2006). A comparison between non-monotonic formalisms. In *VIII Workshop de Investigadores en Ciencias de la Computación*.

Dondio, P. (2017). Propagating degrees of truth on an argumentation framework: an abstract account of fuzzy argumentation. In *Proceedings of the Symposium on Applied Computing* (pp. 995–1002). ACM.

Dondio, P. (2018). Ranking semantics based on subgraphs analysis. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* AAMAS '18 (pp. 1132–1140). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Dondio, P., & Longo, L. (2011). Trust-based techniques for collective intelligence in social search systems. In *Next generation data technologies for collective computational intelligence* (pp. 113–135). Springer.

Dondio, P., & Longo, L. (2014). Computing trust as a form of presumptive reasoning. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on* (pp. 274–281). IEEE volume 2.

Dubois, D., & Prade, H. (1998). Possibility theory: qualitative and quantitative aspects. In *Quantified representation of uncertainty and imprecision* (pp. 169–226). Springer.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, *77*, 321–358.

Dunne, P. E., Hunter, A., McBurney, P., Parsons, S., & Wooldridge, M. (2011). Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, *175*, 457–486.

Durkin, J., & Durkin, J. (1998). *Expert systems: design and development*. Prentice Hall PTR.

Dutilh Novaes, C., & Veluwenkamp, H. (2017). Reasoning biases, non-monotonic logics and belief revision. *Theoria*, *83*, 29–52.

Eggemeier, F. T. (1988). Properties of workload assessment techniques. *Advances in Psychology*, *52*, 41–62.

El-Azhary, E.-S., Edrees, A., & Rafea, A. (2002). Diagnostic expert system using non-monotonic reasoning. *Expert Systems with Applications*, *23*, 137–144.

Fan, X., & Toni, F. (2015). On computing explanations in argumentation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 1496–1502).

Gabbay, D. M. (1985). Theoretical foundations for non-monotonic reasoning in expert systems. In *Logics and models of concurrent systems* (pp. 439–457). Springer.

Gabbay, D. M., & Guenthner, F. (1984). *Handbook of philosophical logic* volume 4. Dordrecht: Kluwer.

Gegov, A., Gobalakrishnan, N., & Sanders, D. (2014). Rule base compression in fuzzy systems by filtration of non-monotonic rules. *Journal of Intelligent & Fuzzy Systems*, *27*, 2029–2043.

Ginsberg, M. L. (1984). Non-monotonic reasoning using dempster's rule. In *AAAI* (pp. 112–119). volume 84.

Glasspool, D., Fox, J., Oettinger, A., & Smith-Spark, J. (2006). Argumentation in decision support for medical care planning for patients and clinicians. In *AAAI Spring Symposium: Argumentation for Consumers of Healthcare* (pp. 58–63).

Grando, M. A., Moss, L., Sleeman, D., & Kinsella, J. (2013). Argumentation-logic for creating and explaining medical hypotheses. *Artificial intelligence in medicine*, *58*, 1–13.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 904–908). Sage publications Sage CA: Los Angeles, CA volume 50.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (pp. 139–183). Elsevier volume 52.

Hatzilygeroudis, I., & Prentzas, J. (2004). Integrating (rules, neural networks) and cases for knowledge representation and reasoning in expert systems. *Expert Systems with Applications*, *27*, 63–75.

Hellendoorn, H., & Thomas, C. (1993). Defuzzification in fuzzy controllers. *Journal of Intelligent & Fuzzy Systems*, *1*, 109–123.

Hlobil, U. (2018). Choosing your nonmonotonic logic: A shopper' guide. In P. Arazim, & T. Lávika (Eds.), *The Logica Yearbook 2017* (pp. 109–123). London: College Publications.

Horty, J. F., Thomason, R. H., & Touretzky, D. S. (1990). A skeptical theory of inheritance in nonmonotonic semantic networks. *Artificial intelligence*, *42*, 311–348.

Hunter, A., & Williams, M. (2010). Argumentation for aggregating clinical evidence. In *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on* (pp. 361–368). IEEE volume 1.

Hvam, L., Mortensen, N. H., & Riis, J. (2008). Knowledge representation and forms of reasoning for expert systems. *Product Customization*, (pp. 197–217).

Ishibuchi, H., & Nakashima, T. (2001). Effect of rule weights in fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems*, *9*, 506–515.

Konieczny, S., Marquis, P., & Vesic, S. (2015). On supported inference and extension selection in abstract argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty* (pp. 49–59). Springer.

Kowalski, R. A., & Sadri, F. (1991). Logic programs with exceptions. *New Generation Computing*, *9*, 387–400.

Kraus, S., Sycara, K., & Evenchik, A. (1998). Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, *104*, 1 – 69.

Liao, S.-H. (2005). Expert system methodologies and applications - a decade review from 1995 to 2004. *Expert systems with applications*, *28*, 93–103.

Lin, F., & Shoham, Y. (1989). Argument systems: A uniform basis for nonmonotonic reasoning. *Proceedings First International Conference on Principles of Knowledge Representation and Reasoning*, (pp. 245–255).

Longo, L. (2012). Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In *User Modeling, Adaptation, and Personalization* (pp. 369–373). Springer.

Longo, L. (2014). *Formalising Human Mental Workload as a Defeasible Computational Concept*. Ph.D. thesis Trinity College Dublin.

Longo, L. (2015). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology*, *34*, 758–786.

Longo, L. (2016). Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. In *Machine Learning for Health Informatics* (pp. 183–208). Springer.

Longo, L. (2017). Subjective usability, mental workload assessments and their impact on objective human performance. In *IFIP Conference on Human-Computer Interaction* (pp. 202–223). Springer.

Longo, L. (2018a). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE*, *13*, 1–36.

Longo, L. (2018b). On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: A case study in a third-level course. In *Proceedings of the 10th International Conference on Computer Supported Education - Volume 2: CSEDU* (pp. 166–178). INSTICC SciTePress.

Longo, L. (2018c). On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: A case study in a third-level course. In *Proceedings of the 10th International Conference on Computer Supported Education, CSEDU 2018, Funchal, Madeira, Portugal, March 15-17, 2018, Volume 2.* (pp. 166–178).

Longo, L., & Dondio, P. (2014). Defeasible reasoning and argument-based systems in medical fields: An informal overview. In *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on* (pp. 376–381). IEEE.

Longo, L., & Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, Volume I* (pp. 345–352).

Longo, L., & Orru, G. (2019). An evaluation of the reliability, validity and sensitivity of three human mental workload measures under different instructional conditions in third-level education. In B. M. McLaren, R. Reilly, S. Zvacek, & J. Uhomoibhi (Eds.), *Computer Supported Education* (pp. 384–413). Cham: Springer International Publishing.

Mamdani, E. H. (1974). Application of fuzzy algorithms for control of simple dynamic plant. In *Proceedings of the institution of electrical engineers* (pp. 1585–1588). IET volume 121.

Mardani, A., Jusoh, A., & Zavadskas, E. K. (2015). Fuzzy multiple criteria decision-making techniques and applications–two decades review from 1994 to 2014. *Expert Systems with Applications*, *42*, 4126–4148.

Martınez, D. C., Garcıa, A. J., & Simari, G. R. (2008). An abstract argumentation framework with varied-strength attacks. In *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR'08)* (pp. 135–144).

Matt, P.-A., Morgem, M., & Toni, F. (2010). Combining statistics and arguments to compute trust. In *9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada* (pp. 209–216). ACM volume 1.

McCarthy, J. (1980). CircumscriptionâĂŤa form of non-monotonic reasoning. *Artificial intelligence*, *13*, 27–39.

Mitra, R. S., & Basu, A. (1997). Knowledge representation in mickey: An expert system for designing microprocessor-based systems. *Systems, Man*

*and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *27*, 467–479.

Modgil, S. (2009). Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, *173*, 901–934.

Moore, R. C. (1985). Semantical considerations on nonmonotonic logic. *Artificial intelligence*, *25*, 75–94.

Morgenstern, L. (1998). Inheritance comes of age: Applying nonmonotonic techniques to problems in industry. *Artificial Intelligence*, *103*, 237–271.

Morgenstern, L., & Singh, M. (1997). An expert system using nonmonotonic techniques for benefits inquiry in the insurance industry. In *IJCAI (1)* (pp. 655–661).

Moustafa, K., Luz, S., & Longo, L. (2017). Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques. In L. Longo, & M. C. Leva (Eds.), *Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers* (pp. 30–50). Springer.

Nauck, D., Klawonn, F., & Kruse, R. (1997). *Foundations of neuro-fuzzy systems*. John Wiley & Sons, Inc.

Nohria, R. (2015). Medical expert system-a comprehensive review. *International Journal of Computer Applications*, *130*, 44–50.

Nute, D., Mann, R. I., & Brewer, B. F. (1990). Controlling expert system recommendations with defeasible logic. *Decision Support Systems*, *6*, 153–164.

O'Donnell, R., & Eggemeier, F. (1986). *Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. KR Boff, L. Kaufman and JP Thomas*. John Wiley and Sons, Inc.

Parsons, S., & Hunter, A. (1998). A review of uncertainty handling formalisms. In *Applications of uncertainty formalisms* (pp. 8–37). Springer.

Parsons, S., McBurney, P., & Sklar, E. (2010). Reasoning about trust using argumentation: A position paper. In *Workshop on Argumentation in Multi-Agent Systems* (pp. 159–170).

Passino, K. M., Yurkovich, S., & Reinfrank, M. (1998). *Fuzzy control* volume 42. Citeseer.

Patkar, V., Hurt, C., Steele, R., Love, S., Purushotham, A., Williams, M., Thomson, R., & Fox, J. (2006). Evidence-based guidelines and decision support services: a discussion and evaluation in triple assessment of suspected breast cancer. *British Journal of Cancer*, *95*, 1490.

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology*, *5*, 1–11.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument and Computation*, *1*, 93–124.

Prakken, H., & Sartor, G. (2002). The role of logic in computational models of legal argument: A critical survey. In A. C. Kakas, & F. Sadri (Eds.), *Computational Logic: Logic Programming and Beyond: Essays in Honour of Robert A. Kowalski Part II* (pp. 342–381). Berlin, Heidelberg: Springer Berlin Heidelberg.

Prakken, H., & Vreeswijk, G. (2001). Logics for defeasible argumentation. In D. M. Gabbay, & F. Guenthner (Eds.), *Handbook of Philosophical Logic* (pp. 219–318). Dordrecht: Springer Netherlands.

Precup, R.-E., & Hellendoorn, H. (2011). A survey on industrial applications of fuzzy control. *Computers in industry*, *62*, 213–226.

Rahwan, I., & Simari, G. R. (2009). *Argumentation in artificial intelligence* volume 47. Springer.

Reiter, R. (1980). A logic for default reasoning. *Artificial intelligence*, *13*, 81–132.

Reiter, R. (1988). Nonmonotonic reasoning. In *Exploring artificial intelligence* (pp. 439–481). Elsevier.

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 L. Iliadis, & I. Maglogiannis (Eds.), *Artificial Intelligence Applications and Innovations* (pp. 215–229). Cham: Springer International Publishing.

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* (pp. 126–140).

Rizzo, L., & Longo, L. (2018). A qualitative investigation of the degree of explainability of defeasible argumentation and non-monotonic fuzzy reasoning. In *26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 138–149).

Rizzo, L., & Longo, L. (2019). Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study. In *2nd Workshop on Advances In Argumentation In Artificial Intelligence* (pp. 11–26).

Rizzo, L., Majnaric, L., Dondio, P., & Longo, L. (2018a). An investigation of argumentation theory for the prediction of survival in elderly using biomarkers. In *Int. Conf. on Artificial Intelligence Applications and Innovations* (pp. 385–397). Springer.

Rizzo, L., Majnaric, L., & Longo, L. (2018b). A comparative study of defeasible argumentation and non-monotonic fuzzy reasoning for elderly survival prediction using biomarkers. In *AI*IA 2018 - Advances in Artificial Intelligence*

- *XVIIth Int. Conference of the Italian Association for Artificial Intelligence* (pp. 197–209).

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*, 61–86.

Siler, W., & Buckley, J. J. (2005). *Fuzzy expert systems and fuzzy reasoning*. John Wiley & Sons.

Singholi, A. K., & Agarwal, D. (2018). Review of expert system and its application in robotics. In *Intelligent Communication, Control and Devices* (pp. 1253–1265). Springer.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., Cowell, R. G. et al. (1993). Bayesian analysis in expert systems. *Statistical science*, *8*, 219–247.

Spielberger, C., Weiner, I., & Craighead, W. (2010). *The corsini encyclopedia of psychology*. Wiley.

Takagi, T., & Sugeno, M. (1993). Fuzzy identification of systems and its applications to modeling and control. In *Readings in Fuzzy Sets for Intelligent Systems* (pp. 387–403). Elsevier.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Tracy, J. P., & Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. *Usability and Information Design,*, (pp. 256–260).

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*, 358–381.

Tsukamoto, Y. (1979). An approach to fuzzy reasoning method. In M. M. Gupta, R. K. Ragade, & R. R. Yager (Eds.), *Advances in Fuzzy Set Theory and Applications* (pp. 407–428). Amsterdam: NorthHolland.

Walton, D. (2013). *Argumentation schemes for presumptive reasoning*. Routledge.

Wickens, C. D. (1991). Processing resources and attention. *Multiple-task performance*, (pp. 3–34).

Yang, K. H., Olson, D., & Kim, J. (2004). Comparison of first order predicate logic, fuzzy logic and non-monotonic logic as knowledge representation methodology. *Expert Systems with Applications*, *27*, 501–519.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*, 1–17.

Zadeh, L. A. et al. (1965). Fuzzy sets. *Information and control*, *8*, 338–353.

Zeng, Z., Miao, C., Leung, C., & Chin, J. J. (2018). Building more explainable artificial intelligence with argumentation. In *Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 8044–8045).

## Appendix A. Knowledge bases

In this appendix three knowledge-bases in the field of human mental workload are described. These knowledge-bases are built upon subjective measures of mental workload measurement. In other words, they rely on the subjective feedback (in this case questionnaires) provided by humans engaging with an underlying task. For each knowledge base the following are defined:

1. *Features*: A set of features (attributes) believed to influence mental workload and its assessment (with the aid of an expert);
2. *Questions*: A set of questions for quantitatively quantifying the above features;
3. *Mapping*: A map between natural language terms and numerical ranges (for instance "$low = [0, 33]$").
4. *Inferential rules*: A list of inferential IF-THEN rules employing natural language terms of item 3.

5. *Contradictions*: A list of contradictions and exceptions for rules of item 4 in three possible forms:

   - IF Rule A THEN not Rule B.
   - Rule A and Rule B cannot coexist.
   - IF premises THEN not Rule A.

6. *Graphical representation*: A graphical representation of rules and contradictions of items 4 and 5.

At the end of the section a set of fuzzy membership functions is also provided. These can be used to compute the membership grade of natural language terms (defined in 3).

*Appendix A.1. Knowledge base A*

Table A.11: Questions associated to the NASA Task Load Index and employed as features of the knowledge-base A (Hart & Staveland, 1988).

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

Table A.12: Natural language terms and associated numerical ranges employed to reason with features in knowledge base A.

| **Features** | | **MWL** | |
|---|---|---|---|
| **Terms** | **Range** | **Terms** | **Range** |
| Low | [0, 33) | Underload | [0, 33) |
| Medium Lower | [33, 50) | Fitting minus load | [33, 50) |
| Medium Upper | [50, 67) | Fitting plus load | [50, 67) |
| High | [67, 100] | Overload | [67, 100] |

Table A.13: The pairwise comparison procedure of the Nasa Task Load Index instrument (Hart & Staveland, 1988). This comparison is employed for the definition of weights for each feature. The number of times a feature is selected represents its respective weight.

| Pair | feature 1 | | | | feature 2 |
|---|---|---|---|---|---|
| 1 | temporal demand | ☐ | OR | ☐ | frustration |
| 2 | performance | ☐ | OR | ☐ | mental demand |
| 3 | mental demand | ☐ | OR | ☐ | physical demand |
| 4 | frustration | ☐ | OR | ☐ | performance |
| 5 | temporal demand | ☐ | OR | ☐ | effort |
| 6 | physical demand | ☐ | OR | ☐ | frustration |
| 7 | performance | ☐ | OR | ☐ | temporal demand |
| 8 | mental demand | ☐ | OR | ☐ | effort |
| 9 | physical demand | ☐ | OR | ☐ | temporal demand |
| 10 | frustration | ☐ | OR | ☐ | effort |
| 11 | physical demand | ☐ | OR | ☐ | performance |
| 12 | temporal demand | ☐ | OR | ☐ | mental demand |
| 13 | effort | ☐ | OR | ☐ | physical demand |
| 14 | frustration | ☐ | OR | ☐ | mental demand |
| 15 | performance | ☐ | OR | ☐ | effort |

Table A.14: (fuzzy) IF-THEN rules for knowledge base A designed by a domain expert believed to influence mental workload and its assessment.

| **Label** | **Internal structure** |
|---|---|
| MD1 | *low* `mental demand` **THEN** *underload* mwl |
| MD2 | *medium lower* `mental demand` **THEN** *fitting minus load* mwl |
| MD3 | *medium upper* `mental demand` **THEN** *fitting plus load* mwl |
| MD4 | *high* `mental demand` **THEN** *overload* mwl |
| TD1 | *low* `temporal demand` **THEN** *underload* mwl |
| TD2 | *medium lower* `temporal demand` **THEN** *fitting minus load* mwl |
| TD3 | *medium upper* `temporal demand` **THEN** *fitting plus load* mwl |
| TD4 | *high* `temporal demand` **THEN** *overload* mwl |
| EF1 | *low* `effort` **THEN** *underload* mwl |
| EF2 | *medium lower* `effort` **THEN** *fitting minus load* mwl |
| EF3 | *medium upper* `effort` **THEN** *fitting plus load* mwl |
| EF4 | *high* `effort` **THEN** *overload* mwl |
| PF1 | *low* `performance` **THEN** *overload* mwl |
| PF2 | *medium lower* `performance` **THEN** *fitting plus load* mwl |
| PF3 | *medium upper* `performance` **THEN** *fitting minus load* mwl |
| PF4 | *high* `performance` **THEN** *underload* mwl |
| FR1 | *low* `frustration` **THEN** *underload* mwl |
| FR2 | *high* `frustration` **THEN** *overload* mwl |

Table A.15: Contradictions associated to knowledge base A designed by a domain expert believed to influence mental workload and its assessment.

| Label | Internal structure |
|---|---|
| R1 | **IF** *high* `performance` **THEN not** FR2 |
| R2 | **IF** *low* `performance` **THEN not** FR1 |
| C1 | MD1 AND FR2 *cannot coexist* |
| C2 | TD1 AND FR2 *cannot coexist* |
| C3 | FR1 AND MD4 *cannot coexist* |
| C4 | FR1 AND TD4 *cannot coexist* |
| C5 | FR1 AND EF4 *cannot coexist* |
| C6 | EF1 AND FR2 *cannot coexist* |
| C7 | EF1 AND MD4 *cannot coexist* |
| R3 | **IF** EF4 **THEN not** MD1 |

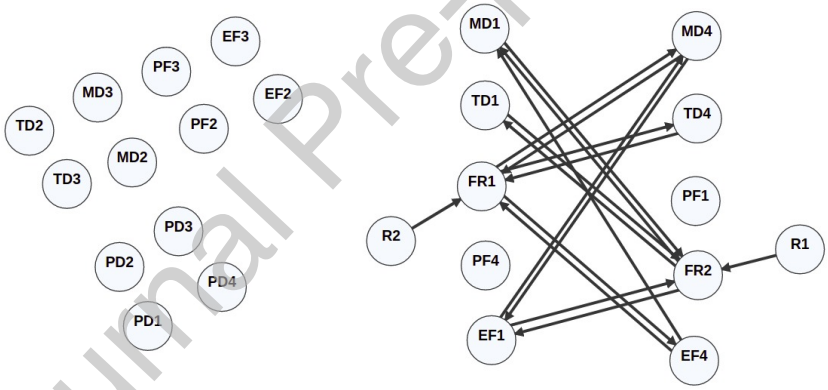



Figure A.23: Graphical representation of knowledge base A. Nodes can represent (fuzzy) IF-THEN rules or premises of contradictions. Arrows represent contradictions between two rules.

*Appendix A.2. Knowledge base B*

Features employed in this knowledge base are the same ones listed in Table A.17. Natural language terms and associated numerical ranges are the same ones listed in Table A.12. The remaining information for modelling and assessing mental workload by this knowledge base are described in the following tables and figures.

Table A.16: (fuzzy) IF-THEN rules for knowledge base B designed by domain expert for inference of mental workload. The same principle of `mental demand` applies to the attributes `temporal demand` (TD), `physical demand` (PD), `solving and deciding` (SD), `selection of response` (SR), `task and space` (TS), `verbal material` (VM), `visual resources` (VR), `auditory resources` (AR), `manual response` (MR), `speech response` (SPR), `effort` (EF), `parallelism` (PR), and `context bias` (CB), forming 52 other rules.

| **Label** | **Internal structure** |
|---|---|
| MD1 | **IF** *low* `mental demand` **THEN** *Underload* |
| MD2 | **IF** *medium lower* `mental demand` **THEN** *Fitting minus* |
| MD3 | **IF** *medium upper* `mental demand` **THEN** *Fitting plus* |
| MD4 | **IF** *high* `mental demand` **THEN** *Overload* |
| PS1 | **IF** *low* `frustration` **THEN** *Underload* |
| PS2 | **IF** *high* `frustration` **THEN** *Overload* |
| MV1 | **IF** *low* `motivation` **THEN** *Underload* |
| PK1 | **IF** *low* `past knowledge` **THEN** *Overload* |
| PK2 | **IF** *high* `past knowledge` **THEN** *Underload* |
| SK1 | **IF** *low* `skills` **THEN** *Overload* |
| SK2 | **IF** *high* `skills` **THEN** *Underload* |
| PF1 | **IF** *low* `performance` **THEN** *Overload* |
| PF2 | **IF** *medium lower* perf. **THEN** *Fitting minus* |
| PF3 | **IF** *medium upper* perf. **THEN** *Fitting plus* |
| PF4 | **IF** *high* `performance` **THEN** *Underload* |

Table A.17: Features and respective questions for their measurement employed in knowledge base B. They were originally proposed in (Longo, 2014).

| **Feature** | **Question** |
|---|---|
| Mental demand | How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy (low mental demand) or complex (high mental demand)? |
| Temporal demand | How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely (low temporal demand) or rapid and frantic (high temporal demand)? |
| Effort | How much conscious mental effort or concentration was required? Was the task almost automatic (low effort) or it required total attention (high effort)? |
| Performance | How successful do you think you were in accomplishing the goal of the task? How satisfied were you with your performance in accomplishing the goal? |
| Frustration | How secure, gratified, content, relaxed and complacent (low psychological stress) versus insecure, discouraged, irritated, stressed and annoyed (high psychological stress) did you feel during the task? |
| Solving and deciding | How much attention was required for activities like remembering, problem-solving, decision-making and perceiving (eg. detecting, recognizing and identifying objects)? |
| Selection of response | How much attention was required for selecting the proper response channel and its execution? (manual - keyboard/mouse, or speech - voice) |
| Task and space | How much attention was required for spatial processing (spatially pay attention around you)? |
| Verbal material | How much attention was required for verbal material (eg. reading or processing linguistic material or listening to verbal conversations)? |
| Visual resources | How much attention was required for executing the task based on the information visually received (through eyes)? |
| Auditory resources | How much attention was required for executing the task based on the information auditorily received (ears)? |
| Manual Response | How much attention was required for manually respond to the task (eg. keyboard/mouse usage)? |
| Speech response | How much attention was required for producing the speech response(eg. engaging in a conversation or talk or answering questions)? |
| Context bias | How often interruptions on the task occurred? Were distractions (mobile, questions, noise, etc.) not important (low context bias) or did they influence your task (high context bias)? |
| Past knowledge | How much experience do you have in performing the task or similar tasks on the same website? |
| Skill | Did your skills have no influence (low) or did they help to execute the task (high)? |
| Motivation | Were you motivated to complete the task? |
| Parallelism | Did you perform just this task (low parallelism) or were you doing other parallel tasks (high parallelism) (eg. multiple tabs/windows/programs)? |
| Arousal | Were you aroused during the task? Were you sleepy, tired (low arousal) or fully awake and activated (high arousal)? |
| Task difficult | $\frac{1}{8}$((solving/deciding) + (auditory resources) + (manual response) + (speech response ) + (selection of response) + (task/space) + (verbal material ) + (visual resources)) |
| Physical demand | How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious? |

Table A.18: Contradictions for knowledge base B designed by domain expert for inference of mental workload.

| Label | Internal structure |
|---|---|
| AD1a | **IF** *low* `arousal` AND *low* `task difficulty` **THEN not** PF4 |
| AD1b | **IF** *low* `arousal` AND *low* `task difficulty` **THEN not** PF3 |
| AD1c | **IF** *low* `arousal` AND *low* `task difficulty` **THEN not** PF2 |
| AD2a | **IF** *low* `arousal` AND *high* `task difficulty` **THEN not** PF4 |
| AD2b | **IF** *low* `arousal` AND *high* `task difficulty` **THEN not** PF3 |
| AD2c | **IF** *low* `arousal` AND *high* `task difficulty` **THEN not** PF2 |
| AD3a | **IF** *medium lower* `arousal` AND *low* `task difficulty` **THEN not** PF1 |
| AD3b | **IF** *medium lower* `arousal` AND *low* `task difficulty` **THEN not** PF4 |
| AD4a | **IF** *medium lower* `arousal` AND *high* `task difficulty` **THEN not** PF1 |
| AD4b | **IF** *medium lower* `arousal` AND *high* `task difficulty` **THEN not** PF3 |
| AD4c | **IF** *medium lower* `arousal` AND *high* `task difficulty` **THEN not** PF4 |
| AD4d | **IF** *medium upper* `arousal` AND *high* `task difficulty` **THEN not** PF1 |
| AD4e | **IF** *medium upper* `arousal` AND *high* `task difficulty` **THEN not** PF3 |
| AD4f | **IF** *medium upper* `arousal` AND *high* `task difficulty` **THEN not** PF4 |
| AD5a | **IF** *medium upper* `arousal` AND *low* `task difficulty` **THEN not** PF1 |
| AD5b | **IF** *medium upper* `arousal` AND *low* `task difficulty` **THEN not** PF2 |
| AD5c | **IF** *medium upper* `arousal` AND *low* `task difficulty` **THEN not** PF3 |
| AD5d | **IF** *high* `arousal` AND *low* `task difficulty` **THEN not** PF1 |
| AD5e | **IF** *high* `arousal` AND *low* `task difficulty` **THEN not** PF2 |
| AD5f | **IF** *high* `arousal` AND *low* `task difficulty` **THEN not** PF3 |
| AD6a | **IF** *high* `arousal` AND *high* `task difficulty` **THEN not** PF2 |
| AD6b | **IF** *high* `arousal` AND *high* `task difficulty` **THEN not** PF3 |
| AD6c | **IF** *high* `arousal` AND `task difficulty` **THEN not** PF4 |
| MV2 | **IF** *low* `motivation` **THEN not** EF3 |
| MV3 | **IF** *low* `motivation` **THEN not** EF4 |
| MV4 | **IF** *high* `motivation` **THEN not** EF1 |
| MV5 | **IF** *high* `motivation` **THEN not** EF2 |
| DS1 | **IF** *high* `task difficulty` AND *high* `skills` **THEN not** EF4 |
| DS2 | **IF** *high* `task difficulty` AND *high* `skills` AND *low* `effort` **THEN not** PF1 |
| DS3 | **IF** *high* `task difficulty` AND *high* `skills` AND *medium lower* `effort` **THEN not** PF1 |
| DS4 | **IF** *high* `task difficulty` AND *high* `skills` AND *medium upper* `effort` **THEN not** PF1 |
| R1 | MD1 AND SD4 *cannot coexist* |
| R2 | MD4 AND SD1 *cannot coexist* |
| R3 | PK1 AND SK2 *cannot coexist* |
| R4 | PK2 AND SK1 *cannot coexist* |
| R5 | PK1 AND EF1 *cannot coexist* |
| R6 | PK2 AND EF4 *cannot coexist* |
| R7 | SK1 AND EF1 *cannot coexist* |
| R8 | SK2 AND EF4 *cannot coexist* |
| R9 | CB4 AND PS1 *cannot coexist* |

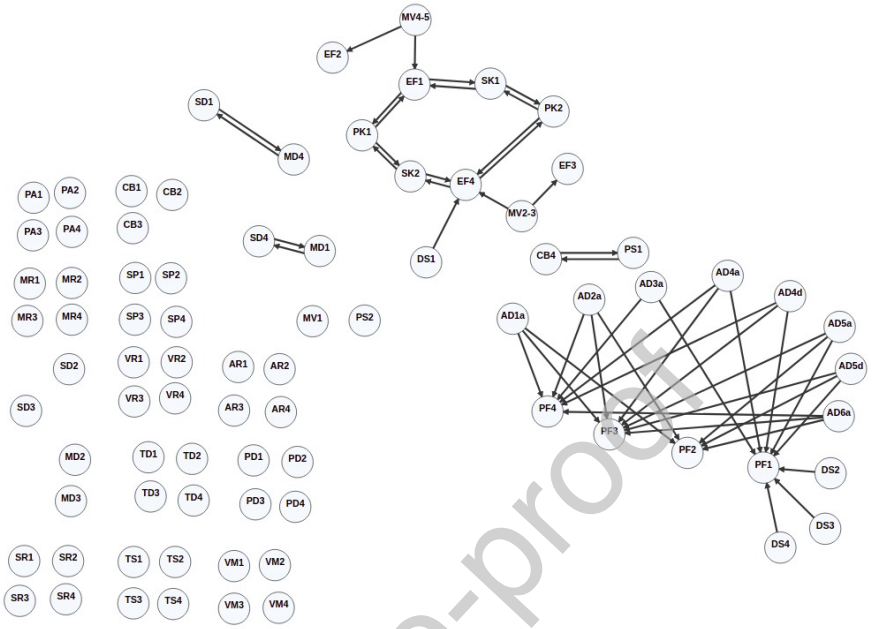



Figure A.24: Graphical representation of knowledge base B. Nodes can represent (fuzzy) IF-THEN rules or premises of contradictions. Arrows represent contradictions between two rules.

*Appendix A.3. Knowledge base C*

This knowledge base is a mix of knowledge bases A and B. The elements required by it are defined as following:

- The features employed are listed in Table A.17.
- Natural language terms and associated numerical ranges are listed in Table A.12.
- IF-THEN rules are listed in Table A.16.
- Contradictions are from both Tables A.18 and A.15.
- The graphical representation of the knowledge base is depicted in Fig. A.25.

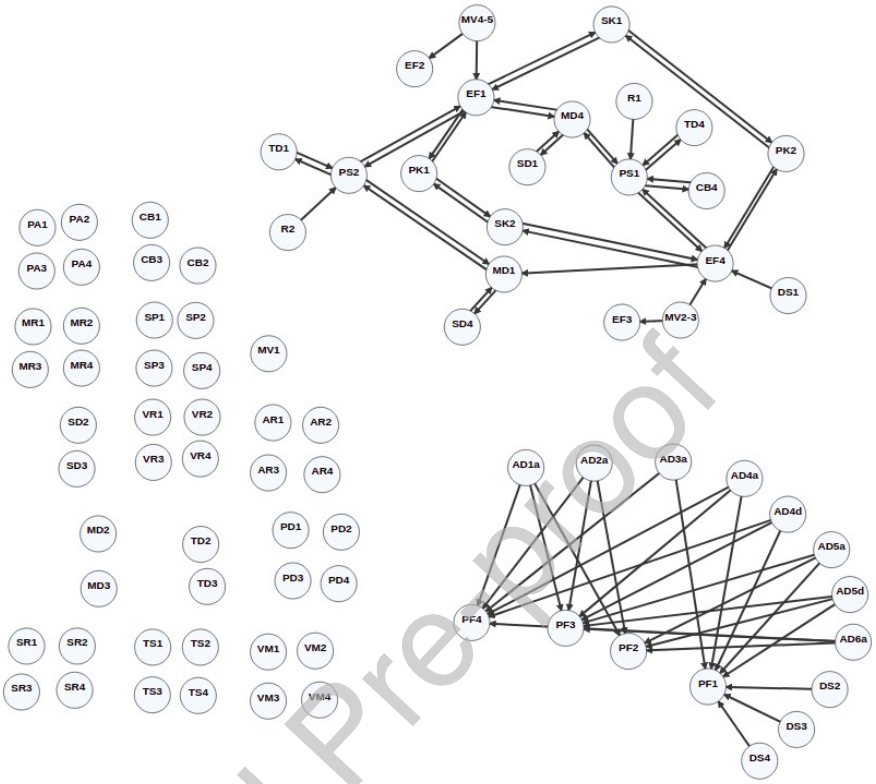
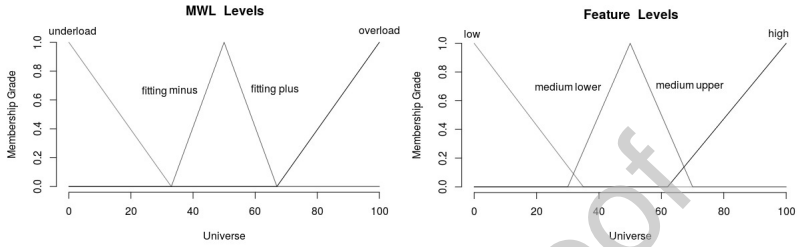



Figure A.25: Graphical representation of knowledge base C. Nodes can represent (fuzzy) IF-THEN rules or premises of contradictions. Arrows represent contradictions between two rules.

*Appendix A.4. Fuzzy membership functions*

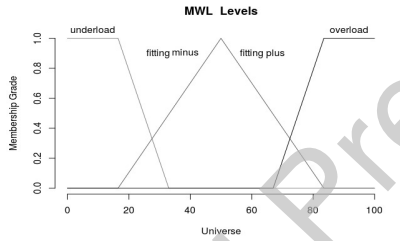
Fig. A.26 depicts the possible fuzzy membership functions employed for modelling the natural language terms listed in Table A.12



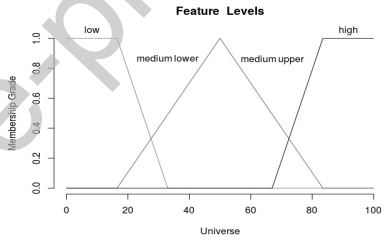

(a) Triangular MWL levels



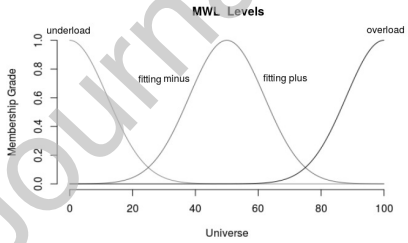

(b) Triangular feature levels



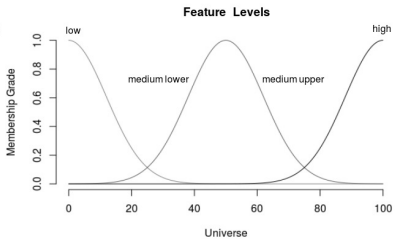

(c) Trapezoid and triangular MWL levels




(d) Trapezoid and triangular feature levels




(e) Gaussian MWL levels




(f) Gaussian feature levels

Figure A.26: Employed fuzzy membership functions for different MWL and feature levels.

## Appendix B. List of information seeking web-based tasks

Table B.19: List of experimental web-based tasks employed for measurement of imposed mental workload. Each website had two interfaces: the original one and one slightly modified, generating two tasks for each description. These tasks were first designed and employed in (Longo, 2014).

| Task | Description | Task condition | Web-site |
|---|---|---|---|
| $T_{1.1}, T_{1.2}$ | Find out how many people live in Sidney | Simple search | Wikipedia |
| $T_{2.1}, T_{2.2}$ | Read `simple.wikipedia.org/wiki/Grammar` | No goals, no time pressure | Wikipedia |
| $T_{3.1}, T_{3.2}$ | Find out the difference (in years) between the year of the foundation of the Apple Computer Inc. and the year of the $14^{th}$ FIFA world cup | Dual-task and mental arithmetical calculations | Google |
| $T_{4.1}, T_{4.2}$ | Find out the difference (in years) between the foundation of the Microsoft Corp. & the year of the $23^{rd}$ Olympic games | Dual-task and mental arithmetical calculations | Google |
| $T_{5.1}, T_{5.2}$ | Find out the year of birth of the $1^{st}$ wife of the founder of playboy | Single task + time pressure (2-min limit). Each 30 secs user is warned of time left | Google |
| $T_{6.1}, T_{6.2}$ | Find out the name of the man (interpreted by Johnny Deep) in the video `www.youtube.com/watch?v=FfTPS-TFQ_c` | Constant demand on visual and auditory modalities. Participant can replay the video if required | Youtube |
| $T_{7.1}.T_{7.2}$ | a) Play the song `www.youtube.com/watch?v=Rb5G1eRIj6c`. While listening to it, b) find out the result of the polynomial equation $p(x)$, with $x = 7$ contained in the wikipedia article `http://it.wikipedia.org/wiki/Polinomi` | Demand on visual modality and inference on auditory modality. The song is extremely irritating | Wikipedia |
| $T_{8.1}, T_{8.2}$ | Find out how many times Stewie jumps in the video `www.youtube.com/watch?v=TSe9gbdkQ8s` | Demand on visual resource + external interference: user is distracted twice & can replay video | Youtube |
| $T_{9.1}, T_{9.2}$ | Find out the age of the blue fish in the video `www.youtube.com/watch?v=H4BNbHBcnDI` | Demand on visual and auditory modality, plus time-pressure: 150-sec limit. User can replay the video. There is no answer. | Youtube |

## Appendix C. List of models built using each reasoning approach

Table C.20: Designed argument-based models and their parameters across each layer.

| Model | Exp. (Table 3) | Layer 1 Arguments | Layer 2 Conflicts | Layer 3 Attack relation | Layer 4 Semantics | Layer 5 Accrual |
|---|---|---|---|---|---|---|
| A1 | $E_a$ | KB1 (Appendix A.1) | | Binary | Grounded | average |
| A2 | $E_a$ | KB1 (Appendix A.1) | | Strength of arg. | Grounded | w. average |
| A3 | $E_a$ | KB1 (Appendix A.1) | | Binary | Preferred | card. + average |
| A4 | $E_a$ | KB1 (Appendix A.1) | | Strength of arg. | Preferred | card. + w. average |
| A5 | $E_b$ | KB2 (Appendix A.2) | | Binary | Grounded | average |
| A6 | $E_b$ | KB2 (Appendix A.2) | | Binary | Preferred | card. + average |
| A7 | $E_c$ | KB3 (Appendix A.3) | | Binary | Grounded | average |
| A8 | $E_c$ | KB3 (Appendix A.3) | | Binary | Preferred | card. + average |

Table C.21: Designed expert system models and their parameters.

| Model | Knowledge-base (App. A) | Heuristic (p. 23) | Experiment (Table 3) |
|---|---|---|---|
| E1 | KB1 | $h_1$ | $E_a$ |
| E2 | KB1 | $h_2$ | $E_a$ |
| E3 | KB1 | $h_3$ | $E_a$ |
| E4 | KB1 | $h_4$ | $E_a$ |
| E5 | KB2 | $h_1$ | $E_b$ |
| E6 | KB2 | $h_3$ | $E_b$ |
| E7 | KB3 | $h_1$ | $E_c$ |
| E8 | KB3 | $h_3$ | $E_c$ |

Table C.22: Designed fuzzy reasoning models and their parameters.

| Model | Operators | Defuzzification method | Rule weight | KB (App. A) | FMF (App. A.4) | Experiment (Table 3) |
|---|---|---|---|---|---|---|
| $FL1$ | Zadeh | Centroid | no | KB1 | Triangular | $E_a$ |
| $FL2$ | Zadeh | Mean of max | no | KB1 | Triangular | $E_a$ |
| $FL3$ | Product | Centroid | no | KB1 | Triangular | $E_a$ |
| $FL4$ | Product | Mean of max | no | KB1 | Triangular | $E_a$ |
| $FL5$ | Łukasiewicz | Centroid | no | KB1 | Triangular | $E_a$ |
| $FL6$ | Łukasiewicz | Mean of max | no | KB1 | Triangular | $E_a$ |
| $FL7$ | Zadeh | Centroid | yes | KB1 | Triangular | $E_a$ |
| $FL8$ | Zadeh | Mean of max | yes | KB1 | Triangular | $E_a$ |
| $FL9$ | Product | Centroid | yes | KB1 | Triangular | $E_a$ |
| $FL10$ | Product | Mean of max | yes | KB1 | Triangular | $E_a$ |
| $FL11$ | Łukasiewicz | Centroid | yes | KB1 | Triangular | $E_a$ |
| $FL12$ | Łukasiewicz | Mean of max | yes | KB1 | Triangular | $E_a$ |
| $FL13$ | Zadeh | Centroid | no | KB2 | Trapezoid | $E_b$ |
| $FL14$ | Zadeh | Mean of max | no | KB2 | Trapezoid | $E_b$ |
| $FL15$ | Product | Centroid | no | KB2 | Trapezoid | $E_b$ |
| $FL16$ | Product | Mean of max | no | KB2 | Trapezoid | $E_b$ |
| $FL17$ | Łukasiewicz | Centroid | no | KB2 | Trapezoid | $E_b$ |
| $FL18$ | Łukasiewicz | Mean of max | no | KB2 | Trapezoid | $E_b$ |
| $FL19$ | Zadeh | Centroid | no | KB3 | Trapezoid | $E_c$ |
| $FL20$ | Zadeh | Mean of max | no | KB3 | Trapezoid | $E_c$ |
| $FL21$ | Product | Centroid | no | KB3 | Trapezoid | $E_c$ |
| $FL22$ | Product | Mean of max | no | KB3 | Trapezoid | $E_c$ |
| $FL23$ | Łukasiewicz | Centroid | no | KB3 | Trapezoid | $E_c$ |
| $FL24$ | Łukasiewicz | Mean of max | no | KB3 | Trapezoid | $E_c$ |
| $FC1$ | Zadeh | Centroid | no | KB1 | Gaussian | $E_a$ |
| $FC2$ | Zadeh | Mean of max | no | KB1 | Gaussian | $E_a$ |
| $FC3$ | Product | Centroid | no | KB1 | Gaussian | $E_a$ |
| $FC4$ | Product | Mean of max | no | KB1 | Gaussian | $E_a$ |
| $FC5$ | Łukasiewicz | Centroid | no | KB1 | Gaussian | $E_a$ |
| $FC6$ | Łukasiewicz | Mean of max | no | KB1 | Gaussian | $E_a$ |
| $FC7$ | Zadeh | Centroid | yes | KB1 | Gaussian | $E_a$ |
| $FC8$ | Zadeh | Mean of max | yes | KB1 | Gaussian | $E_a$ |
| $FC9$ | Product | Centroid | yes | KB1 | Gaussian | $E_a$ |
| $FC10$ | Product | Mean of max | yes | KB1 | Gaussian | $E_a$ |
| $FC11$ | Łukasiewicz | Centroid | yes | KB1 | Gaussian | $E_a$ |
| $FC12$ | Łukasiewicz | Mean of max | yes | KB1 | Gaussian | $E_a$ |
| $FC13$ | Zadeh | Centroid | no | KB2 | Gaussian | $E_b$ |
| $FC14$ | Zadeh | Mean of max | no | KB2 | Gaussian | $E_b$ |
| $FC15$ | Product | Centroid | no | KB2 | Gaussian | $E_b$ |
| $FC16$ | Product | Mean of max | no | KB2 | Gaussian | $E_b$ |
| $FC17$ | Łukasiewicz | Centroid | no | KB2 | Gaussian | $E_b$ |
| $FC18$ | Łukasiewicz | Mean of max | no | KB2 | Gaussian | $E_b$ |
| $FC19$ | Zadeh | Centroid | no | KB3 | Gaussian | $E_c$ |
| $FC20$ | Zadeh | Mean of max | no | KB3 | Gaussian | $E_c$ |
| $FC21$ | Product | Centroid | no | KB3 | Gaussian | $E_c$ |
| $FC22$ | Product | Mean of max | no | KB3 | Gaussian | $E_c$ |
| $FC23$ | Łukasiewicz | Centroid | no | KB3 | Gaussian | $E_c$ |
| $FC24$ | Łukasiewicz | Mean of max | no | KB3 | Gaussian | $E_c$ |

## Appendix D. Density plots

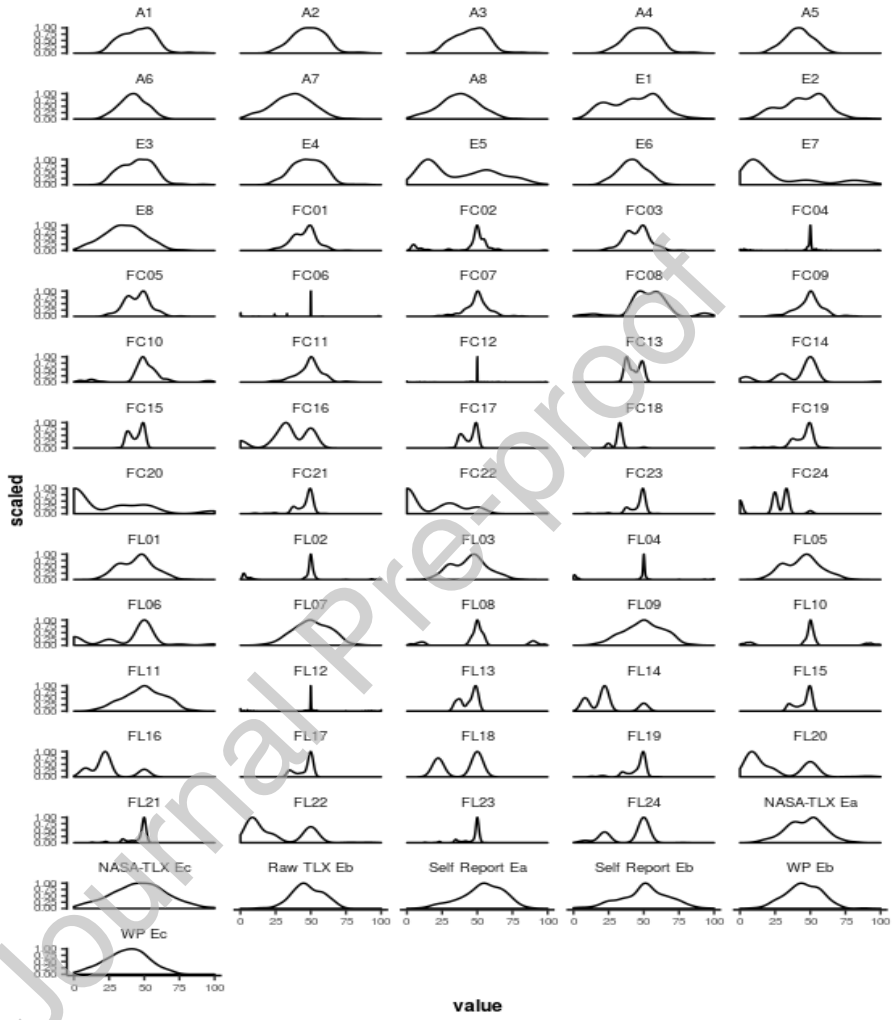



Figure D.27: Density plots of inferred MWL scalars by all designed models and baseline instruments. A{1-8} are argument-based models. FC{01-24} are fuzzy reasoning models of Gaussian fuzzy membership functions. FL{01-24} are fuzzy reasoning models of linear fuzzy membership functions. E{1-8} are expert system models. Other graphs are the result of baseline models (NASA-TLX, Raw TLX, WP and Self Report) in the different experiments ($E_a$, $E_b$ and $E_c$).

**Author statement**

Lucas Rizzo: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Visualization, Data curation. Luca Longo: Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision, Data curation

**Declaration of interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**