Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study

Lucas Rizzo and Luca Longo*

The ADAPT global centre of excellence for digital content and media innovation
School of Computing, Dublin Institute of Technology, Dublin, Ireland
lucas.rizzo@mydit.ie,luca.longo@dit.ie*

Abstract. Inferences through knowledge driven approaches have been researched extensively in the field of Artificial Intelligence. Among such approaches argumentation theory has recently shown appealing properties for inference under uncertainty and conflicting evidence. Nonetheless, there is a lack of studies which examine its inferential capacity over other quantitative theories of reasoning under uncertainty with real-world knowledge-bases. This study is focused on a comparison between argumentation theory and non-monotonic fuzzy reasoning when applied to modeling the construct of human mental workload (MWL). Different argument-based and non-monotonic fuzzy reasoning models, aimed at inferring the MWL imposed by a selection of learning tasks, in a third-level context, have been designed. These models are built upon knowledge-bases that contain uncertain and conflicting evidence provided by human experts. An analysis of the convergent and face validity of such models has been performed. Results suggest a superior inferential capacity of argument-based models over fuzzy reasoning-based models.

Keywords: Argumentation Theory, Non-monotonic Reasoning, Fuzzy Logics, Mental workload, Defeasible Reasoning

1 Introduction

Uncertainty is inevitable in many real-world domains. Several theories in the field of Artificial Intelligence (AI) have been studied for dealing with quantitative reasoning under uncertainty, such as Probability calculus and its variations: Possibility Theory and Imprecise Probabilities, Dempster-Shafer Theory, Argumentation Theory and Multi-valued Logics like Fuzzy Logics. More specifically, Fuzzy Reasoning [29] and computational Argumentation Theory (AT) [2] have been extensively used in practical domains such as medicine, pharmaceutical industry and engineering [14, 12, 18]. On one hand, AT allows the construction of computational models for the implementation of defeasible reasoning, or reasoning when a conclusion can be withdrawn in the light of new evidence. On the other hand, Fuzzy Reasoning allows the creation of models that can include

a robust representation of linguistic information and can produce rational inferences when this is incomplete, inconsistent or ambiguous. While some works have proposed fuzzy argumentation frameworks [6], building upon the two fields, there is a lack of research devoted to the analysis of the inferential capacity of AT in the context of quantitative reasoning under uncertainty. Thus, an empirical investigation is proposed here whereby the inferential capacity of AT and non-monotonic fuzzy reasoning is compared. To achieve this goal, three knowledge bases, built with the aid of an expert in the field of Mental Workload (MWL), are considered. In this study, the inferential capacity is quantified in terms of the validity of the mental workload indexes produced by constructed models. The specific research question under investigation is: *to what extent can defeasible reasoning, implemented via argumentation theory, allow the construction of models with a superior inferential capacity when compared to models implemented with non-monotonic fuzzy reasoning?*

The rest of the paper continues with section 2 presenting related work on fuzzy reasoning, AT and with a short description of the construct of MWL. The design of a comparative experiment and methodologies for the development of argument-based and fuzzy-reasoning based models are detailed in section 3. Section 4 introduces the results followed by a discussion. Section 5 concludes the research by highlighting its contribution and proposing future work.

2 Related work

Reasoning and explanation under incomplete and uncertain knowledge have been investigated for several decades in AI. On one hand, classical propositional logic has demonstrated to be inadequate, due to its monotonicity property, for dealing with real-world argumentative activities often involving inconsistent and conflicting information [24]. On the other hand, defeasible reasoning has emerged as a good alternative for non-monotonic activities [5, 15]. In monotonic reasoning, the knowledge base may only grow with new facts in a monotonic fashion and a previous conclusion cannot be retracted. Instead, reasoning is non-monotonic when a conclusion can be retracted in the light of new evidence. It relies on the idea that a claim can be derived from premises partially specified, but in the case of an exception arising the conclusion can be withdrawn [14].

2.1 Non-monotonic fuzzy reasoning

Fuzzy reasoning, as proposed by [29], is built upon the concept of membership functions. This is a particular function that assigns to each proposition or linguistic term a grade of membership in the range $[0, 1] \in \mathbb{R}$. Fuzzy sets are formed by fuzzy propositions and have similar notions to classical set theory such as inclusion, union and intersection. A fuzzy control 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 [22]. This module is responsible for returning the fuzzy information into the original domain of the problem and

producing a final inference. Some works have suggested different extensions of such systems that incorporate a non-monotonic layer for reasoning under uncertainty and with conflicting information. Unfortunately, these are sporadic and not backed up by empirical research. For example, in [4], conflicting rules have their conclusions aggregated by an averaging function; in [11] a rule base compression method is proposed for the reduction of non-monotonic rules; and in [27], a third approach can be found. Here, as proposed in [27], Possibility Theory [8] is included into the fuzzy reasoning system to handle conflicting instructions. In Possibility Theory, differently from traditional fuzzy systems, truth values can be represented by *possibility* and *necessity*. The first indicates the extent to which data fail to refute its truth while the second indicates the extent to which data supports its truth. Both are values between $[0, 1] \in \mathbb{R}$. This theory is employed in this study for the development of a non-monotonic fuzzy reasoning system, being detailed in section 3, useful for comparison purposes.

2.2 Argumentation theory

Classical argumentation, from its roots within philosophy and psychology, deals with the study of how arguments or assertions are defined, discussed and solved in case of divergent opinions. In AI, argumentation refers to that body of literature that focuses on techniques for constructing computational models of arguments. Such models have become increasingly important for operationalising non-monotonic reasoning [5, 1]. Example of application areas include dialogue and negotiation [1], knowledge representation [25] and decision-making in healthcare [12, 17, 16]. Argumentation systems are usually formed by several parts. These can range from the definition of the internal structure of arguments and the resolution of the conflicts between arguments to possible resolution strategies for reaching justifiable conclusions. A good summary of these components and their role was presented in [14] and depicted in figure 1. Such structure has been already adopted in previous studies [26] and has helped with the internal organisation of novel argument-based systems. Unfortunately, one of the main issues surrounding argumentation theory is the lack of studies devoted to the examination of its impact on the quality of the inferences produced by reasoning models built upon it. This research is an attempt to investigate this issue and the aforementioned multi-layer structure (figure 1) is adopted.

2.3 Mental workload

Mental workload (MWL) can be intuitively described as the amount of cognitive activity exerted to accomplish a specific task under a finite period of time [3]. There are different classes of methods that have been proposed for measuring MWL [10]: self-reporting, primary task performance and physiological methods. In this work the class of self-reporting measures is adopted. This class relies on the analysis of the subjective feedback provided by humans interacting with an underlying system or on a certain cognitive task. Among well known methods, the NASA-Task Load Index (NASA-TLX) has been largely employed in the last

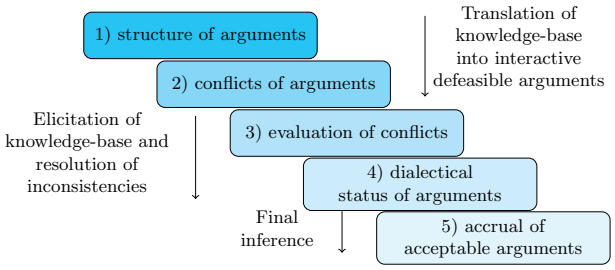

Fig. 1: Five layers upon which argumentation systems are generally built [14].

few decades [13]. It is a combination of six factors believed to influence mental workload: mental, temporal and physical demand, stress, effort and performance. Each factor is quantified with a subjective judgement coupled with a weight w computed via a paired comparison procedure. The questionnaire designed for the quantification of each factor can be found in [13]. Eventually, the final MWL score is computed as a weighted average, considering the subjective rating associated to each attribute d_i (for the 6 dimensions) and the correspondent weights w_i: $\text{TLX}_{\text{MWL}} = \left(\sum_{i=1}^{6} d_i \times w_i\right)\frac{1}{15}$. Several criteria have been proposed and widely used in psychology for the validation of measures of mental workload [21] such as: reliability, validity, sensitivity and diagnosticity among others. This paper focuses particularly on *validity* and in details on two forms:
− *face validity* – it determines the extent to which a measure of MWL appears effective in terms of its stated aims (measuring mental workload);
− *convergent validity* – it refers to the extent to which different MWL measures that should be theoretically related, are in fact related [28].

3 Design and methodology

A primary research study was designed and it included a comparison between the inferential capacity of AT and non-monotonic fuzzy reasoning considering three knowledge-bases produced within the MWL domain. It is demonstrated how these knowledge-bases, built by an expert upon the features extracted from the original NASA-TLX mental workload assessment technique (as per section 2.3), can be translated into defeasible argument-based models and into non-monotonic fuzzy reasoning models. Three main parts compose the non-monotonic fuzzy reasoning models: a fuzzification module, an inference engine and a defuzzification module. Argument-based models are defined as in figure 1 (section 2.2). A comparison of the inference produced by fuzzy reasoning and AT was made in terms of their differences in convergent and face validity. Employed data was composed by the answers of the NASA-TLX questionnaire from 213 students who performed different learning tasks in third-level classes. The overall design of the research is summarised in figure 2. Due to space constraints, the full knowledge-bases produced by the expert with the aid of the authors can be seen online[1].

[1] `https://doi.org/10.6084/m9.figshare.6979865.v1`

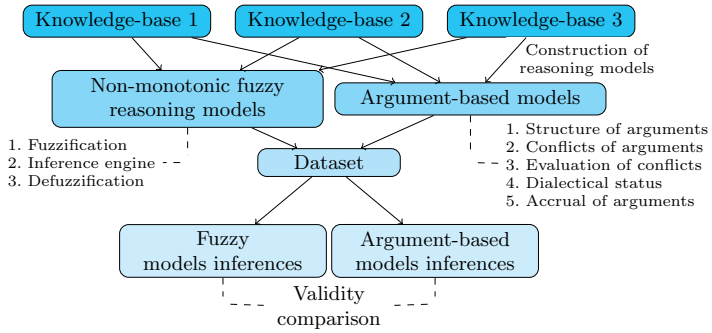

Fig. 2: Evaluation strategy schema

3.1 Non-monotonic fuzzy reasoning models

Fuzzification module The knowledge-bases of the interviewed expert can be represented by rules of the form *"IF ... THEN ..."*. The antecedent (before THEN) is a set of premises associated to a number of workload attributes, while the consequent (after THEN) is associated to a possible MWL level. Examples:

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

Each MWL level (consequent of a rule) was described by a number of FMFs in different ways (figure 3). According to the domain expert's knowledge two options were designed: from [0, 100] having 4 membership functions associated to it and from [0, 20] having 5 membership functions. Fuzzy membership functions (FMF) were also defined for all linguistic variables present in the knowledge-base such as *low* `mental demand` and *low* `effort`. Figure 4 show some examples of FMFs designed following the expert's opinion. Their inputs were normalised according to their possible minimum and maximum values to follow the same universe adopted for the MWL levels.

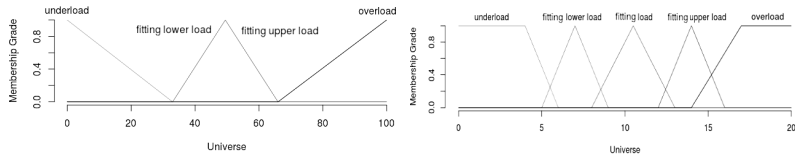

Fig. 3: Example of membership functions for the MWL levels

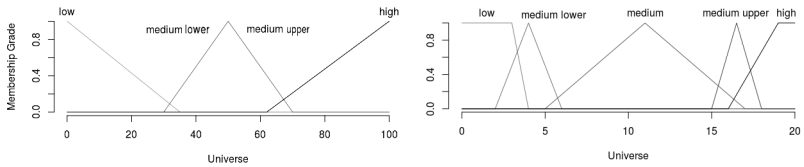

Fig. 4: Example of membership functions for the attribute 'mental demand'

Inference engine Once the knowledge-base of the expert is fully translated into rules within the fuzzification module, fuzzy inferences can be performed. Unfortunately, a high amount of contradicting information is provided by the expert in the knowledge-bases which needs to be firstly solved. For example, the expert expressed the following contradiction in natural language: 'If high effort then mental demand cannot be low'. This information indicates that if `effort` is *high* then any rule whose antecedent contains "*low* `mental demand`" is being refuted and should be re-evaluated in order to change or not its truth value. An example is given by the following rule:

- Exception 1: *high* `effort` *refutes* Rule 1

Exceptions can be tackled by Possibility Theory, as implemented in [27] for fuzzy reasoning with rule based systems. In this case truth values are represented by *possibility* (Pos) and *necessity* (Nec) as defined in Section 3.1. Both are values between [0, 1] $\in \mathbb{R}$. Possibility of a proposition can also be seen as the upper bound of the 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 refutes A is derivable in [27] and given by:

$$Nec(A) = min(Nec(A), \neg Nec(Q_1), \ldots, \neg Nec(Q_n)) \quad (1)$$

Where $\neg Nec(Q) = 1 - Nec(Q)$. In this study, there is no addition of supporting information but only attempts to refute information. Thus, equation (1) can deal with the contradictions in the knowledge-bases. For instance, the truth value of Rule 1, supposing that it is refuted only by Exception 1, is given by:

- Truth value of 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 attribute `mental demand`. For instance, if `mental demand` = 1, then Nec(*low* `mental demand`) = 1, according to the membership function *low* of figure 4. Also, for instance, if Nec(*high* `effort`) = 0 note that Exception 1 has no impact on Rule 1 and if Nec(*high* `effort`) = 1 the new truth value of Rule 1 is 0. Values between 1 and 0 indicates that Rule 1 is partially refuted. The truth value of Rule 1 represents the truth value of `underload` in this particular rule.

It is important to highlight that the theory developed in [27] had in mind a multi-step forward-chaining reasoning system. In this study, the 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 refutations. More exactly, it is possible to define a tree structure in which the consequent of a refutation is the antecedent of the next refutation. In this way, equation (1) 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 that is not the case here. For instance suppose the following IF-THEN rules and their refutations:

- Rule 3: **IF** *low* `temporal demand` **THEN** *underload*

- Rule 4: **IF** *high* `frustration` **THEN** *overload*
- Exception 2: *low* `temporal demand` *refutes* Rule 4
- Exception 3: *high* `frustration` *refutes* Rule 3

In this case it is not clear if exceptions 2 or 3 should be solved first. Given that there is no information on the knowledge-bases to decide whether an attribute (premise of a rule) or an exception is more important, here they are solved simultaneously. Firstly, the truth value of all rules are stored before solving any cyclic exceptions. For instance, the truth values of Rule 3 and 4 are:

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

Having a mechanism to solve conflicts, fuzzy operators can be applied on antecedents of IF-THEN rules and for the aggregation of the consequents (MWL levels) across the rules. Three known operators are selected for investigation: *Zadeh, Product* and *Lukasiewicz*. Table 1 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 Rule 1 and Rule 3 infer *underload* is "Nec(Rule 1) OR Nec(Rule 3)".

Table 1: 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)
Lukasiewicz	max(a + b - 1, 0)	min(a + b, 1)
Product	$a.b$	a + b - $a.b$

Defuzzification module The output of the inference engine is a graphic representation of the aggregation of consequents (MWL levels), as depicted in figure 5 with an example. Several methods can be used for calculating a single defuzzified scalar. Two are selected here: *mean of max* and *centroid*. The first returns the average of all elements (here MWL levels) with maximal membership grade. The second returns the coordinates (x, y) of the center of gravity of the geometric shape formed by the aggregation of the membership functions associated to each MWL level. The defuzzified scalar is represented then by the x coordinate of the centroid. Finally, a set of models is constructed with different fuzzy logic operators and defuzzification techniques, as listed in table 2.

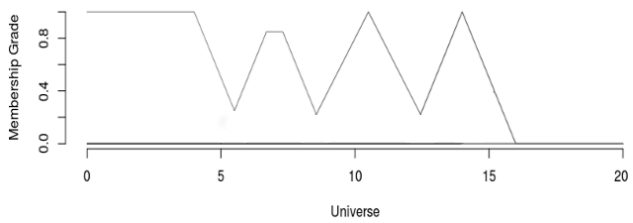

Fig. 5: An example of the defuzzification process whereby an aggregation of 5 membership functions associated to 5 MWL level. The final truth values in this example are: *underload* = 1, *fitting lower* = 0.83, *fitting load* = 1, *fitting upper* = 1 and *overload* = 0. The coordinates of the centroid are (6.89, 0.39) and the mean of max is 7.12.

Table 2: Summary of the designed fuzzy reasoning models

Model	**Operators**	**Defuzzification method**	**Attribute levels**	**Index levels**	**Knowledge-base**[2]
$F1$	Zadeh	Centroid	Figure 3 left	Figure 4 left	1
$F2$	Zadeh	Mean of max	Figure 3 left	Figure 4 left	1
$F3$	Product	Centroid	Figure 3 left	Figure 4 left	1
$F4$	Product	Mean of max	Figure 3 left	Figure 4 left	1
$F5$	Lukasiewicz	Centroid	Figure 3 left	Figure 4 left	1
$F6$	Lukasiewicz	Mean of max	Figure 3 left	Figure 4 left	1
$F7$	Zadeh	Centroid	Figure 3 left	Figure 4 left	2
$F8$	Zadeh	Mean of max	Figure 3 left	Figure 4 left	2
$F9$	Product	Centroid	Figure 3 left	Figure 4 left	2
$F10$	Product	Mean of max	Figure 3 left	Figure 4 left	2
$F11$	Lukasiewicz	Centroid	Figure 3 left	Figure 4 left	2
$F12$	Lukasiewicz	Mean of max	Figure 3 left	Figure 4 left	2
$F13$	Zadeh	Centroid	Figure 3 right	Figure 4 right	3
$F14$	Zadeh	Mean of max	Figure 3 right	Figure 4 right	3
$F15$	Product	Centroid	Figure 3 right	Figure 4 right	3
$F16$	Product	Mean of max	Figure 3 right	Figure 4 right	3
$F17$	Lukasiewicz	Centroid	Figure 3 right	Figure 4 right	3
$F18$	Lukasiewicz	Mean of max	Figure 3 right	Figure 4 right	3

3.2 Argument-based models

The definition of argument based-models follows the 5 layer modelling approach proposed in [14] and depicted on figure 1 (section 2.2).

Layer 1 - Definition of the structure of arguments The first step focuses on the construction of *forecast arguments* as it follows:

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

This structure is composed by a set of premises built upon the features considered in the NASA-TLX mental workload assessment instrument and a conclusion (MWL level) derivable by applying an inference rule $\rightarrow$. The categories

[2] `https://doi.org/10.6084/m9.figshare.6979865.v1`

associated to these conclusions are the same as the ones described in section 3.1. However, since no notion of gradualism is considered here, they are strictly bounded in well defined ranges (example, *low mental demand* in one knowledge-base is defined in the range $[0, 33) \in \mathbf{R}$). An example of a forecast argument is given below (it matches Rule 1 of section 3.1):

– ARG 1: *low* `mental demand` → `underload`

Layer 2 - Definition of the conflicts of arguments In order to evaluate inconsistencies and invalid arguments, *mitigating arguments* [19] are defined. These are formed by a set of premises and an undercutting inference ⇒ to an argument B (forecast or mitigating):

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

Both forecast and mitigating arguments follow a similar notion of *defeasible rules*, as defined in [23]. Informally, if their premises hold then presumably their conclusions also hold. In addition, mitigating arguments can be of different types. In this research, the notion of *undercutting attack* is employed for the resolution of conflicts. It defines an exception, where the application of the knowledge carried in some argument is no longer allowed. Contradictions, such as in section 3.1, represent the information necessary for the construction of undercutting attacks. For example, the corresponding mitigating argument that can be constructed from Exception 1 (section 3.1) through an undercutting attack is:

– UA1: *high* `effort` ⇒ ARG 1

All the designed arguments and attacks can now be seen as an *argumentation framework* (AF), as depicted in figure 6.

Layer 3 - Evaluation of the conflicts of arguments At this stage an AF can be elicited with data. Forecast and mitigating arguments can be activated or discarded, based on whether their premises evaluate true or false. Attacks between activated arguments are considered valid, while the others are discarded. Contrarily to fuzzy systems, there is no partial refutation, so a successful attack always refutes its target. From the activated forecast/mitigating arguments and valid attacks, a *sub-argumentation framework* emerges (sub-AF), as in figure 7 (this is equivalent to the Abstract Argumentation proposed by Dung [9]).

Layer 4 - Definition of the dialectical status of arguments Given a sub-AF acceptability semantics [7, 9] are applied in order to accept or reject its arguments. Each record of the dataset instantiates a different sub-AF, thus semantics have to applied for each different case. These are aimed at evaluating which arguments are defeated. An argument A is *defeated* by B if there is a valid attack from A to B [9]. Not only that, but it is also necessary to evaluate if the defeaters are defeated themselves. A set of non defeated arguments is called *extension* (conflict free set of arguments). Extensions are in turn used in the 5th layer of the diagram of figure 1, 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*

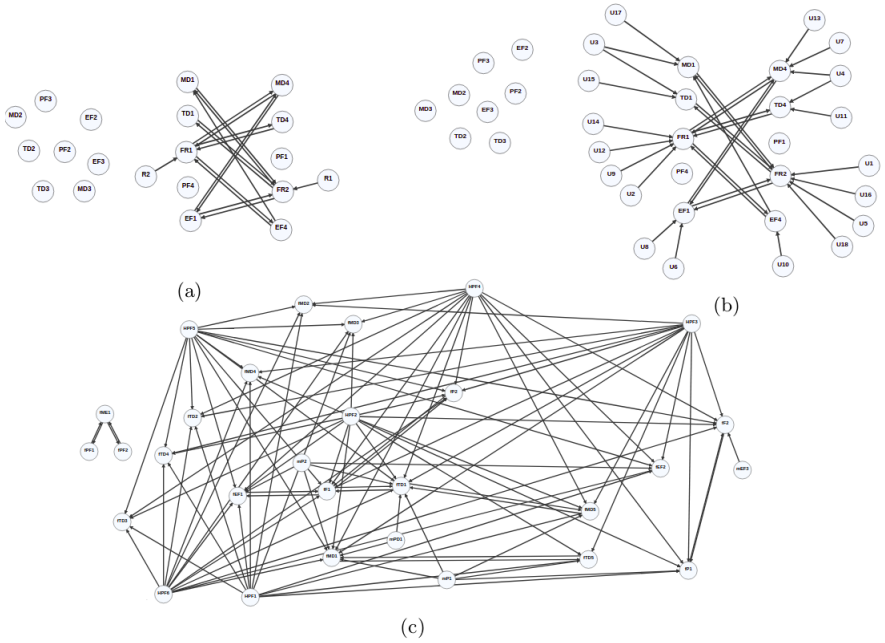

Fig. 6: Three knowledge-bases encoded as interactive arguments. Further details: `https://doi.org/10.6084/m9.figshare.6979865.v1`

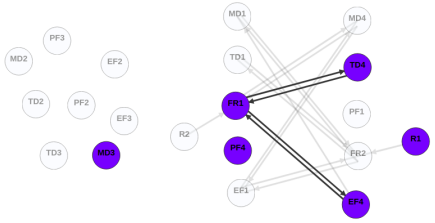

Fig. 7: Example of a sub-Argumentation framework (sub-AF): activated arguments (blue nodes). Light nodes and edges are still part of the original knowledge-base but, since not activated, they are not considered in the next layers.

(AAF) as proposed by Dung [9]. 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 $\subseteq$)

Layer 5 - Accrual of acceptable arguments Eventually, in the last step of the reasoning process, a final inference has to be produced for practical purposes. In case multiple extensions are computed, one extension might be preferred over the others. In this study, the cardinality of an extension (number of accepted

arguments) is used as a mechanism for the quantification of its credibility. Intuitively, a larger conflict-free extension of arguments might be seen as more credible 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 had their role 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. This scalar is essentially a linear relationship from the range of the argument's premise to the range of the argument's conclusion. For instance, if argument ARG 1 is activated by the lowest value of the `mental demand` range, then its final scalar will be the correspondent lowest value in its conclusion's range. 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, the average is considered. Table 3 summarises the design of the argument-based models.

Table 3: Designed argument-based models and their parameters across each layer.

Model	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
	Arguments	Conflicts	Conflict evaluation	Semantics	Accrual
A1	KB1	figure 6 (a)	binary	Grounded	cardinality + average
A2	KB1	figure 6 (a)		Preferred	
A3	KB2	figure 6 (b)		Grounded	
A4	KB2	figure 6 (b)		Preferred	
A5	KB3	figure 6 (c)		Grounded	
A6	KB3	figure 6 (c)		Preferred	

3.3 Participant and procedures

A number of third-level classes have been delivered to students at Dublin Institute of Technology. After each class, students had to fill in the questionnaire associated to the NASA-TLX instrument (as described in section 2.3). Students were from 23 different countries (age 21-74, mean 30.9, std= 7.67). Four different topics of the module 'research methods' were delivered in different semesters during the academic terms 2015-2017, as per table 4. Three different delivery methods were used: 1) traditional direct instruction, using slides projected to a white board; 2) multimedia video of content (same as in 1) projected to a white board; 3) constructivist collaborative activity added to 2. Summary statistics can be found in table 4. Beside completing the NASA-TLX questionnaire, participants were required to fill in another scale providing an indication of their experienced mental workload (figure 8). This was designed as a baseline and as a form of ground truth. It is believe that only the person executing the task can provide a precise account of the mental workload experienced [20].

Table 4: Number of students across topics and delivery methods topics

Topic	**Duration (Mins)**	**Delivery method**		
		1	**2**	**3**
Science	[18, 62]	13	36	6
Scientific method	[20, 46]	19	15	15
Research planning	[10, 68]	22	22	15
Literature review	[19, 55]	20	21	9

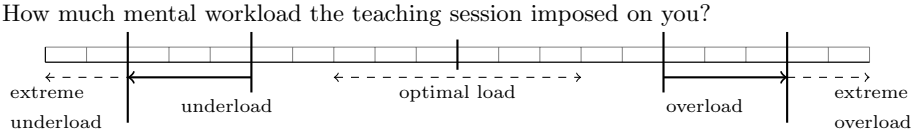

Fig. 8: Baseline self-reporting measure of Mental Workload

In details, in order to evaluate the inferential capacity of the models (built in tables 2 and 3) two forms of the validity of their inferences (scalar values) were adopted. As suggested in section 2.3, these are *convergent validity* and *face validity*. The former has been assessed through an analysis of the correlation of the inferences, produced by designed models, and the scores produced by the Nasa-Task Load Index. The latter has been assessed through an investigation of the error of designed models against the mental workload scores reported by students, using the scale of figure 8. Table 5 summarises these two forms of validity and the statistical test associated to them.

Table 5: Convergent and face validity, and associated statistical tests.

Validity	Definition	Statistical test
Convergent	It refers to the extent to which different MWL measures that should be theoretically related, are in fact related	Correlation coefficient
Face	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]

4 Results

The answers of the NASA-TLX questionnaire were used to elicit the designed non-monotonic fuzzy reasoning and argument-based models (tables 2, 3).

Convergent validity Figure 9 depicts the Spearman correlation coefficients of the inferences of the designed models and the NASA-TLX indexes. This statis-

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

tical test was used because the assumptions behind the Pearson correlation were not met. Moderate to high correlation (coefficients: 0.50-0.76) were generally observed. This indicates that, the assumption of the theoretical relationship between the NASA-TLX measure, known to fairly model the construct of mental workload, and the designed models in fact exists. As a consequence, it can be said that the non-monotonic fuzzy reasoning models and the argument-based models are fairly modelling mental workload in the designed experiment, regardless of the operators used to aggregate premises of fuzzy rules (Zadeh, Product, Lukasiewicz) or the semantics used in layer 4 (grounded/preferred).

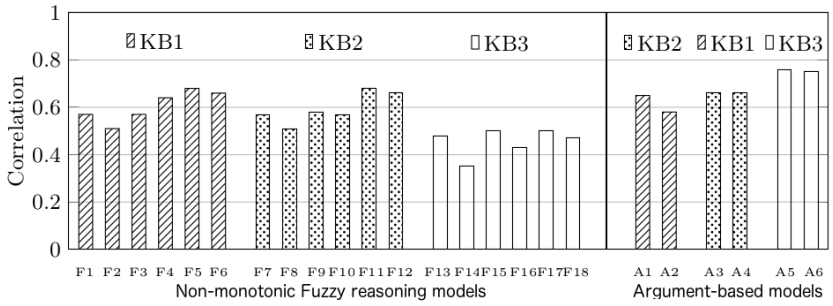

Fig. 9: Spearmans coefficients of inferences of models and NASA-TLX scores ($p < 0.05$)

Face validity Figure 10 depicts the mean squared errors of the inferences (scalar values) of each designed model. As it is easy to observe, argument-based models had a lower error when compared to the baseline instrument (NASA-TLX). The average of MSEs associated to fuzzy reasoning models (F1-18) was 407.75 while the average of MSEs of argument-based models (A1-6) was 229.83. Additionally, among the fuzzy reasoning models, the difference in the scores of those employing the *centroid* as defuzzification method (labelled with an upper dot) appear to be better than the ones employing the *mean of max*. As for argument-based models, there is only a slight difference across knowledge-bases and no significant difference among semantics (grounded/preferred). It is important to highlight that knowledge-base KB3 is certainly richer than KB1 and KB2, as it carries more interacting pieces of knowledge (figure 6). Despite this higher richness, the mean squared error did not decrease significantly when compared to KB1 and KB2 both for the fuzzy reasoning or argument-based models. However, when comparing the mean squared error of the fuzzy reasoning models against the ones of the argument-based models, it can be stated that the latter have a better inferential capacity over the former for the specific tasks and dataset employed. This statement holds regardless of the fuzzy operators employed in the fuzzy engines (fuzzy reasoning models); the semantics adopted in the conflict resolution layer (argument-based models); and the knowledge-bases considered.

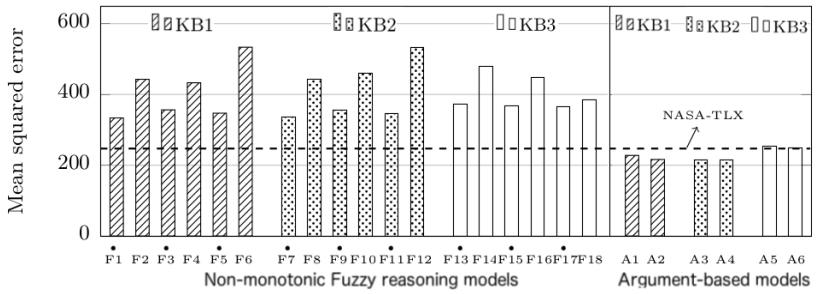

Fig. 10: Mean squared error of each model against the self-reported MWL

Argumentation, overall, had a better face validity and was superior in approximating the target (self-reported mental workload reported by students) than non-monotonic fuzzy reasoning. An analysis of the convergent validity of the models showed that their inferences can be considered valid. They are positively and moderately correlated to the well known (NASA-TLX), thus likely modelling mental workload too. A negative or null correlation would have implied the invalidity of the models since they would have probably modelled another construct. With a good convergent validity, the findings from the analysis of the face validity can be considered more reliable. This analysis indicated a better inferential capacity of the argument-based models over the fuzzy reasoning models for the selected tasks and data, despite the internal configuration and underlying knowledge-base employed. Argument-based models consistently showed a significantly lower mean squared error (difference between self-reported MWL and the inferences by designed models) over fuzzy reasoning models, in addition to a slight improvement also against the baseline instrument (NASA-TLX). This demonstrates the potential of argumentation as a modelling tool for knowledge-bases characterised by uncertainty, partiality and conflictual info.

5 Conclusion and future work

This study presented a comparison between non-monotonic fuzzy reasoning and non-nonotonic (defeasible) argumentation using three different knowledge-bases coded from an expert in the domain of mental workload. A primary research has been conducted including the construction of computational models using these two non-monotonic reasoning approaches to represent the construct of mental workload and to allow its assessment (inference). Such models were elicited with data provided by the NASA-Task Load Index questionnaire that was filled in by students who performed a set of learning tasks in a third-level context. The output of these models was a single scalar representing a level of mental workload that was used for comparison purposes. The selected metrics for evaluation of the inferential capacity of constructed models were convergent and face validity. Findings indicated how both the models built with the non-monotonic fuzzy reasoning mechanism and defeasible argumentation had a good convergent validity

with the NASA-TLX, confirming mental workload was actually the construct being modelled. However, the argument-based models had a significantly better face validity over the non-monotonic fuzzy reasoning models for the selected tasks and data. The novelty of this research lies in the quantification of the impact of argumentation through a novel empirical research in a real-world context employing primary data gathered from humans. Future work will concentrate on replicating this experiment by considering additional knowledge-bases and by extending the comparison of argumentation with other reasoning approaches such as expert systems. Moreover, the creation of inference models adopting fuzzy reasoning and argumentation such as in [6] is envisioned.

Acknowledgments

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.

References

1. Bench-Capon, T.J., Dunne, P.E.: Argumentation in artificial intelligence. Artificial intelligence 171(10-15), 619–641 (2007)
2. Bryant, D., Krause, P.: A review of current defeasible reasoning implementations. The Knowledge Engineering Review 23(3), 227–260 (2008)
3. Cain, B.: A review of the mental workload literature. Tech. rep., Defence research and development Toronto (Canada) (2007)
4. Castro, J.L., Trillas, E., Zurita, J.M.: Non-monotonic fuzzy reasoning. Fuzzy Sets and Systems 94(2), 217–225 (1998)
5. Chesñevar, C.I., Maguitman, A.G., Loui, R.P.: Logical models of argument. ACM Computing Surveys (CSUR) 32(4), 337–383 (2000)
6. Dondio, P.: 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 (2017)
7. Dondio, P.: Ranking semantics based on subgraphs analysis. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1132–1140. AAMAS '18, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2018), http://dl.acm.org/citation.cfm?id=3237383.3237864
8. Dubois, D., Prade, H.: Possibility theory: qualitative and quantitative aspects. In: Quantified representation of uncertainty and imprecision, pp. 169–226 (1998)
9. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence 77(2), 321–358 (1995)
10. Eggemeier, F.T.: Properties of workload assessment techniques. Advances in Psychology 52, 41–62 (1988)
11. Gegov, A., Gobalakrishnan, N., Sanders, D.: Rule base compression in fuzzy systems by filtration of non-monotonic rules. Journal of Intelligent & Fuzzy Systems 27(4), 2029–2043 (2014)

12. Glasspool, D., Fox, J., Oettinger, A., Smith-Spark, J.: Argumentation in decision support for medical care planning for patients and clinicians. In: AAAI Spring Symposium: Argumentation for Consumers of Healthcare. pp. 58–63 (2006)
13. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in Psychology 52(C), 139–183 (1988)
14. Longo, L.: Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. In: Machine Learning for Health Informatics, pp. 183–208. Springer (2016)
15. Longo, L., Dondio, P.: 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 (2014)
16. Longo, L., Hederman, L.: Argumentation Theory for Decision Support in Health-Care: A Comparison with Machine Learning, pp. 168–180. Springer, Cham (2013)
17. Longo, L., Kane, B., Hederman, L.: Argumentation theory in health care. In: Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on. pp. 1–6. IEEE (2012)
18. Mardani, A., Jusoh, A., Zavadskas, E.K.: Fuzzy multiple criteria decision-making techniques and applications–two decades review from 1994 to 2014. Expert Systems with Applications 42(8), 4126–4148 (2015)
19. Matt, P.A., Morgem, M., Toni, F.: Combining statistics and arguments to compute trust. In: 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada. vol. 1, pp. 209–216. ACM (May 2010)
20. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: Int. Symposium on Human Mental Workload: Models and Applications. pp. 30–50 (2017)
21. O'Donnell, R., Eggemeier, F.: 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 (1986)
22. Passino, K.M., Yurkovich, S., Reinfrank, M.: Fuzzy control, vol. 20. Citeseer (1998)
23. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument and Computation 1(2), 93–124 (2010)
24. Reiter, R.: A logic for default reasoning. Artificial intelligence 13(1-2), 81–132 (1980)
25. Rizzo, L., Longo, L.: 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 (2017)
26. Rizzo, L., Majnaric, L., Dondio, P., Longo, L.: An investigation of argumentation theory for the prediction of survival in elderly using biomarkers. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 385–397. Springer (2018)
27. Siler, W., Buckley, J.J.: Fuzzy expert systems and fuzzy reasoning. John Wiley & Sons (2005)
28. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics 39(3), 358–381 (1996)
29. Zadeh, L.A., et al.: Fuzzy sets. Information and control 8(3), 338–353 (1965)