International Journal of Approximate Reasoning 164 (2024) 109065

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

A preferential interpretation of MultiLayer Perceptrons in a conditional logic with typicality

Mario Alviano [a], Francesco Bartoli [b], Marco Botta [b], Roberto Esposito [b], Laura Giordano [c], Daniele Theseider Dupré [c,*]

[a] *Università della Calabria, Italy*
[b] *Università di Torino, Italy*
[c] *Università del Piemonte Orientale, Italy*

ARTICLE INFO

Keywords:
Conditional logics
Description logics
Many-valued logics
Neural networks

ABSTRACT

In this paper we investigate the relationships between a multipreferential semantics for defeasible reasoning in knowledge representation and a multilayer neural network model. Weighted knowledge bases for a simple description logic with typicality are considered under a (many-valued) "concept-wise" multipreference semantics. The semantics is used to provide a preferential interpretation of MultiLayer Perceptrons (MLPs). A model checking and an entailment based approach are exploited in the verification of conditional properties of MLPs.

1. Introduction

Preferential approaches to commonsense reasoning [1–9] have their roots in conditional logics [10,11], and have been used to provide axiomatic foundations of non-monotonic or defeasible reasoning. They have been extended to Description Logics (DLs) [12], to deal with inheritance with exceptions in ontologies, by allowing for non-strict forms of inclusions, called *typicality or defeasible inclusions*, with different preferential semantics [13–15], and closure constructions [16–22]. Preferential extensions of DLs allow reasoning with exceptions through the identification of *prototypical properties* of individuals or classes of individuals.

In recent work, a "concept-wise" multi-preferential semantics has been proposed as a semantics of ranked knowledge bases (KBs) in a lightweight description logic [23], in which defeasible or typicality inclusions of the form $\mathbf{T}(C) \sqsubseteq D$ (meaning "the typical C's are D's" or "normally C's are D's") are given a rank, a natural number representing their strength. This two-valued concept-wise multi-preferential semantics, which takes into account preferences with respect to different concepts, has been shown to have some desirable properties from the knowledge representation point of view [23,24], and has also been used to develop a preferential interpretation for Self-Organising Maps [25], psychologically and biologically plausible neural network models.

The idea underlying the multi-preferential semantics is that different preferences should be associated to different concepts and, for instance, for two individuals Tom and Bob, and two concepts, *Swimmer* and *Student*, Tom might be more typical than Bob as a swimmer ($tom <_{Swimmer} bob$) but less typical than Bob as a student ($bob <_{Student} tom$).

In this paper, we focus on *weighted defeasible knowledge bases* (KBs), i.e., KBs in which typicality inclusions (conditionals) have a positive or negative *weight*, a real number representing the plausibility of the property. For instance, one may want to represent a

* Corresponding author.
E-mail address: dtd@uniupo.it (D. Theseider Dupré).

https://doi.org/10.1016/j.ijar.2023.109065
Received 3 May 2023; Received in revised form 8 September 2023; Accepted 22 October 2023

Available online 7 November 2023

situation in which students are normally young and use to have classes, while they usually do not have a scholarship. In a weighted knowledge base these defeasible properties of students may be represented through some weighted typicality inclusions such as:

$$\mathbf{T}(Student) \sqsubseteq Young, 80$$

$$\mathbf{T}(Student) \sqsubseteq \exists hasClasses.\top, 90$$

$$\mathbf{T}(Student) \sqsubseteq \exists hasScholarship.\top, -20$$

where negative weights represent implausible properties, so that, in this example, it is rather implausible for students to have a scholarship, while it is quite plausible for them being young and having classes (with having classes slightly more plausible than being young). Given such properties, a student Bob, who is young, has classes and has no scholarship, can be regarded as being more typical than a student Tom who is not young, but has classes and has a scholarship, so that $bob <_{Student} tom$. Similarly, for concept *Swimmer* the prototypical elements can be characterized by a set of features and, hence, a set of typicality inclusions with their weights. In our approach such features (such as being young or having classes) are as well represented as concepts in the description logic.

We do not assume that concepts are crisp, but that a domain element (Bob) may belong to a concept (e.g., Young) to some degree. Hence, we build our approach on fuzzy description logics, which have been widely studied in the literature (see, for instance, [26–30]). We develop a fuzzy formulation of the concept-wise multi-preferential semantics for weighted KBs in the description logic $\mathcal{ALC}$, based on a *non-crisp* interpretation of typicality concepts (this different choice with respect to previous work, as we will see, has some impact on the properties of entailment). We start from a fuzzy extension of $\mathcal{ALC}$ [28], and further extend it with multiple preferences and with a non-crisp notion of typicality. The resulting fuzzy description logic with typicality is called $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$. To provide a semantics for weighted KBs, we introduce three different closure constructions for $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$, the *coherent*, the *faithful* and the *φ-coherent* multi-preferential semantics for weighted knowledge bases. Such constructions are similar in spirit to other semantic constructions adopted in the logics of commonsense reasoning, such as the *lexicographic closure* [31] and *c-representations* [9,32], but exploit multiple preference relations associated to concepts.

While similar (but different) semantic constructions for weighted knowledge bases have been considered in previous work based on different description logics,[1] here we aim at a uniform formulation of the three semantics for $\mathcal{ALC}$, under the assumption that the interpretation of typicality is non-crisp. This allows us to study their mutual relationships, and to prove additional properties of multi-preferential entailment for the different semantics.

In particular, we show that any φ-coherent model of a weighted KB is a faithful (resp., coherent) model of the KB under suitable conditions, and that the notions of entailment under the different semantics satisfies (for some choice of fuzzy combination functions) all the *KLM properties of a preferential consequence relation* [4,6], as well as other properties of the typicality operator studied in [15] for $\mathcal{ALC}+\mathbf{T}$, a two-valued typicality extension of $\mathcal{ALC}$. This contribution of the work extends the preliminary results on the properties of the typicality logic investigated in [34] under the faithful semantics. In that case, the faithful semantics failed to satisfy all KLM properties of a preferential consequence relation, a negative result which is now overcome by adopting a non-crisp interpretation of typicality.

The proposed (fuzzy) many-valued multi-preferential semantics are used in providing a *logical characterization* of Multilayer Perceptrons (MLPs) [36], which can be used for post-hoc verification. We will see that the input-output behavior of a multilayer network $\mathcal{N}$ can be captured by a preferential interpretation $I^{\Delta}_{\mathcal{N}}$ built over a set of input stimuli Δ (e.g., the training set), through a simple construction, which exploits the activity level of units for the input stimuli, thus allowing for the verification of properties of the network by *model checking* over the preferential interpretation. We show that properties formalized as fuzzy typicality inclusions in the boolean fragment of $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ can be verified on the interpretation $I^{\Delta}_{\mathcal{N}}$ in polynomial time in the size of $I^{\Delta}_{\mathcal{N}}$ and in the size of the property. This is another contribution of the paper.

A logical characterization of a trained multi-layer networks $\mathcal{N}$ is established by proving that the preferential interpretation $I^{\Delta}_{\mathcal{N}}$, describing the network behavior over a set Δ of input stimuli, is indeed a φ-coherent model of the weighted knowledge base $K^{\mathcal{N}}$ and, vice-versa, that any φ-coherent model of the knowledge base $K^{\mathcal{N}}$ captures the behavior of the network over some set Δ of input stimuli. This strengthens the result in [33] that the interpretation $I^{\Delta}_{\mathcal{N}}$ is a coherent model of $K^{\mathcal{N}}$.

Undecidability results for fuzzy DLs with general inclusion axioms [37,30] has led to consider a finitely-valued notion of the φ-coherent semantics, the φ_n-coherent semantics [38]. In this paper, we prove that the φ_n-coherent semantics is indeed an *approximation* of the φ-coherent semantics, which provides a full path from the definition of the fuzzy typicality logic with its semantics to the verification of properties of feedforward neural networks, which is based on proof methods developed for φ_n-coherent entailment [39] and on a Datalog encoding of the model checking approach [40]. In the experimentation, we exploit both the entailment-based approach and the model-checking approach in the verification of properties of trained multilayer feedforward networks. Such properties, expressed as fuzzy typicality inclusions, rely on typicality in order to describe what the network has learned to be a typical member of a class. The experiments extend and complement the ones reported in [38,40,41].

The schedule of the paper is the following. Section 2 contains the preliminaries about the description logic $\mathcal{ALC}$ and its fuzzy version. Section 3 defines a (monotonic) extension of fuzzy $\mathcal{ALC}$ (called $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$) including a fuzzy notion of typicality. Section 4

[1] In particular, the coherent semantics was first introduced in [33] for weighted $\mathcal{EL}^{\perp}$ KBs under a fuzzy semantics; the faithful semantics was considered in [34] for weighted $\mathcal{ALC}$ KBs in the fuzzy case; the φ-coherent semantics was first proposed as an argumentation semantics in [35]. In all cases the interpretation of typicality was crisp.

introduces *weighted knowledge bases* in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ and their closure constructions through the notions of *faithful*, *coherent* and φ-*coherent* (fuzzy) multi-preferential models. It establishes their relationships and defines the associated notions of entailment. Section 5 studies the properties of $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ and its closures, and proves that, for Gödel fuzzy combination functions, 1-entailment satisfies all KLM properties of a preferential consequence relation [6] (properties that also extend to k-entailment, except for Cautious Monotonicity), while Rational Monotonicity does not hold. Further properties of the notion of typicality are also studied. Section 6 establishes the relationships between multi-preferential semantics and multilayer networks. It is proven that a multilayer network can be interpreted as a (fuzzy) multi-preferential interpretation (Section 6.2), and that the network itself can be regarded as a weighted knowledge base in the boolean fragment of $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ (Section 6.3). This allows both a *model-checking approach* and an *entailment approach* to be exploited for property verification. Section 7 proves that the φ-coherent models can be approximated in the finitely-valued case, which justifies the use of the φ-coherent semantics over a finite domain for verification. Section 8 reports about experiments in the verification of properties of feedforward neural networks for the recognition of basic emotions, based on both the entailment and the model-checking approaches. The networks considered for entailment are significantly larger than the ones considered in [23,41], implying a much larger search space for solving. How model checking and entailment can be used together, and can be seen as complementary, is also pointed out. Section 9 concludes the paper with a discussion of related work and open issues.

The paper extends the work in [40,41] by substantiating it with several technical contributions, including the above mentioned ones.

2. The description logic $\mathcal{ALC}$ and fuzzy $\mathcal{ALC}$

Fuzzy description logics have been widely studied in the literature for representing vagueness in description logics, e.g., by [26–28,30], based on the idea that concepts and roles can be interpreted as fuzzy sets and fuzzy relations. In fuzzy DLs, formulas have a truth degree from a truth space S, usually the interval $[0,1]$, as in Mathematical Fuzzy Logic [42]. The finitely many-valued case is also well studied for DLs [29,43–45].

In this section we recall the syntax and semantics of the description logic $\mathcal{ALC}$ [12] and of its fuzzy extension [28]. We will also consider a finitely many-valued fragment of $\mathcal{ALC}$ with typicality.

2.1. $\mathcal{ALC}$

Let N_C be a set of concept names, N_R a set of role names and N_I a set of individual names. The set of $\mathcal{ALC}$ *concepts* (or, simply, concepts) can be defined inductively as follows:

- $A \in N_C$, $\top$ and $\bot$ are concepts;
- if C and D are concepts, and $r \in N_R$, then $C \sqcap D$, $C \sqcup D$, $\neg C$, $\forall r.C$, $\exists r.C$ are concepts.

A knowledge base (KB) K is a pair $(\mathcal{T}, \mathcal{A})$, where $\mathcal{T}$ is a TBox and $\mathcal{A}$ is an ABox. The TBox $\mathcal{T}$ is a set of concept inclusions (or subsumptions) $C \sqsubseteq D$, where C, D are concepts. The ABox $\mathcal{A}$ is a set of assertions of the form $C(a)$ and $r(a,b)$ where C is a concept, a and b are individual names in N_I and r a role name in N_R.

An $\mathcal{ALC}$ *interpretation* is defined as a pair $I = \langle \Delta, \cdot^I \rangle$ where: Δ is a domain—a set whose elements are denoted by $x, y, z, \ldots$—and $\cdot^I$ is an extension function that maps each concept name $C \in N_C$ to a set $C^I \subseteq \Delta$, each role name $r \in N_R$ to a binary relation $r^I \subseteq \Delta \times \Delta$, and each individual name $a \in N_I$ to an element $a^I \in \Delta$. It is extended to complex concepts as follows:

$$\bot^I = \emptyset,$$
$$\top^I = \Delta,$$
$$(\neg C)^I = \Delta \backslash C^I,$$
$$(C \sqcap D)^I = C^I \cap D^I,$$
$$(C \sqcup D)^I = C^I \cup D^I,$$
$$(\exists r.C)^I = \{x \in \Delta \mid \exists y.(x,y) \in r^I \text{ and } y \in C^I\},$$
$$(\forall r.C)^I = \{x \in \Delta \mid \forall y.(x,y) \in r^I \Rightarrow y \in C^I\}.$$

The notion of satisfiability of a KB in an interpretation and the notion of entailment are defined as follows:

Definition 1 *(Satisfiability and entailment).* Given an $\mathcal{ALC}$ interpretation $I = \langle \Delta, \cdot^I \rangle$:
- I satisfies an inclusion $C \sqsubseteq D$ if $C^I \subseteq D^I$;
- I satisfies an assertion $C(a)$ (resp., $r(a,b)$) if $a^I \in C^I$ (resp., $(a^I, b^I) \in r^I$).

Given a knowledge base $K = (\mathcal{T}, \mathcal{A})$, an interpretation I satisfies $\mathcal{T}$ (resp. $\mathcal{A}$) if I satisfies all inclusions in $\mathcal{T}$ (resp. all assertions in $\mathcal{A}$); I is a *model* of K if I satisfies $\mathcal{T}$ and $\mathcal{A}$.

A subsumption $F = C \sqsubseteq D$ (resp., an assertion $C(a)$, $r(a,b)$), is entailed by K, written $K \models F$, if for all models $I = \langle \Delta, \cdot^I \rangle$ of K, I satisfies F.

Table 1
Properties for t-norms and s-norms.

Axiom	T-norm	S-norm
Tautology/contradiction	$a \otimes 0 = 0$	$a \oplus 1 = 1$
Identity	$a \otimes 1 = a$	$a \oplus 0 = a$
Commutativity	$a \otimes b = b \otimes a$	$a \oplus b = b \oplus a$
Associativity	$(a \otimes b) \otimes c = a \otimes (b \otimes c)$	$(a \oplus b) \oplus c = a \oplus (b \oplus c)$
Monotonicity	if $b \leq c$, then $a \otimes b \leq a \otimes c$	if $b \leq c$, then $a \oplus b \leq a \oplus c$

Table 2
Properties for implication and negation functions.

Axiom	Implication function	Negation function
Tautology/contradiction	$0 \rhd b = 1, a \rhd 1 = 1, 1 \rhd 0 = 0$	$\ominus 0 = 1, \ominus 1 = 0$
Antitonicity	if $a \leq b$, then $a \rhd c \geq b \rhd c$	if $a \leq b$, then $\ominus a \geq \ominus b$
Monotonicity	if $b \leq c$, then $a \rhd b \leq a \rhd c$	

Given a knowledge base K, the *subsumption* problem is the problem of deciding whether an inclusion $C \sqsubseteq D$ is entailed by K.

2.2. Fuzzy $\mathcal{ALC}$ and a finitely-valued $\mathcal{ALC}$

We shortly recall the semantics of a fuzzy extension of $\mathcal{ALC}$, referring to the survey by Lukasiewicz and Straccia [28]. We limit our consideration to a few features of a fuzzy DL and, in particular, we omit considering datatypes.

A *fuzzy interpretation* for $\mathcal{ALC}$ is a pair $I = \langle \Delta, \cdot^I \rangle$ where: Δ is a non-empty domain and $\cdot^I$ is *fuzzy interpretation function* that assigns to each concept name $A \in N_C$ a function $A^I : \Delta \to [0,1]$, to each role name $r \in N_R$ a function $r^I : \Delta \times \Delta \to [0,1]$, and to each individual name $a \in N_I$ an element $a^I \in \Delta$. A domain element $x \in \Delta$ belongs to the extension of A to some degree in $[0,1]$, i.e., A^I is a fuzzy set.

The interpretation function $\cdot^I$ is extended to complex concepts as follows:

$$\top^I(x) = 1, \qquad \bot^I(x) = 0,$$
$$(\neg C)^I(x) = \ominus C^I(x),$$
$$(C \sqcap D)^I(x) = C^I(x) \otimes D^I(x),$$
$$(C \sqcup D)^I(x) = C^I(x) \oplus D^I(x),$$
$$(\exists r.C)^I(x) = \sup_{y \in \Delta} r^I(x,y) \otimes C^I(y),$$
$$(\forall r.C)^I(x) = \inf_{y \in \Delta} r^I(x,y) \rhd C^I(y),$$

where $x \in \Delta$, and $\otimes$, $\oplus$, $\rhd$ and $\ominus$ are arbitrary but fixed triangular norm (or *t-norm*), triangular co-norm (or *s-norm*), implication function, and negation function, chosen among the combination functions of some fuzzy logic. In particular, in Gödel logic $a \otimes b = min\{a,b\}$, $a \oplus b = max\{a,b\}$, $a \rhd b = 1$ *if* $a \leq b$ *and* b *otherwise*; $\ominus a = 1$ *if* $a = 0$ *and* 0 *otherwise*. In Łukasiewicz logic, $a \otimes b = max\{a + b - 1, 0\}$, $a \oplus b = min\{a + b, 1\}$, $a \rhd b = min\{1 - a + b, 1\}$ and $\ominus a = 1 - a$. In Product Logic, $a \otimes b = a \cdot b$, $a \oplus b = a + b - a \cdot b$, $a \rhd b = min\{1, b/a\}$ and $\ominus a = 1$ *if* $a = 0$ *and* 0 *otherwise*.[2] Following [28], we will not commit to a specific choice of combination functions, but in Tables 1 and 2 we report their main properties (from Tables 1 and 2 in [28]).

The interpretation function $\cdot^I$ is also extended to non-fuzzy axioms (i.e., to strict inclusions and assertions of an $\mathcal{ALC}$ knowledge base) as follows:

$$(C \sqsubseteq D)^I = \inf_{x \in \Delta} C^I(x) \rhd D^I(x)$$
$$(C(a))^I = C^I(a^I)$$
$$(R(a,b))^I = R^I(a^I, b^I).$$

A *fuzzy $\mathcal{ALC}$ knowledge base* K is a pair $(\mathcal{T}, \mathcal{A})$ where $\mathcal{T}$ is a fuzzy TBox and $\mathcal{A}$ a fuzzy ABox. A fuzzy TBox is a set of *fuzzy concept inclusions* of the form $C \sqsubseteq D \,\theta\, n$, where $C \sqsubseteq D$ is an $\mathcal{ALC}$ concept inclusion axiom, $\theta \in \{\geq, \leq, >, <\}$ and $n \in [0,1]$. A fuzzy ABox $\mathcal{A}$ is a set of *fuzzy assertions* of the form $C(a)\theta n$ or $r(a,b)\theta n$, where C is an $\mathcal{ALC}$ concept, $r \in N_R$, $a, b \in N_I$, $\theta \in \{\geq, \leq, >, <\}$ and $n \in [0,1]$. Following Bobillo and Straccia [46], we assume that fuzzy interpretations are *witnessed*, i.e., the sup and inf are attained at some point of the involved domain.

We refer to fuzzy concept inclusions and fuzzy assertions as *fuzzy axioms*.

[2] Let us mention that any continuous t-norm can be expressed as an ordinal sum of copies of these three t-norms.

Example 1. Let us consider the fuzzy concepts *Tall* (tall individuals) and $\exists hasFriend.Tall$ (the individuals having a tall friend), where hasFriend might as well be a fuzzy role, as a domain individual Bob may be friend of Mary to a given degree, e.g., $hasFriend^I(bob^I, mary^I) = 0.7$. Similarly, we may consider the fuzzy concept $\exists hasParent.Tall$.

For instance, we may have fuzzy assertions such as $hasFriend(bob, mary) \geq 0.5$ or $Tall(mary) \geq 0.8$ in the ABox $\mathcal{A}_f$, and fuzzy concept inclusions such as $\exists hasParent.Tall \sqsubseteq Tall \geq 0.7$ (an individual having at least a tall parent is tall, holding to a degree greater than or equal to 0.7) or $\forall hasFriend.Nerd \sqsubseteq Nerd \geq 0.8$ (an individual having all nerd friends is a nerd, holding to a degree greater than 0.8) in the TBox $\mathcal{T}_f$.

Let us assume Gödel logic, and that Bob has parents Mary and Tom. Consider an interpretation I such that:

$$hasParent^I(bob^I, mary^I) = 1,\ hasParent^I(bob^I, tom^I) = 1,$$

$$hasParent^I(bob^I, z) = 0,\ \textit{for all } z \in \Delta \textit{ with } z \neq mary^I, tom^I,$$

$$hasParent^I(x, z) = 0,\ \textit{for all } x, z \in \Delta \textit{ with } x \neq bob^I,$$

$$Tall^I(bob^I) = 0.8,\ Tall^I(mary^I) = 0.5,\ Tall^I(tom^I) = 0.9,$$

$$Tall^I(x) = 0.5,\ \textit{for all } x \in \Delta \textit{ with } x \neq bob^I, mary^I, tom^I.$$

As an example, we show that $(\exists hasParent.Tall \sqsubseteq Tall)^I = 0.8$ holds. We have to show that $inf_{x\in\Delta}(\exists hasParent.Tall)^I(x) \rhd Tall^I(x) = 0.8$. Note that:

$$\begin{aligned}(\exists hasParent.Tall)^I(x) &= \sup_{y\in\Delta} hasParent^I(x,y) \otimes Tall^I(y)\\ &= \sup_{y\in\Delta} min\{hasParent^I(x,y), Tall^I(y)\}.\end{aligned}$$

In particular, for $x = bob^I$ we have:

$$\begin{aligned}(\exists hasParent.Tall)^I(bob^I) &= \sup_{y\in\Delta} min\{hasParent^I(bob^I, y), Tall^I(y)\},\\ &= \sup_{y\in\Delta} min\{hasParent^I(bob^I, y), Tall^I(y)\} = 0.9.\end{aligned}$$

In fact, we have three possible cases for y, $y = mary^I$, $y = tom^I$ and $y \neq mary^I, tom^I$:

$$min\{hasParent^I(bob^I, mary^I), Tall^I(mary^I)\} = min\{1, 0.5\} = 0.5$$

$$min\{hasParent^I(bob^I, tom^I), Tall^I(tom^I)\} = min\{1, 0.9\} = 0.9$$

$$min\{hasParent^I(bob^I, y), Tall^I(y)\} = min\{0, Tall^I(y)\} = 0,$$

for $y \neq mary^I, tom^I$. We take the maximum among the values, then

$$(\exists hasParent.Tall)^I(bob^I) \rhd Tall^I(bob^I) = 0.9 \rhd 0.8 = 0.8.$$

In a similar way, one can see that, for all $x \neq bob^I$,

$$(\exists hasParent.Tall)^I(x) \rhd Tall^I(x) = 0 \rhd 0.5 = 1,$$

where $(\exists hasParent.Tall)^I(x) = 0$ as, for $x \neq bob^I$, $hasParent^I(x, z) = 0$ for all $z \in \Delta$.

Thus: $inf_{x\in\Delta}(\exists hasParent.Tall)^I(x) \rhd Tall^I(x) = 0.8$.

The notions of satisfiability of a KB in a fuzzy interpretation and of entailment are defined in the natural way.

Definition 2 *(Satisfiability and entailment for fuzzy KBs).* A fuzzy interpretation I satisfies a fuzzy $\mathcal{ALC}$ axiom E (denoted $I \models E$), as follows:

- I satisfies a fuzzy inclusion axiom $C \sqsubseteq D\ \theta\ n$ if $(C \sqsubseteq D)^I\ \theta\ n$;
- I satisfies a fuzzy assertion $C(a)\ \theta\ n$ if $C^I(a^I)\theta\ n$;
- I satisfies a fuzzy assertion $r(a, b)\ \theta\ n$ if $r^I(a^I, b^I)\theta\ n$,

where $\theta \in \{\geq, \leq, >, <\}$.

Given a fuzzy $\mathcal{ALC}$ knowledge base $K = (\mathcal{T}, \mathcal{A})$, a fuzzy interpretation I satisfies $\mathcal{T}$ (resp. $\mathcal{A}$) if I satisfies all fuzzy inclusions in $\mathcal{T}$ (resp. all fuzzy assertions in $\mathcal{A}$). A fuzzy interpretation I is a *model* of K if I satisfies $\mathcal{T}$ and $\mathcal{A}$. A fuzzy axiom E is *entailed* by a fuzzy knowledge base K, written $K \models E$, if for all models $I = \langle\Delta, \cdot^I\rangle$ of K, I satisfies E.

Example 2. Referring to the interpretation I in Example 1, we have seen that $(\exists hasParent.Tall \sqsubseteq Tall)^I = 0.8$ holds. Then, we can conclude that axiom $\exists hasParent.Tall \sqsubseteq Tall \geq 0.7$ is satisfied in I.

For the finitely many-valued case, we restrict to the boolean fragment $\mathcal{LC}$ of $\mathcal{ALC}$ with no roles (and no universal and existential restrictions). We assume the truth space to be $C_n = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$, for an integer $n \geq 1$.

A *finitely many-valued interpretation* for $\mathcal{ALC}$ is a pair $I = \langle \Delta, \cdot^I \rangle$ where: Δ is a non-empty domain and $\cdot^I$ is an *interpretation function* that assigns to each $a \in N_I$ a value $a^I \in \Delta$, and to each $A \in N_C$ a function $A^I : \Delta \to C_n$ and to each role name $r \in N_R$ a function $r^I : \Delta \times \Delta \to C_n$. In particular, in [38] we have considered two finitely many-valued fragments on $\mathcal{ALC}$, the one based on Łukasiewicz logic, and the other based on Gödel logic extended with a standard involutive negation $\ominus a = 1 - a$. Such fragments are defined along the lines of the finitely many-valued extension of description logic $\mathcal{SROIQ}$ [43], of the logic GZ $\mathcal{SROIQ}$ [44], and of the logic $\mathcal{ALC}^*(S)$ [29].

In the following, we will use $\mathcal{ALC}_n$ to refer to a finitely-valued extension of $\mathcal{ALC}$ interpreted over the truth space C_n, without committing to a specific choice of combination functions. In Section 8, we will mainly refer to $G_n\mathcal{LC}$, the boolean fragment of $\mathcal{ALC}_n$ based on Gödel logic, and we will assume that the interpretation of negated concepts exploits involutive negation, i.e., $(\neg C)^I(x) = \ominus C^I(x) = 1 - C^I(x)$.

3. Fuzzy $\mathcal{ALC}$ with typicality: $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$

In this section, we extend fuzzy $\mathcal{ALC}$ with typicality concepts of the form $\mathbf{T}(C)$, where C is a concept in fuzzy $\mathcal{ALC}$. The idea is similar to the extension of $\mathcal{ALC}$ with typicality [15], but transposed to the fuzzy case. The extension allows for the definition of *fuzzy typicality inclusions* of the form $\mathbf{T}(C) \sqsubseteq D\ \theta\ n$, meaning that typical C-elements are D-elements with a degree greater than n. A typicality inclusion $\mathbf{T}(C) \sqsubseteq D$, as in the two-valued case, stands for a KLM conditional implication $C \mathrel{|\!\sim} D$ [4,6], but now it has an associated degree.

We call $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ the extension of fuzzy $\mathcal{ALC}$ with typicality. As in the two-valued case, such as in $\mathcal{SROIQ}^{\mathbf{P}}\mathbf{T}$, a preferential extension of $\mathcal{SROIQ}$ with typicality [47], or in the propositional typicality logic, PTL [48], the typicality concept may be allowed to freely occur within inclusions and assertions, while the nesting of the typicality operator is not allowed.

In the definition of the semantics for $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$, we diverge from the choice in [33,38] and consider a fuzzy interpretation for the typicality operator, rather than a crisp one. This will allow us to prove that all the properties of a preferential consequence relation hold for a notion of entailment.

Observe that, in a fuzzy $\mathcal{ALC}$ interpretation $I = \langle \Delta, \cdot^I \rangle$, the degree of membership $C^I(x)$ of the domain elements x in a concept C induces a preference relation $<_C$ on Δ as follows:

$$x <_C y \text{ iff } C^I(x) > C^I(y) \tag{1}$$

Each preference $<_C$ has the properties of preference relations in KLM-style ranked interpretations [6], that is, $<_C$ is a modular and well-founded strict partial order. Let us recall that, $<_C$ is *well-founded* if there is no infinite descending chain $x_1 <_C x_0$, $x_2 <_C x_1$, $x_3 <_C x_2, \ldots$ of domain elements; $<_C$ is *modular* if, for all $x, y, z \in \Delta$, $x <_C y$ implies ($x <_C z$ or $z <_C y$). Well-foundedness holds for the induced preference $<_C$ defined by condition (1) under the assumption that fuzzy interpretations are witnessed [46] (see Section 2) or that Δ is finite.

While each preference relation $<_C$ has the properties of a preference relation in KLM rational interpretations [6] (also called ranked interpretations), here there are multiple preferences and, therefore, fuzzy interpretations can be regarded as *multipreferential* interpretations, which have also been studied in the two-valued case [23,49,21].

Each preference relation $<_C$ captures the relative typicality of domain elements wrt concept C and may be used to identify the *typical C-elements*. We will regard typical C-elements as the domain elements x that are preferred with respect to relation $<_C$ among those such that $C^I(x) \neq 0$.

For an interpretation I, let $C^I_{>0}$ be the crisp set containing all domain elements x such that $C^I(x) > 0$, that is, $C^I_{>0} = \{x \in \Delta \mid C^I(x) > 0\}$. The (fuzzy) interpretation of typicality concepts $\mathbf{T}(C)$ in I is:

$$(\mathbf{T}(C))^I(x) = \begin{cases} C^I(x) & \text{if } x \in min_{<_C}(C^I_{>0}) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $min_{<_C}(S) = \{u : u \in S \text{ and } \nexists z \in S \text{ s.t. } z <_C u\}$. When $(\mathbf{T}(C))^I(x) > 0$, we say that x is a typical C-element in I. Note that all typical C-elements have the same membership degree in concept C.

Observe also that, if $C^I(x) > 0$ for some $x \in \Delta$, $min_{<_C}(C^I_{>0})$ is non-empty (and the size of the fuzzy concept $(\mathbf{T}(C))^I$ is greater than zero). This generalizes the property that, in the crisp case, $C^I \neq \emptyset$ implies $(\mathbf{T}(C))^I \neq \emptyset$.

Let us define a *fuzzy multi-preferential interpretation* for $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ (shortly, an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation) as follows:

Definition 3 *($\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation).* An $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ is fuzzy $\mathcal{ALC}$ interpretation, equipped with the valuation of typicality concepts given by condition (2) above.

The fuzzy interpretation $I = \langle \Delta, \cdot^I \rangle$ implicitly defines a multi-preferential interpretation, where any concept C is associated to a preference relation $<_C$. This is different from the two-valued multi-preferential semantics in [23], where only a subset of distinguished concepts have an associated preference, and a notion of global preference $<$ is introduced to define the interpretation of the typicality concept $\mathbf{T}(C)$, for an arbitrary C. Here, we do not need to introduce a notion of global preference. The interpretation of any $\mathcal{ALC}$

concept C is defined compositionally from the interpretation of atomic concepts, and the preference relation $<_C$ associated to C is defined from C^I.

The notions of *satisfiability* in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$, *model* of an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base, and $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ *entailment* can be defined in a similar way as in fuzzy $\mathcal{ALC}$ (see Section 2). In particular, given an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base K, a fuzzy concept inclusion $\mathbf{T}(C) \sqsubseteq D\ \theta\ k$ (with $\theta \in \{\geq, \leq, >, <\}$ and $k \in [0,1]$) is *entailed from* K in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ (written $K \models_{\mathcal{ALC}^{\mathbf{F}}\mathbf{T}} \mathbf{T}(C) \sqsubseteq D\ \theta\ k$) if $\mathbf{T}(C) \sqsubseteq D\ \theta k$ is satisfied in all $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ models I of the knowledge base K. In the following, we will refer to the entailment of $\mathbf{T}(C) \sqsubseteq D \geq k$ as *k-entailment* and, as a special case, for $k = 1$, as *1-entailment*.

As an example of satisfiability, the fuzzy concept inclusion $\langle \mathbf{T}(C) \sqsubseteq D \geq k \rangle$ is satisfied in a fuzzy interpretation $I = \langle \Delta, \cdot^I \rangle$ if $\inf_{x \in \Delta}(\mathbf{T}(C))^I(x) \rhd D^I(x) \geq k$ holds, which can be evaluated based on the combination functions of some specific fuzzy logic.

As in the two-valued case, the typicality operator $\mathbf{T}$ introduced in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ is non-monotonic in the following sense: for a given knowledge base K, from the fact that $C \sqsubseteq D$ is 1-entailed from K, we cannot conclude that $\mathbf{T}(C) \sqsubseteq \mathbf{T}(D)$ is 1-entailed from K. Nevertheless, the logic $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ is monotonic, that is, for two $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge bases K and K', and a fuzzy axiom E, if $K \subseteq K'$, and $K \models_{\mathcal{ALC}^{\mathbf{F}}\mathbf{T}} E$ then $K' \models_{\mathcal{ALC}^{\mathbf{F}}\mathbf{T}} E$. $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ is a fuzzy relative of the monotonic logic $\mathcal{ALC} + \mathbf{T}$ [15].

Although, as we will see, the KLM postulates of a preferential consequence relation [6] can be reformulated and hold for $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$, this typicality extension of fuzzy $\mathcal{ALC}$ is rather weak. Similarly, in the two-valued case, the preferential extension of $\mathcal{ALC}$ with typicality, $\mathcal{ALC} + \mathbf{T}$ [15], and the rational extension of $\mathcal{ALC}$ with defeasible inclusions [14] do not allow to deal with *irrelevance*. From the fact that birds normally fly, one would like to be able to conclude that normally yellow birds fly, the color being irrelevant to flying.

In the two-valued case, this has led to the definition of non-monotonic defeasible Description Logics [16–18,50,51,21], which build on some closure construction (such as the rational closure [6] and the lexicographic closure [31] in KLM framework) or some notion of minimal entailment [52]. In the next section, we introduce a notion of weighted knowledge base and strengthen $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ by considering some different closure constructions, starting from the notion of coherent preferential interpretation introduced in [33], and we discuss their properties.

4. Weighted knowledge bases and closure constructions

To overcome the weakness of rational closure (as well as of preferential entailment), Lehmann introduced the lexicographic closure of a conditional knowledge base [31] which strengthens the rational closure by allowing further inferences. From the semantic point of view, in the propositional case, a preference relation is defined on the set of propositional interpretations, so that the interpretations satisfying conditionals with higher rank are preferred to the interpretations satisfying conditionals with lower rank and, in case of contradictory defaults with the same rank, interpretations satisfying more defaults with that rank are preferred. The ranks of conditionals used by the lexicographic closure construction are the ones computed by the rational closure construction [6], which capture specificity: the higher is the rank, the more specific is the default. In other cases, the ranks may be part of the knowledge base specification, such as for ranked knowledge bases in Brewka's framework of basic preference descriptions [53], or might be learned from empirical data, as we will see in the following.

In this section, we consider weighted (fuzzy) knowledge bases, where typicality inclusions are associated to weights, and develop a (semantic) closure construction to strengthen $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ entailment, which leads to some variants of the notion of fuzzy coherent multi-preferential model in [33]. The construction also relates to the definition of Kern-Isberner's c-representations [9,32] which also include penalty points for falsified conditionals, and to the algebraic semi-qualitative approach to conditionals by Weydert [54].

A *weighted* $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ *knowledge base* K, over a set $\mathcal{C} = \{C_1, \ldots, C_k\}$ of distinguished $\mathcal{ALC}$ concepts, is a tuple $\langle \mathcal{T}_f, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_k}, \mathcal{A}_f \rangle$, where $\mathcal{T}_f$ is a set of fuzzy $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ inclusion axioms, $\mathcal{A}_f$ is a set of fuzzy $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ assertions and, for each $C_i \in \mathcal{C}$, $\mathcal{T}_{C_i} = \{(d^i_h, w^i_h)\}$ is a (non-empty) set of weighted typicality inclusions $d^i_h = \mathbf{T}(C_i) \sqsubseteq D_{i,h}$ for C_i, indexed by h, where each inclusion d^i_h has weight w^i_h, a real number. As in [33], the typicality operator is assumed to occur only on the left hand side of a weighted typicality inclusion, and the *distinguished concepts* are those concepts C_i occurring on the l.h.s. of some typicality inclusion $\mathbf{T}(C_i) \sqsubseteq D$ in $\mathcal{T}_{C_i}$. Arbitrary $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ inclusions and assertions may belong to $\mathcal{T}_f$ and $\mathcal{A}_f$.

Example 3. Consider the weighted knowledge base $K = \langle \mathcal{T}_f, \mathcal{T}_{Bird}, \mathcal{T}_{Penguin}, \mathcal{T}_{Canary}, \mathcal{A}_f \rangle$, over the set of distinguished concepts $\mathcal{C} = \{Bird, Penguin, Canary\}$, and assume the combination functions as in Gödel fuzzy logic. The Tbox $\mathcal{T}_f$ contains the inclusions:

$$Yellow \sqcap Black \sqsubseteq \bot \geq 1 \quad Yellow \sqcap Red \sqsubseteq \bot \geq 1 \quad Black \sqcap Red \sqsubseteq \bot \geq 1$$

the ABox $\mathcal{A}_f$ contains the following assertions:

$$Red(reddy) \geq 1,\ \exists hasWings.Small(reddy) \geq 1,\ Fly(reddy) \geq 1$$

$$Black(opus) \geq 1,\ \exists hasWings.Long(opus) \geq 1,\ Fly(opus) \leq 0,$$

the weighted TBox $\mathcal{T}_{Bird}$ contains the following weighted defeasible inclusions:

(d_1) $\mathbf{T}(Bird) \sqsubseteq Fly$, *+20*

(d_2) $\mathbf{T}(Bird) \sqsubseteq \exists hasWings.\top$, *+50*

(d_3) $\mathbf{T}(Bird) \sqsubseteq \exists hasFeathering.\top$, *+50;*

and $\mathcal{T}_{Penguin}$ and $\mathcal{T}_{Canary}$ contain, respectively, the following inclusions:

(d_4) $\mathbf{T}(Penguin) \sqsubseteq Bird$, $+100$
(d_5) $\mathbf{T}(Penguin) \sqsubseteq Fly$, -70
(d_6) $\mathbf{T}(Penguin) \sqsubseteq Black$, $+50$;
(d_7) $\mathbf{T}(Canary) \sqsubseteq Bird$, $+100$
(d_8) $\mathbf{T}(Canary) \sqsubseteq Yellow$, $+30$
(d_9) $\mathbf{T}(Canary) \sqsubseteq Red$, $+20$

The intended meaning is that a bird normally has wings, has feathers and flies, but having wings and having feathers (both with weight 50) for a bird is more plausible than flying (weight 20), although flying is regarded as being plausible. For a penguin, flying is not plausible (inclusion (d_5) has negative weight -70), while being a bird and being black are very plausible properties of prototypical penguins, as (d_4) and (d_6) have positive weights (100 and 50, respectively). Similar considerations can be done for concept *Canary*.

Consider an interpretation I, satisfying both TBox and ABox axiams, in which Reddy is red, has small wings, has feathers and flies (suppose all with degree 1) and Opus has long wings, has feathers (with degree 1), is black with degree 0.8 and does not fly ($Fly^I(opus^I) = 0$). Considering the weights of defeasible inclusions, we might expect Reddy to be more typical than Opus as a bird, but less typical than Opus as a penguin in the interpretation I.

The fuzzy axioms in TBox $\mathcal{T}_f$ define strict constraints, e.g., for the second one, in any interpretation I, for each domain element x, the value of $(Red \sqcap Black)^I(x)$ must be 0 (and, hence, $Black^I(reddy^I) = 0$). Note also that, as ABox $\mathcal{A}_f$ contains the assertion $\exists hasWings.Small(reddy) \geq 1$, then $(\exists hasWings.Small)^I(reddy^I) = 1$ holds. Hence, there is a domain element $y \in \Delta$ such that $hasWings^I(reddy^I, y) = 1$ and $Small^I(y) = 1$. Thus, it follows that $(\exists hasWings.\top)^I(reddy^I) = 1$, and hence the assertion $\exists hasWings.\top(reddy) \geq 1$ is as well satisfied in I.

We define the semantics of weighted knowledge bases as the one above through a *semantic closure construction*, similar in spirit to Lehmann's lexicographic closure [31], but exploiting weights and based on multiple preferences. The construction allows a subset of the $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretations to be selected, the interpretations whose induced preference relations $<_{C_i}$, for the distinguished concepts C_i, faithfully represent the defeasible part of the knowledge base K.

Let $\mathcal{T}_{C_i} = \{(d^i_h, w^i_h)\}$ be the set of weighted typicality inclusions $d^i_h = \mathbf{T}(C_i) \sqsubseteq D_{i,h}$ associated to the distinguished concept C_i, and let $I = \langle \Delta, \cdot^I \rangle$ be a fuzzy $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation. In the two-valued case, we would associate to each domain element $x \in \Delta$ and each distinguished concept C_i, a weight $W_i(x)$ of x wrt C_i in I, by summing the weights of the defeasible inclusions satisfied by x. However, as I is a fuzzy interpretation, we do not only distinguish between the typicality inclusions satisfied or falsified by x; we also need to consider, for all inclusions $\mathbf{T}(C_i) \sqsubseteq D_{i,h} \in \mathcal{T}_{C_i}$, the degree of membership of x in $D_{i,h}$. Furthermore, in comparing the weight of domain elements with respect to $<_{C_i}$, we want to give higher preference to the domain elements having a membership degree in C_i greater than 0, with respect to those elements whose degree of membership in C_i is 0.

For each domain element $x \in \Delta$ and distinguished concept C_i, *the weight* $W_i(x)$ *of* x *wrt* C_i in the $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ is defined as follows:

$$W_i(x) = \begin{cases} \sum_h w^i_h \, D^I_{i,h}(x) & \text{if } C^I_i(x) > 0 \\ -\infty & \text{otherwise} \end{cases} \tag{3}$$

where $-\infty$ is added at the bottom of all real values.

The value of $W_i(x)$ is $-\infty$ when x is not a C-element (i.e., $C^I_i(x) = 0$). Otherwise, $C^I_i(x) > 0$ and the higher is the sum $W_i(x)$, the more typical is the element x relative to concept C_i. How much x satisfies a typicality property $\mathbf{T}(C_i) \sqsubseteq D_{i,h}$ depends on the value of $D^I_{i,h}(x) \in [0,1]$, which is weighted by w^i_h in the sum. In the two-valued case, $D^I_{i,h}(x) \in \{0,1\}$, and $W_i(x)$ is the sum of the weights of the typicality inclusions for C satisfied by x, if x is a C-element, and is $-\infty$, otherwise.

Example 4. Let us continue Example 3. In the $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation I, it holds that: $Fly^I(reddy^I) = (\exists has\_Wings.\top)^I(reddy^I) = (\exists has\_Feathering.\top)^I(reddy^I) = Red^I(reddy^I) = 1$, i.e., Reddy flies, has wings and feathers and is red (and hence $Black^I(reddy^I) = 0$). For Opus it holds that: $Fly^I(opus^I) = 0$, $Black^I(opus^I) = 0.8$ and $(\exists has\_Wings.\top)^I(opus^I) = (\exists has\_Feathering.\top)^I(opus^I) = 1$, i.e., Opus does not fly, is black with degree 0.8, it has wings and feathers.

Let us further assume that $Bird^I(reddy^I) = 1$ and $Bird^I(opus^I) = 0.8$. Considering the weights of the typicality inclusions for *Bird*:

$$W_{Bird}(reddy^I) = 20 + 50 + 50 = 120$$

$$W_{Bird}(opus^I) = 0 + 50 + 50 = 100$$

which suggests that Reddy should be more typical as a bird than Opus.

On the other hand, if we suppose $Penguin^I(reddy^I) = 0.2$ and $Penguin^I(opus^I) = 0.8$, we have:

$$W_{Penguin}(reddy) = 100 - 70 + 0 = 30$$

$$W_{Penguin}(opus) = 0.8 \times 100 - 0 + 0.8 \times 50 = 120.$$

This suggests that Reddy should be less typical as a penguin than Opus.

We have seen in Section 3 that each fuzzy interpretation I induces a preference relation for each concept and, in particular, it induces a preference $<_{C_i}$ for each distinguished concept C_i. We further require that, if $x <_{C_i} y$, then x must be more typical than y wrt C_i, that is, the weight $W_i(x)$ of x wrt C_i should be higher than the weight $W_i(y)$ of y wrt C_i (and x should satisfy more properties or

more plausible properties of typical C_i-elements with respect to y). This leads to the following definition of *faithful multi-preferential model* of a weighted a $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base.

Definition 4 *(Faithful (fuzzy) multi-preferential model of K).* Let $K = \langle \mathcal{T}_f, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_k}, \mathcal{A}_f \rangle$ be a weighted $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base over $\mathcal{C}$. A *faithful (fuzzy) multi-preferential model* (fm-model) of K is a fuzzy $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ s.t.:

- I satisfies the fuzzy inclusions in $\mathcal{T}_f$ and the fuzzy assertions in $\mathcal{A}_f$;
- for all $C_i \in \mathcal{C}$, the preference $<_{C_i}$ *is faithful to* $\mathcal{T}_{C_i}$, that is:

$$x <_{C_i} y \Rightarrow W_i(x) > W_i(y) \tag{4}$$

Example 5. Referring to Example 4 above, clearly, $reddy <_{Bird} opus$, as $Bird^I(reddy) = 1$ and $Bird^I(opus) = 0.8$, while $opus <_{Penguin} reddy$, as $Penguin^I(reddy) = 0.2$ and $Penguin^I(opus) = 0.8$. For the interpretation I to be faithful, it is necessary that the conditions $W_{Bird}(reddy) > W_{Bird}(opus)$ and $W_{Penguin}(opus) > W_{Penguin}(reddy)$ hold with respect to interpretation I. This is true, as we have seen in Example 4. On the contrary, if we had $Penguin^I(reddy) = 0.9$, the interpretation I would not be faithful (as it assigns to *reddy* a membership degree in concept *Penguin* higher than the one for *opus*).

Let us now consider two alternative closure constructions, by introducing the notions of *coherent* and of φ-*coherent* models.

The notion of *coherent (fuzzy) multi-preferential model* of K, can be defined as in Definition 4 above, but replacing the faithfulness condition (4), with the following stronger coherence condition:

$$x <_{C_i} y \quad \text{iff} \quad W_i(x) > W_i(y) \tag{5}$$

This is a reformulation of the notion of coherent (fuzzy) multi-preferential model from [33], but here we do not restrict to a crisp interpretation of typicality concepts $\mathbf{T}(C)$.

The weaker notion of faithfulness determines a larger class of fuzzy multi-preferential models of a weighted knowledge base, compared to the class of coherent models. As we will see in Section 6, this also allows a larger class of monotone non-decreasing activation functions in neural network models to be captured.

The notion of φ-*coherence* of a fuzzy interpretation I wrt a KB, first introduced in [35], exploits a function φ from $\mathbb{R}$ to the interval $[0, 1]$, i.e., $\varphi : \mathbb{R} \to [0, 1]$. We actually allow for possibly different functions $\varphi_i : \mathbb{R} \to [0, 1]$, one for each concept $C_i \in \mathcal{C}$. As we will see, φ or the φ_i are intended to represent the activation function(s) for units in a neural network $\mathcal{N}$.

Definition 5 *(φ-coherence).* Let $K = \langle \mathcal{T}_f, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_k}, \mathcal{A}_f \rangle$ be a weighted $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base, and φ a collection of functions $\varphi_i : \mathbb{R} \to [0, 1]$, for $i = 1, \ldots, k$. A fuzzy $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ is φ-*coherent* if, for all concepts $C_i \in \mathcal{C}$ and $x \in \Delta$,

$$C_i^I(x) = \varphi_i\Big(\sum_h w_h^i \, D_{i,h}^I(x)\Big) \tag{6}$$

where $\mathcal{T}_{C_i} = \{(\mathbf{T}(C_i) \sqsubseteq D_{i,h}, w_h^i)\}$ is the set of weighted conditionals for C_i.

Observe that, for all x such that $C_i(x) > 0$, condition (6) above corresponds to condition $C_i^I(x) = \varphi_i(W_i(x))$, for all distinguished concepts $C_i \in \mathcal{C}$. While in coherent and faithful models the notion of weight $W_i(x)$ considers, as a special case, the case $C_i(x) = 0$, condition (6) imposes the same constraint to all domain elements x.

Coherent (resp., φ-*coherent*) *multi-preferential models* of a knowledge base K, can be defined similarly to faithful models in Definition 4. We provide explicitly the definition of φ-coherent model of K.

Definition 6 *(φ-coherent (fuzzy) multi-preferential model of K).* Let $K = \langle \mathcal{T}_f, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_k}, \mathcal{A}_f \rangle$ be a weighted $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base over $\mathcal{C}$. A φ-*coherent (fuzzy) multi-preferential model* (or, simply, φ-coherent model) of K is a fuzzy $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ s.t.:

- I satisfies the fuzzy inclusions in $\mathcal{T}_f$ and the fuzzy assertions in $\mathcal{A}_f$;
- for all the distinguished concepts $C_i \in \mathcal{C}$, for all $x \in \Delta$,

$$C_i^I(x) = \varphi_i\Big(\sum_h w_h^i \, D_{i,h}^I(x)\Big)$$

where $\mathcal{T}_{C_i} = \{(\mathbf{T}(C_i) \sqsubseteq D_{i,h}, w_h^i)\}$ is the set of all the weighted conditionals for C_i.

The following proposition establishes the relationships between φ-coherent, faithful and coherent fuzzy multi-preferential models of a weighted conditional knowledge base K.

Proposition 1. *Let K be a weighted conditional $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base and let $\varphi_i : \mathbb{R} \to [0, 1]$, for all $i = 1, \ldots, k$. The following statements hold:*

(1) Any coherent model of K is a faithful model of K;
(2) If the φ_i are monotonically non-decreasing *functions, a φ-coherent multi-preferential model I of K is also a faithful model of K;*
(3) If the φ_i are monotonically increasing *functions, a φ-coherent multi-preferential model I of K is also a coherent-model of K.*

Proof. Item (1) directly follows from the definition, as the coherence condition (5) is stronger than the faithfulness condition (4).

For item (2), let us assume that the φ_i are monotonically non-decreasing functions and that $I = \langle \Delta, \cdot^I \rangle$ is a φ-coherent fuzzy multi-preferential model of K. In particular, for all distinguished concepts C_i and $z \in \Delta$, s.t. $C_i^I(z) > 0$, $C_i^I(z) = \varphi_i(W_i(z))$. To prove condition (4), i.e., that $x <_{C_i} y \Rightarrow W_i(x) > W_i(y)$ holds, let us assume that, for some $x, y \in \Delta$, $x <_{C_i} y$ holds, i.e., $C_i^I(x) > C_i^I(y)$. This also implies $C_i^I(x) > 0$. If $C_i^I(y) = 0$, $W_i(y) = -\infty$, and the thesis follows. If $C_i^I(y) > 0$, both the equalities $C_i^I(x) = \varphi_i(W_i(x))$ and $C_i^I(y) = \varphi_i(W_i(y))$ hold. Suppose that $W_i(x) > W_i(y)$ does not hold, i.e., that $W_i(x) \leq W_i(y)$. As φ_i is monotonically non-decreasing, $\varphi_i(W_i(x)) \leq \varphi_i(W_i(y))$. Hence, by the equalities above, $C_i^I(x) \leq C_i^I(y)$, which contradicts the assumption that $C_i^I(x) > C_i^I(y)$.

For item (3), assume that the φ_i are monotonically increasing functions and that $I = \langle \Delta, \cdot^I \rangle$ is a φ-coherent multi-preferential model of K. Then, equality (6) holds. In particular, for all distinguished concept C_i and $z \in \Delta$, s.t. $C_i^I(z) > 0$, $C_i^I(z) = \varphi_i(W_i(z))$.

We have to prove that condition (5) holds, i.e., that $x <_{C_i} y$ iff $W_i(x) > W_i(y)$. The "only if" direction holds with the same proof as for item (2), as φ_i is as well monotonically non-decreasing. To prove the "if" direction, assume that $W_i(x) > W_i(y)$ holds, for some $x, y \in \Delta$. If $W_i(y) = -\infty$, it must be that $C_i^I(y) = 0$ and $C_i^I(x) > 0$, and hence $C^I(x) > C^I(y)$ follows. If $W_i(y) \neq -\infty$, $C_i^I(y) > 0$ and $C_i^I(x) > 0$. From $W_i(x) > W_i(y)$, as φ_i is monotonically increasing, $\varphi_i(W_i(x)) > \varphi_i(W_i(y))$. Hence, $C_i^I(x) > C_i^I(y)$. □

The notions of *faithful/coherent/φ-coherent multi-preferential entailment* from a weighted $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base K can be defined as expected.

Definition 7 *(Faithful/coherent/φ-coherent entailment).* A fuzzy axiom E is *faithfully entailed* (resp., *coherently/φ-coherently entailed*) from a fuzzy weighted knowledge base K (for short $K \models_{fm/cm/\varphi} E$) if, for all faithful models (resp., coherent/φ-coherent-models) $I = \langle \Delta, \cdot^I \rangle$ of K, I satisfies E.

As usual in preferential semantics, a stronger notion of entailment can be obtained by restricting to a specific subset of models, namely, to *canonical models*, which are large enough to contain all the relevant domain elements. More precisely, in the two-valued case, a canonical model contains a domain element for each possible valuation of concepts which is present in some model of K [18,23]. The notion of canonical model can be extended to the many-valued case.

Definition 8. Given a weighted knowledge base $K = \langle \mathcal{T}_f, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_k}, \mathcal{A}_f \rangle$ a faithful/coherent/φ-coherent $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ model $I = \langle \Delta, \cdot^I \rangle$ of K is *canonical* if, for each faithful/coherent/φ-coherent model $I' = \langle \Delta', \cdot' \rangle$ of K, and for each $x \in \Delta'$, there is an element $y \in \Delta$ such that $A^I(y) = A^{I'}(x)$, for all concept names A occurring in K.

That is, a canonical faithful/coherent/φ-coherent model for K contains a domain element y corresponding to each domain element x in any faithful/ coherent/φ-coherent interpretation I' of K, and y has the same membership degree as x in all named concepts A occurring in K. Note that, as in the two-valued case for defeasible $\mathcal{ALC}$ [22] (and similarly for $\mathcal{ALC}$ with typicality [18]), also in the many-valued case two $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ models of the knowledge base can be combined by taking the disjoint union of their domains, to construct a new (larger) model of the KB (and the proof is similar to the one in the two-valued case [22]). This property guarantees that, if a faithful/coherent/φ-coherent $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ model of a knowledge base K exists, a canonical faithful/coherent/φ-coherent $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ model of K also exists.

A notion of *canonical faithful/coherent/φ-coherent entailment* can be defined. In the following, we assume that axiom E only contains concept names occurring in the knowledge base K.

Definition 9 *(Canonical entailment).* Given a weighted $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base K, a fuzzy axiom E is *canonically entailed* from K in the faithful/ coherent/φ-coherent semantics if, for all *canonical faithful/coherent/ φ-coherent models* $I = \langle \Delta, \cdot^I \rangle$ of K, I satisfies E.

In Sections 5 and 6 we study the KLM properties of $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ and its relationships with MLPs, under the different closure constructions.

5. The KLM properties of $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ and its closures

In this section we investigate whether the KLM postulates of a preferential consequence relation [4,6] are satisfied by entailment in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ as well as in the coherent and faithful semantics.

The satisfiability of KLM postulates of rational or preferential consequence relations [4,6] has been studied for $\mathcal{ALC}$ with defeasible inclusions and typicality inclusions in the two-valued case [14,15,55]. The KLM postulates of a preferential consequence relation (namely, Reflexivity, Left Logical Equivalence, Right Weakening, And, Or, Cautious Monotonicity) can be reformulated for $\mathcal{ALC}$ with typicality, by considering that a typicality inclusion $\mathbf{T}(C) \sqsubseteq D$ stands for a conditional $C \mathrel{|\!\sim} D$ in KLM preferential logics, by the following properties, expressed as inference rules:

$(REFL)$ $\mathbf{T}(C) \sqsubseteq C$
(LLE) If $\models A \equiv B$ and $\mathbf{T}(A) \sqsubseteq C$, then $\mathbf{T}(B) \sqsubseteq C$
(RW) If $\models C \sqsubseteq D$ and $\mathbf{T}(A) \sqsubseteq C$, then $\mathbf{T}(A) \sqsubseteq D$
(AND) If $\mathbf{T}(A) \sqsubseteq C$ and $\mathbf{T}(A) \sqsubseteq D$, then $\mathbf{T}(A) \sqsubseteq C \sqcap D$
(OR) If $\mathbf{T}(A) \sqsubseteq C$ and $\mathbf{T}(B) \sqsubseteq C$, then $\mathbf{T}(A \sqcup B) \sqsubseteq C$
(CM) If $\mathbf{T}(A) \sqsubseteq D$ and $\mathbf{T}(A) \sqsubseteq C$, then $\mathbf{T}(A \sqcap D) \sqsubseteq C$

where $\models A \equiv B$ is interpreted as equivalence of concepts A and B in the underlying description logic $\mathcal{ALC}$ (i.e., $A^I = B^I$ in all $\mathcal{ALC}$ interpretations I), while $\models C \sqsubseteq D$ is interpreted as validity of the inclusion $C \sqsubseteq D$ in $\mathcal{ALC}$ (i.e., $A^I \subseteq B^I$ for all $\mathcal{ALC}$ interpretations I).

The interpretation of the postulates is that, given a knowledge base K in $\mathcal{ALC}$ extended with typicality inclusions, the logical consequences of K satisfy the properties above. For instance, reflexivity $(REFL)$ requires that $\mathbf{T}(C) \sqsubseteq C$ is a logical consequence of K; for Left Logical Equivalence (LLE), if A and B are logically equivalent in $\mathcal{ALC}$ and $\mathbf{T}(A) \sqsubseteq C$ is a logical consequence of K, then $\mathbf{T}(B) \sqsubseteq C$ must be as well a logical consequence of K; and so on.

In the following we reformulate these postulates for the fuzzy case and, specifically for $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$. We reinterpret $\models C \sqsubseteq D$ as the requirement that the fuzzy inclusion $C \sqsubseteq D \geq 1$ is valid in fuzzy $\mathcal{ALC}$ (that is, $C \sqsubseteq D \geq 1$ is satisfied in all fuzzy $\mathcal{ALC}$ interpretations), and $\models A \equiv B$ as the requirement that the fuzzy inclusions $A \sqsubseteq B \geq 1$ and $B \sqsubseteq A \geq 1$ are valid in fuzzy $\mathcal{ALC}$. Interpreting inclusions of the form $\mathbf{T}(A) \sqsubseteq C$ as fuzzy inclusions $\mathbf{T}(A) \sqsubseteq C \geq 1$, we reformulate the KLM postulates for 1-entailment:

$(REFL')$ $\mathbf{T}(C) \sqsubseteq C \geq 1$
(LLE') If $\models A \equiv B$ and $\mathbf{T}(A) \sqsubseteq C \geq 1$, then $\mathbf{T}(B) \sqsubseteq C \geq 1$
(RW') If $\models C \sqsubseteq D$ and $\mathbf{T}(A) \sqsubseteq C \geq 1$, then $\mathbf{T}(A) \sqsubseteq D \geq 1$
(AND') If $\mathbf{T}(A) \sqsubseteq C \geq 1$ and $\mathbf{T}(A) \sqsubseteq D \geq 1$,
then $\mathbf{T}(A) \sqsubseteq C \sqcap D \geq 1$
(OR') If $\mathbf{T}(A) \sqsubseteq C \geq 1$ and $\mathbf{T}(B) \sqsubseteq C \geq 1$,
then $\mathbf{T}(A \sqcup B) \sqsubseteq C \geq 1$
(CM') If $\mathbf{T}(A) \sqsubseteq D \geq 1$ and $\mathbf{T}(A) \sqsubseteq C \geq 1$,
then $\mathbf{T}(A \sqcap D) \sqsubseteq C \geq 1$

As an example, the meaning of right weakening (RW') is that, if it holds that $\models C \sqsubseteq D$ (i.e., $C \sqsubseteq D \geq 1$ is valid in fuzzy $\mathcal{ALC}$), and $\mathbf{T}(A) \sqsubseteq C \geq 1$ is entailed from a weighted knowledge base K, then $\mathbf{T}(A) \sqsubseteq D \geq 1$ is also entailed by K.

To prove that all the postulates above hold for the choice of combination functions as in Gödel logic, we prove that each postulate is satisfied in any $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation. For instance, for (RW') this means that, if it holds that $\models C \sqsubseteq D$ (i.e., $C \sqsubseteq D \geq 1$ is valid in fuzzy $\mathcal{ALC}$), then in any $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation I, if $\mathbf{T}(A) \sqsubseteq C \geq 1$ is satisfied in I, then $\mathbf{T}(A) \sqsubseteq D \geq 1$ is also satisfied in I.

Proposition 2. *Under the choice of combination functions as in Gödel logic, any $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ satisfies the postulates $(REFL')$, (LLE'), (RW'), (AND'), (OR') and (CM').*

The proof of Proposition 2 can be found in the Appendix. As a simple consequence, the following corollary states that, for the choice of combination functions as in Gödel logic, 1-entailment in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ satisfies the KLM postulates $(REFL')$, (LLE'), (RW'), (AND'), (OR') and (CM').

Corollary 1. *For the choice of combination functions as in Gödel logic, 1-entailment in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ satisfies the KLM postulates $(REFL')$, (LLE'), (RW'), (AND'), (OR') and (CM').*

Proof (Sketch). Consider, for instance, postulate (AND'). Assume $\mathbf{T}(A) \sqsubseteq C \geq 1$ and $\mathbf{T}(A) \sqsubseteq D \geq 1$ are entailed from a knowledge base K in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$. Then they are satisfied in all $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ models I of K. Hence, by Proposition 2, $\mathbf{T}(A) \sqsubseteq C \sqcap D \geq 1$ is also satisfied in all the models I of K, i.e., $\mathbf{T}(A) \sqsubseteq C \sqcap D \geq 1$ is entailed by K. The proof of all other properties is similar. □

Corollary 1 tells us that, for the choice of combination functions as in Gödel logic, 1-entailment in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ satisfies the properties of a preferential consequence relation. Observe that this result does not depend on the choice of the negation combination function as negation does not occur in the postulates we have considered; in particular, the result holds as well for Gödel logic with standard involutive negation. On the other hand, 1-entailment in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ does not satisfy the Rational Monotonicity postulate, so it does not satisfy all postulates of a rational consequence relation. Let us reformulate the property of Rational Monotonicity in the fuzzy case as follows:
(RM') If $\mathbf{T}(A) \sqsubseteq C \geq 1$ and not $\mathbf{T}(A) \sqsubseteq \neg B \geq 1$, then $\mathbf{T}(A \sqcap B) \sqsubseteq C \geq 1$

Proposition 3. *For the choice of combination functions as in Gödel logic, (RM') does not hold in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ (and the same for Gödel logic with standard involutive negation).*

The proof of the proposition in the Appendix provides a counterexample to Rational Monotonicity for a knowledge base without weighted inclusions.

The postulates for 1-entailment considered above may be violated by other choices of combination functions. For instance, the choice of combination functions as in Product logic or as in Łukasiewicz logics fails to satisfy both postulates (AND') and (OR'). The postulates ($REFL'$), (LLE'), (RW'), (AND'), (OR') and (CM') can as well be formulated for *k-entailment*, by replacing the occurrences of typicality inclusions $\mathbf{T}(A) \sqsubseteq C \geq 1$ with $\mathbf{T}(A) \sqsubseteq C \geq k$. For combination functions as in Gödel logic, all the postulates for k-entailment hold (with a proof similar to the one for 1-entailment), except for Cautious Monotonicity (CM), which does not hold.

For faithful, coherent and φ-coherent entailment, the next corollary also follows from Proposition 2 as a simple consequence, by observing that all faithful, coherent and φ-coherent models of a knowledge base K are $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ models of K.

Corollary 2. *For the choice of combination functions as in Gödel logic, faithful, coherent and φ-coherent entailment in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ satisfy postulates ($REFL'$), (LLE'), (RW'), (AND'), (OR') and (CM') of 1-entailment.*

The results above improve over the previous results in [34], which have been proven for a crisp interpretation of the typicality concept. When the interpretation of $\mathbf{T}(C)$ is either 0 or 1, 1-entailment in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ fails to satisfy the Reflexivity postulate ($REFL'$).

Some further properties of typicality can be obtained by reformulating for the fuzzy case the semantic properties of $\mathcal{ALC}+T$ in [15]. We name the properties $(f_T-1),\dots,(f_T-5)$ after [15]:

(f_T-1) $\quad \mathbf{T}(C) \sqsubseteq C \geq 1$
(f_T-2) if $\mathbf{T}(C) \equiv \bot$, then $C \equiv \bot$
(f_T-3) If $\mathbf{T}(A) \sqsubseteq D \geq 1$, then $\mathbf{T}(A) \equiv \mathbf{T}(A \sqcap D)$
(f_T-4) $\mathbf{T}(A \sqcup B) \sqsubseteq \mathbf{T}(A) \sqcup \mathbf{T}(B) \geq 1$
(f_T-5) $\mathbf{T}(A) \sqcap \mathbf{T}(B) \sqsubseteq \mathbf{T}(A \sqcup B) \geq 1$,

where, for two $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ concepts, $C \equiv D$, stands for $(C \sqsubseteq D \geq 1) \sqcap (D \sqsubseteq C \geq 1)$.

Note that (f_T-1) is ($REFL'$); (f_T-2) is a consequence of well-foundedness of the preference relations; (f_T-3) implies (CM); and (f_T-4) is a reformulation of (OR). It can be proven that properties $(f_T-1),\dots,(f_T-5)$ are satisfied in all $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretations.

Proposition 4. *Under the choice of combination functions as in Gödel logic, any $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation satisfies the postulates $(f_T-1),\dots,(f_T-5)$.*

The proof is similar to the proof of Proposition 2. To conclude this section let us informally describe how fuzzy multi-preferential entailment deals with irrelevance and avoids inheritance blocking, properties which have been considered as desiderata for preferential logics of defeasible reasoning [54,32].

Concerning "irrelevance", let us consider again previous Example 3: if typical birds fly, we would like to conclude that typical yellow birds also fly, as the property of being yellow is irrelevant with respect to flying. Observe, that in Example 4, we can conclude that Reddy is more typical than Opus as a bird ($reddy <_{Bird} opus$), as Opus does not fly, while Reddy flies. The relative typicality of Reddy and Opus wrt *Bird* does not depend on their color (the weighted TBox $\mathcal{T}_{Bird}$ does not refer to a color) and we would obtain the same relative preferences if Reddy were yellow rather than red.

The fuzzy multi-preferential entailment is not subject to the problem called by Pearl the "blockage of property inheritance" problem [5], and by Benferhat et al. the "drowning problem" [7]. This problem affects the rational closure and system Z [5], as well as the rational closure refinements. Roughly speaking, the problem is that property inheritance from classes to subclasses is not guaranteed. If a subclass is exceptional with respect to a superclass for a given property, it does not inherit from that superclass any other property. For instance, referring to the typicality inclusions in Example 4, in the rational closure, typical penguins would not inherit the property of typical birds of having wings, being exceptional to birds concerning flying. On the contrary, in fuzzy multi-preferential models, considering again Example 4, the degree of membership of a domain element x in concept *Bird*, i.e., $Bird^I(x)$, is used to determine the weight of x with respect to *Penguin*. As the weight of typicality inclusion (d_4) is positive, the higher is the value of $Bird^I(x)$, the higher the value of $W_{Penguin}(x)$. Hence, provided the relevant properties of penguins (such as non-flying) remain unaltered, the more typical is x as a bird, the more typical is x as a Penguin. Notice also that the weight $W_{Bird}(x)$ of a domain element x with respect to *Bird* is related to the interpretation of *Bird* in I by the faithfulness condition or by a coherence condition (depending on the semantic construction).

6. A multi-preferential fuzzy interpretation of multilayer perceptrons

In this section, we first shortly introduce multilayer perceptrons. Then we develop a fuzzy multi-preferential interpretation of a neural network, which can be used for post-hoc explanation, based on a model checking approach.

6.1. Multilayer perceptrons

Let us first recall from [36] the model of a *neuron* as an information-processing unit in an artificial neural network. The basic elements are the following:

- a set of *synapses* or *connecting links*, each one characterized by a *weight*; we let x_j be the signal at the input of synapse j connected to neuron k, and w_{kj} the related synaptic weight;

- the adder for summing the input signals to the neuron, weighted by the respective synapses weights: $\sum_{j=1}^{n} w_{kj} x_j$;
- an *activation function* for limiting the amplitude of the output of the neuron (typically, to the interval $[0, 1]$ or $[-1, +1]$).

The logistic, threshold and hyperbolic-tangent functions are examples of activation functions. A neuron k can be described by the following pair of equations: $u_k = \sum_{j=1}^{n} w_{kj} x_j$, and $y_k = \varphi(u_k + b_k)$, where $x_1, \ldots, x_n$ are the input signals and $w_{k1}, \ldots, w_{kn}$ are the weights of neuron k; b_k is the bias, φ the activation function, and y_k is the output signal of neuron k. By adding a new synapse with input $x_0 = +1$ and synaptic weight $w_{k0} = b_k$, one can write:

$$u_k = \sum_{j=0}^{n} w_{kj} x_j \qquad y_k = \varphi(u_k), \tag{7}$$

where u_k is called the *induced local field* of the neuron.

A neural network can then be seen as "a directed graph consisting of nodes with interconnecting synaptic and activation links" [36]: nodes in the graph are the neurons (the processing units) and the weight w_{ij} on the edge from node j to node i represents "the strength of the connection [..] by which unit j transmits information to unit i" [56]. Source nodes (i.e., nodes without incoming edges) produce the input signals to the graph. Neural network models are classified by their synaptic connection topology. In a *feedforward* network the architectural graph is acyclic, while in a *recurrent* network it contains cycles. In a feedforward network neurons are organized in layers. In a *single-layer* network there is an input layer of source nodes and an output layer of computation nodes. In a *multilayer feedforward* network there are one or more hidden layers, whose computation nodes are called *hidden neurons* (or hidden units). The source nodes in the input layer supply the activation pattern (*input vector*) providing the input signals for the first layer computation units. In turn, the output signals of first layer computation units provide the input signals for the second layer computation units, and so on, up to the final output layer of the network, which provides the overall response of the network to the activation pattern. In a recurrent network at least one feedback exists, so that "the output of a node in the system influences in part the input applied to that particular element" [36]. In the following, we do not put restrictions on the topology the network, even though in Section 8 we only report experiments on feedforward networks.

"A major task for a neural network is to learn a model of the world" [36]. In supervised learning, a set of input/output pairs, input signals and corresponding desired response, referred as training data, or training sample, is used to train the network to learn. In particular, the network learns by changing the synaptic weights, through the exposition to the training samples. After the training phase, in the generalization phase, the network is tested with data not seen before. "Thus the neural network not only provides the implicit model of the environment in which it is embedded, but also performs the information-processing function of interest" [36]. In the next section, we aim to make this model explicit as a multi-preferential model.

6.2. A multi-preferential interpretation of MLPs and property verification by model checking

In this section, we show that a fuzzy multi-preferential interpretation (an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation) can be associated to a multilayer network $\mathcal{N}$, based on the activity of the network over a set of input stimuli Δ. Fuzzy and typicality properties of the network can then be verified by model checking over such an interpretation, and used for post-hoc explanation.

Assume that the network $\mathcal{N}$ has been trained and the synaptic weights w_{kj} have been learned. We associate a concept name $C_i \in N_C$ to the units i of interest in $\mathcal{N}$, which may include input, output or hidden units. They are the units we are interested in, for property verification.

We construct a multi-preferential interpretation over a (finite) *domain* Δ of input stimuli; for instance, the input vectors considered so far, for training and generalization, or a subset of it (e.g., the test set). In case the network is not feedforward, we assume that, for each input vector v in Δ, the network reaches a stationary state [36], in which $y_k(v)$ is the activity level of unit k, and equations (7) hold, for all units k. We also assume the activation of units to be in the interval $[0, 1]$.

Let Δ be a finite (non-empty) set of input vectors. We can associate to $\mathcal{N}$ a fuzzy multi-preferential interpretation over Δ, in the boolean fragment of $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$, which contains no roles (i.e., $N_R = \emptyset$) and no individual names (i.e., $N_I = \emptyset$). We refer to the definition of an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation (Definition 3).

Definition 10. The *fuzzy multi-preferential interpretation of a network* $\mathcal{N}$ over a non-empty domain Δ, is the $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I^{\Delta}_{\mathcal{N}} = \langle \Delta, \cdot^I \rangle$ where: the interpretation function $\cdot^I$ satisfies the condition that, for all concept names $C_k \in N_C$ and for all $x \in \Delta$,

$$C_k^I(x) = y_k(x)$$

where $y_k(x)$ is the output signal of neuron k, for input vector x.

As we have seen in section 3, the $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I^{\Delta}_{\mathcal{N}}$ is a multi-preferential interpretation, as the fuzzy interpretation of concepts induces a preference relation associated to each concept. Here, the preferences associated with concepts are those associated with units, and based on the unit activations for the different inputs. More precisely, the preference relation $<_{C_k}$ associated to concept C_k (and to unit k), induced by the interpretation $I^{\Delta}_{\mathcal{N}}$, is determined by the activity of unit k as follows: for $x, x' \in \Delta$,

$$x <_{C_k} x' \text{ iff } y_k(x) > y_k(x'). \tag{8}$$

This allows the set of typical instances of a concept C_k to be identified according to the definition of typicality concepts in Equation (2), by selecting the input stimuli $x \in \Delta$ with the highest activity value $y_k(x)$.

This model provides a multi-preferential interpretation of the network $\mathcal{N}$, based on the input stimuli considered in Δ. For instance, in case the neural network is used for categorization and an output neuron is associated to each category, each concept C_h associated to an output unit h corresponds to a learned category. If $C_h \in N_C$, the preference relation $<_{C_h}$ determines the relative typicality of input stimuli with respect to category C_i. This allows to verify typicality properties concerning categories, such as $\mathbf{T}(C_h) \sqsubseteq D \geq \alpha$ (where D is a boolean concept built from the named concepts in N_C), by *model checking* on the model $I^{\Delta}_{\mathcal{N}}$. According to the semantics of typicality concepts, this would require to identify typical C_h-elements and checking whether they are instances of concept D with a degree greater than α.

For instance, in Section 8 we consider some example neural networks, trained to recognize emotions (*surprise, fear, happiness, anger*) in images of human faces. In that case, we will be interested in understanding which properties have been learned by the network, concerning the relationships between some learned category (e.g., happiness) and some specific features of the image (in the example, facial muscle contractions). To this purpose, we will check properties such as, for instance, $\mathbf{T}(happiness) \sqsubseteq au12 \geq \alpha$ (where $au12$ is the activation of the lip corner puller muscle used for smiling), to verify whether the images recognized by the network as typical instances of happy faces correspond to smiling faces, to some degree.

In general, fuzzy typicality inclusions of the form $\mathbf{T}(C) \sqsubseteq D\theta\alpha$, with C and D boolean concepts, can be verified on the model $I^{\Delta}_{\mathcal{N}}$ in polynomial time in the size of the model $I^{\Delta}_{\mathcal{N}}$ and in the size of the formula.

Consider, for instance, the verification of $\mathbf{T}(C) \sqsubseteq D \geq \alpha$ under the choice of combination functions as in Gödel logic. The verification amounts to check that $\inf_{x\in\Delta} \mathbf{T}(C)^{I^{\Delta}_{\mathcal{N}}}(x) \rhd D^{I^{\Delta}_{\mathcal{N}}}(x) \geq \alpha$, i.e., that for all $x \in \Delta$, $\mathbf{T}(C)^{I^{\Delta}_{\mathcal{N}}}(x) \rhd D^{I^{\Delta}_{\mathcal{N}}}(x) \geq \alpha$ holds. When $\mathbf{T}(C)^{I^{\Delta}_{\mathcal{N}}}(x) = 0$, that is, x is not a typical C-element, $\mathbf{T}(C)^{I^{\Delta}_{\mathcal{N}}}(x) \rhd D^{I^{\Delta}_{\mathcal{N}}}(x) \geq \alpha$ holds trivially.

The identification of typical C-elements in Δ requires: computing the values of $C^{I^{\Delta}_{\mathcal{N}}}(x)$, for all input stimuli $x \in \Delta$ and selecting those y such that the value $C^{I^{\Delta}_{\mathcal{N}}}(y)$ is maximal among the values of $C^{I^{\Delta}_{\mathcal{N}}}(x)$, for all $x \in \Delta$. Then, for all typical C-elements x, one has to verify that $C^{I^{\Delta}_{\mathcal{N}}}(x) \rhd D^{I^{\Delta}_{\mathcal{N}}}(x) \geq \alpha$ holds, which requires to verify that $C^{I^{\Delta}_{\mathcal{N}}}(x) \leq D^{I^{\Delta}_{\mathcal{N}}}(x)$ or $D^{I^{\Delta}_{\mathcal{N}}}(x) \geq \alpha$ hold. In turn, this requires the value of $D^{I^{\Delta}_{\mathcal{N}}}(x)$ to be computed, for all typical C-elements x.

Overall, the verification requires a polynomial number of steps in the size of the model $I^{\Delta}_{\mathcal{N}}$ and in the size of the formula $\mathbf{T}(C) \sqsubseteq D$. Note that, as C and D only contain a polynomial number of subformulas, the values of $C(x)$ and $D(x)$, for some $x \in \Delta$ can be computed in polynomial time. But the evaluation has to be repeated for all elements $x \in \Delta$, and the domain Δ can be very large.

It is easy to see that similar polynomial algorithms can be developed for the verification of inclusions of the form $\mathbf{T}(C) \sqsubseteq D \leq \alpha$ (which require the verification that there is an element $x \in \Delta$, such that $\mathbf{T}(C)^{I^{\Delta}_{\mathcal{N}}}(x) \rhd D^{I^{\Delta}_{\mathcal{N}}}(x) \leq \alpha$ holds), and for the verification of strict inclusions $C \sqsubseteq D\theta\alpha$, according to the choice of the t-norm, s-norm, negation and implication functions. In general, inclusion axioms of the form $C \sqsubseteq D\theta\alpha$ may be considered, where C and D contain (non-nested) occurrences of the typicality operator $\mathbf{T}$.

Proposition 5. *Whether an axiom $C \sqsubseteq D\theta\alpha$ is satisfied in a multi-preferential interpretation $I^{\Delta}_{\mathcal{N}}$, can be decided in polynomial time in the size of $I^{\Delta}_{\mathcal{N}}$ and in the size of $C \sqsubseteq D$.*

The size of model $I^{\Delta}_{\mathcal{N}}$ is $O(|N_C| \times |\Delta|)$: it depends on the number $|N_C|$ of the units in the network that we are considering for property verification, and on the size of the set of input stimuli Δ, which can be very large. Observe, however, that to prove an inclusion $\mathbf{T}(C) \sqsubseteq D\theta\alpha$ (or $C \sqsubseteq D\theta\alpha$) we do not need to consider and build the entire model $I^{\Delta}_{\mathcal{N}}$, but it is sufficient to consider the restriction of the model over the concept names in C and in D, as only the interpretation of the subconcepts occurring in C and in D are needed in the verification.

In Section 8 we report results of the model checking approach in the verification of typicality properties of a multilayer networks, trained to recognize emotions from input features, exploiting a Datalog encoding of the model checking problem developed in [40] for the finite-valued case.

6.3. Multilayer perceptrons as weighted conditional knowledge bases

Another possible approach for reasoning about the properties of a neural network consists in exploiting entailment in the defeasible logic, based on the idea that the neural network $\mathcal{N}$ can be regarded as a defeasible knowledge base $K_{\mathcal{N}}$. In this section, we explore this approach.

Let us introduce a concept name $C_i \in N_C$ for each unit i in the network $\mathcal{N}$ and let $C = \{C_1, \ldots, C_n\}$ be a subset of N_C, namely the set of all concept names $C_i \in N_C$ such that there is at least a synaptic connection between some unit j and unit i. Given the *fuzzy multi-preferential interpretation* $I^{\Delta}_{\mathcal{N}} = \langle \Delta, \cdot^I \rangle$ as defined in Section 6.2, we aim at proving that $I^{\Delta}_{\mathcal{N}}$ is indeed a model of the neural network $\mathcal{N}$ in a logical sense.

A weighted conditional knowledge base $K^{\mathcal{N}}$ can be defined from the neural network $\mathcal{N}$ as follows. For each unit k with incoming edges, we consider all the units $j_1, \ldots, j_m$ whose output signals are the input signals of unit k, with synaptic weights $w_{k,j_1}, \ldots, w_{k,j_m}$. Let C_k be the concept name associated to unit k and $C_{j_1}, \ldots, C_{j_m}$ be the concept names associated to units $j_1, \ldots, j_m$, respectively. We define for each concept $C_k \in C$ a set $\mathcal{T}_{C_k}$ of typicality inclusions, with their associated weights, as follows:

$\mathbf{T}(C_k) \sqsubseteq C_{j_1}$ with w_{k,j_1},
…,
$\mathbf{T}(C_k) \sqsubseteq C_{j_m}$ with w_{k,j_m}

The knowledge base constructed from network $\mathcal{N}$ is defined, from the above set C of distinguished concepts, as the tuple: $K^{\mathcal{N}} = \langle \mathcal{T}_f, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_n}, \mathcal{A}_f \rangle$, where $\mathcal{T}_f = \emptyset$, $\mathcal{A}_f = \emptyset$ and, for each $C_k \in C$, $\mathcal{T}_{C_k}$ is the set of weighted typicality inclusions associated to neuron k as defined above.

$K^{\mathcal{N}}$ is a weighted knowledge base over the set of distinguished concepts $C = \{C_1, \ldots, C_n\}$. For multilayer feedforward networks, $K^{\mathcal{N}}$ corresponds to an acyclic conditional knowledge base, and defines a (defeasible) subsumption hierarchy among concepts. In the more general case, when the network may contain cycles, our characterization is intended to capture the properties of stationary states of the network [36]. We prove that, when a concept name C_k is introduced for each unit k in the network $\mathcal{N}$, the multi-preferential interpretation $I^{\Delta}_{\mathcal{N}}$, defined in Section 6.2, is a φ-coherent multi-preferential model of the weighted knowledge base $K^{\mathcal{N}}$.

Let us refer to the network $\mathcal{N}$ above, in which the output signals of units $j_1, \ldots, j_m$ are the input signals of unit k with synaptic weights $w_{k,j_1}, \ldots, w_{k,j_m}$, respectively. The intuition is that, as concept name C_k is associated to unit k in $\mathcal{N}$ and concept names C_{j_h} are associated to each unit j_h, the following holds: $C^I_k(x)$ corresponds to the activation y_k of unit k for a given input stimulus x, while $C^I_{j_h}(x)$ corresponds to the activation y_{j_h} of unit j_h for the same stimulus. Hence, the sum $\sum_{h=0}^{m} w_{k,j_h} C^I_{j_h}(x)$ corresponds to the *induced local field* u_k of neuron k, and equation $y_k = \varphi(u_k)$ in (7) (which holds for a stationary state, in non-feedforward networks), enforces the φ-coherence condition $C^I_k(x) = \varphi(W_k(x))$, where φ is the activation function of unit k.

Let $I^{\Delta}_{\mathcal{N}} = \langle \Delta, \cdot^I \rangle$ be the fuzzy multi-preferential interpretation of network $\mathcal{N}$ over a domain Δ of input stimuli, as defined in Section 6.2. Assume that N_C contains a concept name C_i for each unit i in the network. We can prove the following proposition.

Proposition 6. *$I^{\Delta}_{\mathcal{N}}$ is a φ-coherent multi-preferential model of the weighted knowledge base $K^{\mathcal{N}}$.*

Proof. Let $\mathcal{N}$ be network such that φ_i is the activation function of unit i in $\mathcal{N}$. Let $K^{\mathcal{N}}$ be the weighted knowledge base over the set of distinguished concepts $C = \{C_1, \ldots, C_n\}$, associated to $\mathcal{N}$ as in the construction above.

Let the fuzzy multi-preferential interpretation $I^{\Delta}_{\mathcal{N}} = \langle \Delta, \cdot^I \rangle$ of $\mathcal{N}$ over a domain Δ be defined according to Definition 10, in Section 6.2, but assuming that N_C contains a concept name C_i for each unit i in the network.

Given the set $\mathcal{T}_{C_k}$ of weighted typicality inclusions for $C_k \in C$ in $K^{\mathcal{N}}$:

$\mathbf{T}(C_k) \sqsubseteq C_{j_1}$ with w_{k,j_1},
…,
$\mathbf{T}(C_k) \sqsubseteq C_{j_m}$ with w_{k,j_m}

by construction, there are units $k, j_1, \ldots, j_m$ in $\mathcal{N}$, such that the output signals of units $j_1, \ldots, j_m$ are the input signals of unit k with synaptic weights $w_{k,j_1}, \ldots, w_{k,j_m}$.

By construction of the fuzzy interpretation $I^{\Delta}_{\mathcal{N}}$, for all $x \in \Delta$ and $C_k \in N_C$, $C_k^{I^{\Delta}_{\mathcal{N}}}(x) = y_k(x)$, i.e., $C_k^{I^{\Delta}_{\mathcal{N}}}(x)$ corresponds to the activation $y_k(x)$ of neuron k for the stimulus x. We have to prove that $I^{\Delta}_{\mathcal{N}}$ satisfies the φ-coherence condition.

Note that, in the construction of $I^{\Delta}_{\mathcal{N}}$, in case the network is not feedforward, we have assumed that, for any input stimulus x in Δ, the network reaches a stationary state, in which (for all k) $y_k(x)$ is the activity level of unit k. Then, equations (7) holds for unit k, i.e.:

$$u_k = \sum_{h=0}^{m} w_{k,j_h} y_{j_h} \qquad\qquad y_k = \varphi_k(u_k),$$

where φ_k is the activation function of unit k. Making input x explicit, it must hold that: $y_k(x) = \varphi_k(\sum_{h=0}^{m} w_{k,j_h} y_{j_h}(x))$, that is to say:

$$C^I_k(x) = \varphi_k\Big(\sum_{h=0}^{m} w_{k,j_h} C^I_{j_h}(x)\Big)$$

As the equation above holds for all concepts $C_k \in C$, and each domain element $x \in \Delta$, the interpretation $I^{\Delta}_{\mathcal{N}}$ satisfies the φ-coherence condition and is a φ-coherent model of $K^{\mathcal{N}}$. □

The next corollaries follow from Proposition 6 and Proposition 1, under the assumptions of Proposition 6, that is: $I^{\Delta}_{\mathcal{N}}$ is a fuzzy multi-preferential interpretation of a network $\mathcal{N}$ built over a domain Δ of input stimuli, as defined in Section 6.2, and N_C contains a concept name C_i for each unit i in $\mathcal{N}$.

Corollary 3. *$I^{\Delta}_{\mathcal{N}}$ is a faithful multi-preferential model of the weighted knowledge base $K^{\mathcal{N}}$, provided the activation functions φ_k of all units are monotone non-decreasing.*

Corollary 4. *$I^{\Delta}_{\mathcal{N}}$ is a coherent multi-preferential model of the weighted knowledge base $K^{\mathcal{N}}$, provided the activation functions φ_k of all units are monotonically increasing.*

Corollary 4 simplifies the formulation of Proposition 1 in [33]. Unlike in [33], here we are considering a non-crisp interpretation for typicality concepts.

By Proposition 6 the interpretation $I^{\Delta}_{\mathcal{N}}$ constructed from the network $\mathcal{N}$, by considering the activations of units over the input stimuli in Δ, is a model of the network in a logical sense, as it is a φ-coherent model of the conditional knowledge base $K^{\mathcal{N}}$ associated to the network.

We can prove that, under the φ-coherent semantics, the knowledge base $K^{\mathcal{N}}$ provides a logical characterization of the neural network $\mathcal{N}$, as the following also holds: given any φ-coherent model $I = \langle \Delta, \cdot^I \rangle$ of the knowledge base $K^{\mathcal{N}}$, each domain element $x \in \Delta$ corresponds to a stationary state of the network $\mathcal{N}$, that is, equations (7) are satisfied when the activity level y_k of unit k is taken to be the value $C^I_k(x)$, for each k.

Proposition 7. *Let $K^{\mathcal{N}}$ be the weighted knowledge base associated to a multilayer network $\mathcal{N}$. Let $I = \langle \Delta, \cdot^I \rangle$ be any φ-coherent model of $K^{\mathcal{N}}$. For all $x \in \Delta$, let $y_j = C^I_j(x)$ be the output signal of unit j, for each unit j. Then, equations (7) hold for any unit k with incoming edges.*

Proof. Consider an element $x \in \Delta$, and let k be a unit with incoming edges such that the output signals of units $j_1, \ldots, j_m$ are the input signals of unit k with synaptic weights $w_{k,j_1}, \ldots, w_{k,j_m}$.

By construction of $K^{\mathcal{N}}$, from the φ-coherence condition, it must hold that:

$$C^I_k(x) = \varphi_k\Big(\sum_{h=0}^{m} w_{k,j_h} C^I_{j_h}(x)\Big).$$

Hence, $C^I_k(x) = \varphi_k(u_k)$, and $u_k = \sum_{h=0}^{m} w_{k,j_h} C^I_{j_h}(x)$.

As from the hypothesis $y_k = C^I_k(x)$ and, for all h, $y_{j_h} = C^I_{j_h}(x)$ it holds:

$$u_k = \sum_{h=0}^{m} w_{k,j_h} y_{j_h} \qquad\qquad y_k = \varphi_k(u_k).$$

That is, equations (7) are satisfied. □

Let us observe that any *canonical* φ-coherent model of $K^{\mathcal{N}}$ contains all the stationary states of the network $\mathcal{N}$. For feedforward networks, a canonical model describes the activity of all units in the network for all the (possibly infinitely many) input stimuli.

Proof methods for reasoning in the φ-coherent multi-preferential semantics have been developed in [38,39], for the fragment $\mathcal{LC}$ of $\mathcal{ALC}$ without roles and role restrictions, based on the finitely many-valued Gödel description logic or Łukasiewicz description logic, extended with typicality. More precisely, an Answer Set Programming encoding of an approximation of φ-coherent entailment (called φ_n-coherent entailment) has been developed for the boolean fragment $\mathcal{LC}_n\mathbf{T}$ of $\mathcal{LC}$ plus typicality, over the truth space $\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$, for an integer $n \geq 1$. The study of the finitely-valued case, is indeed motivated by the undecidability results for fuzzy description logics with general inclusion axioms [37,30].

In the next section, we prove that, under suitable conditions, the φ-coherent semantics in the finitely-valued case is indeed an approximation of the φ-coherent semantics in the fuzzy case.

7. Approximating φ-coherent models in the finitely-valued case

While in Sections 3 and 4 we have defined a fuzzy $\mathcal{ALC}$ with typicality and its closure constructions, in a similar way, one can define a *finitely many-valued $\mathcal{ALC}$ with typicality*, $\mathcal{ALC}_n\mathbf{T}$, by building on the finitely-valued description logic $\mathcal{ALC}_n$, and taking $C_n = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$ (for $n \geq 1$) as the truth value space.

The idea is that of approximating function φ with a function φ_n over the truth space C_n, by developing a φ_n-coherent semantics, which is indeed an approximation of the φ-coherent semantics (under some conditions). In the following, for simplicity, we consider a single function φ, rather than a different function φ_i for each unit i, but the results generalize to the case of multiple functions.

Let us assume that φ is continuous function $\varphi : \mathbb{R} \to [0,1]$, and that the chosen t-norm, s-norm and negation function in $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ are continuous as well. We define the φ_n-coherent semantics as follows.

Values $v \in [0,1]$ are approximated to the nearest value in C_n:

$$[v]^n = \begin{cases} 0 & \text{if } v \leq \frac{1}{2n} \\ \frac{i}{n} & \text{if } \frac{2i-1}{2n} < v \leq \frac{2i+1}{2n}, \text{ for } 0 < i < n \\ 1 & \text{if } \frac{2n-1}{2n} < v \end{cases} \tag{9}$$

For an integer $n \geq 1$, let $\varphi_n : \mathbb{R} \to C_n$ be defined as:

$$\varphi_n(z) = [\varphi(z)]^n,$$

for all $z \in \mathbb{R}$. The notions of φ_n-*coherent model* and φ_n-*coherent entailment* can be defined similarly to φ-*coherent model* and φ-*coherent entailment*, by replacing φ with φ_n in Definitions 6 and 7.

Observe that the sequence of functions $(\varphi_n)_{n\in\mathbb{N}}$ uniformly converges to function φ, i.e., for all $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$| \varphi_n(z) - \varphi(z) | < \varepsilon, \ \forall z \in \mathbb{R} \tag{10}$$

Indeed, from the definition of φ_n, $|\varphi_n(z) - \varphi(z)| \leq \frac{1}{2n}$. We can get $|\varphi_n(z) - \varphi(z)| \leq \frac{1}{2n} < \varepsilon$, by choosing $n_0 = \lceil \frac{1}{2\varepsilon} \rceil + 1$.

For any $v \in \mathbb{R}$, $lim_{n\to\infty}[v]^n = v$. Hence, for any concept name $A \in N_C$, fuzzy $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ and $x \in \Delta$, $lim_{n\to\infty}[A^I(x)]^n = A^I(x)$. As we are considering continuous combination functions, for any concept D_j, $lim_{n\to\infty}[D_j^I(x)]^n = D_j^I(x)$. Let $W_i^I(x) = \sum_h w_h^i \ D_{i,h}^I(x)$ and let $W_i^{I,n}(x) = \sum_h w_h^i \ [D_{i,h}^I(x)]^n$. As $W_i^{I,n}(x)$ is continuous in $[D_{i,1}^I(x)]^n, \ldots, [D_{i,k}^I(x)]^n$, and φ is as well continuous, their composition is a continuous function, and:

$$lim_{n\to\infty}\varphi(W_i^{I,n}(x)) = \varphi(W_i^I(x)) \tag{11}$$

that is, for all $\varepsilon > 0$ there is an $m_0 \in \mathbb{N}$ such that, for all $n \geq m_0$,

$$| \varphi(W_i^{I,n}(x)) - \varphi(W_i^I(x)) | < \varepsilon.$$

Therefore the following lemma holds.

Lemma 1. *Given a continuous function $\varphi : \mathbb{R} \to [0,1]$, and an $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ interpretation I, $lim_{n\to\infty}\varphi_n(W_i^{I,n}(x)) = \varphi(W_i^I(x))$, for all $i = 1, \ldots, k$.*

Given an $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$, we can define an $\mathcal{LC}_n^{\mathbf{F}}\mathbf{T}$ interpretation $I_n = \langle \Delta, \cdot^{I_n} \rangle$ over the value space C_n by letting: $C^{I_n}(x) = [C^I(x)]^n$, for all concepts C, and $a^{I_n} = a^I$, for all $a \in N_I$.
We can then prove the following proposition.

Proposition 8. *Let $K = \langle \mathcal{T}, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_k}, \mathcal{A} \rangle$ be a weighted $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ knowledge base, and $\varphi : \mathbb{R} \to [0,1]$ a continuous function.*

(i) If $C_i^I(x) = \varphi(\sum_h w_h^i \ D_{i,h}^I(x))$, then, for all $\varepsilon > 0$, there is a $k_0 \in \mathbb{N}$ such that for all $n \geq k_0$, $|C_i^{I_n}(x) - \varphi_n(\sum_h w_h^i \ D_{i,h}^{I_n}(x))| < \varepsilon$.
(ii) If $C_i^I(x) \neq \varphi(\sum_h w_h^i \ D_{i,h}^I(x))$, then there exist an $\varepsilon > 0$ and a $k_0 \in \mathbb{N}$ such that for all $n \geq k_0$, $|C_i^{I_n}(x) - \varphi_n(\sum_h w_h^i \ D_{i,h}^{I_n}(x))| > \varepsilon$.

Proof. For item (i), assume condition $C_i^I(x) = \varphi(\sum_h w_h^i \ D_{i,h}^I(x))$ holds. As $C_i^{I_n}(x)$ converges to $C_i^I(x)$, and, by Lemma 1, $\varphi_n(\sum_h w_h^i \ D_{i,h}^{I_n}(x))$ converges to $\varphi(\sum_h w_h^i \ D_{i,h}^I(x))$, the thesis follows.

For item (ii), assume $C_i^I(x) \neq \varphi(\sum_h w_h^i \ D_{i,h}^I(x))$, and let $d = |C_i^I(x) - \varphi(\sum_h w_h^i \ D_{i,h}^I(x))|$. Let $\varepsilon = d/3$.

As $C_i^{I_n}(x)$ converges to $C_i^I(x)$, there is an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $|C_i^{I_n}(x) - C_i^I(x)| < \varepsilon = d/3$. By Lemma 1, there is an $m_0 \in \mathbb{N}$ such that for all $n \geq m_0$, $|\varphi_n(\sum_h w_h^i \ D_{i,h}^{I_n}(x)) - \varphi(\sum_h w_h^i \ D_{i,h}^I(x))| < \varepsilon = d/3$.

Let $k_0 = max\{n_0, m_0\}$. Then,

$$|C_i^{I_n}(x) - \varphi_n(\sum_h w_h^i \ D_{i,h}^{I_n}(x)| \geq d/3 = \varepsilon$$

for all $n \geq k_0$. □

Note that the notion of φ_n-coherence may fail to characterize all the stationary states of a network as, although for some $x \in \Delta$ it may hold that $C_i^I(x) = \varphi(\sum_h w_h^i \ D_{i,h}^I(x))$, for all concepts C_i, there is no guarantee that some n_0 exists such that, for all $n > n_0$, $C_i^{I_n}(x) = \varphi_n(\sum_h w_h^i \ D_{i,h}^{I_n}(x))$. Nevertheless, by item (*i*), at the limit the distance between $C_i^{I,n}(x)$ and $\varphi_n(\sum_h w_h^i \ D_{i,h}^{I,n}(x))$ converges to 0.

On the other hand, it may be the case that $C_i^{I_n}(x) = \varphi_n(\sum_h w_h^i \ D_{i,h}^{I_n}(x))$ holds for some n, due to the approximation, while $C_i^I(x) = \varphi(\sum_h w_h^i \ D_{i,h}^I(x))$ does not hold. In such a case, by item (*ii*), there must be a k_0 such that for all values of $n \geq k_0$, the first equality will not hold.

In the following section we exploit the proof methods developed in [39,40] in the verification of the properties of some trained feedforward networks under the φ_n-coherent semantics.

8. An experimentation: model checking and entailment for the verification of facial emotion recognition

While a neural network, once trained, can quickly classify new stimuli (i.e., perform instance checking), other reasoning services such as satisfiability, entailment and model-checking are missing. Such reasoning tasks are useful for validating knowledge that has been learned, including proving whether the network satisfies some (strict or conditional) properties.

In the finitely-valued case, Datalog with weakly stratified negation has been used for developing a model-checking approach for verifying multilayer networks [40]. Still in the finitely-valued case, an ASP-based approach can be exploited for reasoning with weighted conditional KBs under φ_n-coherent entailment [38,39].

Table 3
Results for checking formulae on the test set. The number of counterexamples for $\mathbf{T}(E) \sqsubseteq F \geq k/n$ is provided for $k = 1, \ldots, 4$, as well as the total number of instances of $\mathbf{T}(E)$.

E	F	k = 1	k = 2	k = 3	k = 4	#T(E)
surprise	$au1 \sqcup au2 \sqcup au5$	54	66	79	140	294
surprise	$au1 \sqcup au5 \sqcup au15 \sqcup au20 \sqcup au26$	2	3	6	59	294
fear	$au1 \sqcup au2 \sqcup au4 \sqcup au5$	7	9	10	21	45
fear	$au1 \sqcup au2 \sqcup au4 \sqcup au5 \sqcup au20 \sqcup au26$	0	0	2	9	45
happiness	$au1 \sqcup au6 \sqcup au12 \sqcup au14$	0	0	0	22	255
happiness	$au6 \sqcup au12$	0	0	1	32	255
happiness	$au6 \sqcap au12$	6	15	23	98	255
happiness	$au12$	0	0	1	35	255
anger	$au4 \sqcup au5 \sqcup au7 \sqcup au23$	5	6	7	44	212

Both the entailment and the model-checking approaches have been experimented in the verification of properties of some trained feedforward networks and, in the following, we report some results.

We concentrate, in particular, on the verification of formulae of the form $\mathbf{T}(E) \sqsubseteq F \geq \alpha$ where E is an output class (i.e. one of the possible outputs of classification, or a single output class the network is trained to recognize), and F is a boolean combination of input classes.

The interest for such formulae lies in the fact that a property $\mathbf{T}(E) \sqsubseteq F \geq \alpha$ tells something about the stimuli that are classified as Es with high membership (highest in C_n), and could then be seen as describing what the network intends as a prototypical E.

It might of course be the case that $\mathbf{T}(E) \sqsubseteq F \geq \alpha$ holds, and the corresponding strict version $E \sqsubseteq F \geq \alpha$ does not. Similar considerations apply to inclusions of the form $F \sqsubseteq \mathbf{T}(E)$ and $F \sqsubseteq E$.

8.1. Model checking

Based on the general idea of using model checking for verifying the properties of a neural network, as described in Section 6, in [40] we have developed a Datalog-based approach which builds a multi-valued preferential interpretation of a trained feedforward network $\mathcal{N}$ and, then, verifies the properties of the network for post-hoc explanation.

The Datalog encoding uses weakly stratified negation and contains a component $\Pi(\mathcal{N}, \Delta, n)$ which is intended to build a (single) many-valued, preferential interpretation $I^{\Delta}_{\mathcal{N}}$ with truth degrees in C_n, and a component associated to the formulae to be checked.

The model checking approach has been experimented in the verification of properties of neural networks for the recognition of basic emotions using the Facial Action Coding System (FACS) [57]. The RAF-DB [58] data set contains almost 30000 images labeled with basic emotions or combinations of two emotions. It was used as input to OpenFace 2.0 [59], which detects a subset of the Action Units (AUs) in [57], i.e., facial muscle contractions. The relations between such AUs and emotions, studied by psychologists [60], can be used as a reference for formulae to be verified on neural networks trained to learn such relations.

From the original dataset, the images labeled with a single emotion in the set $\{surprise, fear, happiness, anger\}$ were selected. The dataset, with 4 283 images, was highly unbalanced, then the data was preprocessed by subsampling the larger classes and augmenting the minority ones using standard data augmentation techniques. The processed dataset contains 5 975 images. The images were input to OpenFace 2.0; the output intensities were rescaled in order to make their distribution conformant to the expected one in case AUs are recognized by humans [57]. The resulting 17 AUs were used as input to a neural network trained to classify its input as an instance of the four emotions. The neural network model used is a fully connected feed forward neural network with three hidden layers having 1 800, 1 200, and 600 nodes (all hidden layers use ReLU activation functions, while the softmax function is used in the output layer); the F1 score of the trained network is 0.744 (the data quality was not very high; other emotions were not considered, in order to achieve a reasonable accuracy).

The model checking approach in [40] was easily adapted to the non-crisp notion of typicality we consider in this paper, and applied, using the Clingo ASP solver as Datalog engine, taking as set of input stimuli Δ the test set, containing 1 194 images, and $n = 5$, given that AU intensities, when assigned by humans, are on a scale of five values (plus absence). Table 3 reports some results for the verification of typicality inclusions $\mathbf{T}(E) \sqsubseteq F \geq k/n$ in the finitely-valued Gödel description logic with involutive negation plus typicality $G_n \mathcal{L}C\mathbf{T}$, with the number of typical individuals for the emotion E, and the number of counterexamples for different values of k.[3]

For example, the inclusion axiom $\mathbf{T}(happiness) \sqsubseteq au12 \geq 3/5$ (where $au12$ is the activation of the lip corner puller muscle used for smiling) does not hold in the interpretation $I^{\Delta}_{\mathcal{N}}$, since it has 1 counterexample out of 255 instances of $\mathbf{T}(happiness)$, in fact, there is an instance x such that $(\mathbf{T}(happiness))^{I^{\Delta}_{\mathcal{N}}}(x) \rhd au12^{I^{\Delta}_{\mathcal{N}}}(x) < 3/5$, given that $(\mathbf{T}(happiness))^{I^{\Delta}_{\mathcal{N}}}(x) = 1$ and $au12^{I^{\Delta}_{\mathcal{N}}}(x) = 2/5$. The property holds for $2/5$, i.e., $\mathbf{T}(happiness) \sqsubseteq au12 \geq 2/5$ holds. The formula $\mathbf{T}(happiness) \sqsubseteq au1 \sqcup au6 \sqcup au12 \sqcup au14 \geq 3/5$ also holds; the other action units involved are the activations of the inner brow raiser, cheek raiser, and dimpler.

[3] In [40] conditional probabilities of fuzzy events are also considered, namely $p(F/\mathbf{T}(E))$ of concept F given concept $\mathbf{T}(E)$, based on Zadeh's probability of fuzzy events [61].

The corresponding strict inclusions, $happiness \sqsubseteq au12 \geq k/5$ and $happiness \sqsubseteq au1 \sqcup au6 \sqcup au12 \sqcup au14 \geq k/5$, do not hold even for $k = 1$.

8.2. Entailment

Based on the approximation of the φ-coherence semantics considered in section 7, Answer Set Programming (ASP) has been shown to be suitable for addressing defeasible reasoning in the finitely many-valued case with truth space $C_n = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$ [38]. A $\mathrm{P}^{\mathrm{NP}[\log]}$-completeness result for canonical φ_n-coherent entailment has been proven in [39] and some ASP encodings that deal with weighted knowledge bases with large search spaces have been developed.

The entailment of a typicality inclusion such as $\mathbf{T}(C) \sqsubseteq D\theta\alpha$ from a weighted knowledge base K is considered in the finitely-valued Gödel description logic with involutive negation plus typicality $G_n\mathcal{LC}\mathbf{T}$, introduced in [38] for the boolean fragment $\mathcal{LC}$ of $\mathcal{ALC}$. The verification can be formulated as a problem of computing *preferred answer sets* of an ASP program, considering a single distinguished individual, intended to represent a typical C-element, and selecting, as preferred answer sets, the ones maximizing the membership of the individual in concept C. For the entailment problem, the upper bound in [38] has been improved to $\mathrm{P}^{\mathrm{NP}[\log]}$ by showing an algorithm running in polynomial time and performing *parallel* queries to an NP oracle ($\mathrm{P}^{||\mathrm{NP}}$) [39]. The problem has also been shown to be $\mathrm{P}^{\mathrm{NP}[\log]}$-complete and the proof-of-concept ASP encoding has been redesigned so to obtain the desired multi-preferential semantics by taking advantage of weak constraints. The scalability of the different ASP encodings has been assessed empirically.

The entailment approach has been experimented on the same domain as the model checking approach, for a binary classification task, for the class *happiness* vs other emotions. A set of 8 835 images was used (no augmentation was needed in this case for balancing). The images were input, as in the previous case, to OpenFace 2.0, and 17 resulting AUs were used as input to a fully connected feed forward neural network, with two hidden layers of 50 and 25 nodes, using the logistic activation function for all layers. The F1 score of the trained network is 0.831.

Also in this case the truth space C_5 was used. This means that, with 17 AUs as inputs, the size of the search space for a solver is 6^{17}, i.e., more than 10^{13}. The weighted conditional knowledge base associated to the network contains 2 201 weighted typicality inclusions. The version of the solver in [39] based on weight constraints and order encoding was used. The scalability results in [39] for synthetic knowledge bases are consistent with the theoretical complexity results, showing that there are solved problem instances as well as unsolved ones (within a 30 minutes timeout) for search spaces with sizes from 10^7 to 10^{80}, and KBs containing 500 to 40 000 weighted inclusions.

Consider the formulae:

$$\mathbf{T}(happiness) \sqsubseteq au1 \sqcup au6 \sqcup au12 \sqcup au14 \geq k/5 \tag{12}$$

$$\mathbf{T}(happiness) \sqsubseteq au6 \sqcup au12 \geq k/5 \tag{13}$$

Model checking, applied to the test set for this case (2 651 individuals with 390 instances of $\mathbf{T}(happiness)$), finds that both formulae hold for $k = 3$ and do not hold for $k = 4$. As regards entailment:

- For (12), the solver finds in seconds that it does not hold for $k = 4$, and in minutes that it holds for $k = 1$, while for $k = 2, 3$, it does not provide a result in hours.
- On a variant of the experiment, with the same network structure, but using as inputs AU intensities that are not rescaled (so that the AU values are generally lower wrt the previous case), the solver finds in seconds that (12) does not hold for $k = 2$, and in minutes that it holds for $k = 1$. I.e., in this case the exact separation can be found; note that, in the previous case, the largest k for which the property holds is presumably $k = 2$ or $k = 3$, where the search space for a counterexample is much less constrained wrt the case $k = 1$; such a search space is relevant both for showing that the property does not hold, if a counterexample is found, and that it does hold, if the non-existence of a counterexample can be inferred.
- for (13), the property is found to hold for $k = 1$ and not to hold for $k = 3$, i.e., for such a value, a counterexample is found by entailment, whose search space includes all possible combinations of input vectors, while it is not found by model checking on the (limited) test set.

The network structure for this experiment was chosen to lie in the range considered in the experiments in [39], even though a much smaller network with a single hidden layer of 8 nodes (half of the input nodes) is enough to achieve a similar accuracy for the classification problem; for the resulting knowledge base (with about 150 inclusions) the formulae above can be checked in a few seconds even with a search space of size 6^{17}.

8.3. Further considerations

The entailment approach is definitely more challenging, from the computational point of view, than the model-checking one; for the latter, the verification problem is polynomial in time in the size of the domain Δ and in the size of the formula to be verified.

The two approaches can be combined, as suggested before, with model checking providing a guess for the largest value of k such that a formula $\mathbf{T}(E) \sqsubseteq F \geq k/n$ is entailed.

The entailment approach has been developed for general weighted conditional knowledge bases, which are not required to be acyclic, while in the experimentation we have considered feedforward networks. A multilayer network can be seen as a set of

weighted defeasible inclusions in a simple description logic (only including boolean concepts). However, a weighted conditional knowledge base can be more general. It can be defined for several DLs including roles (as it has been done, for instance, for $\mathcal{EL}^{\bot}$ [33], and for $\mathcal{ALC}$ in this paper), and it allows for general inclusions axioms and assertions. The combination of defeasible inclusions with strict (or fuzzy) inclusions and assertions in a weighted KB allows for the combination of the knowledge acquired from the network and symbolic knowledge in the same formalism. In the entailment based approach this can be exploited, e.g., by adding constraints on the possible inputs through ABox and TBox axioms, e.g., to exclude combinations of input values. For example, $au9 \sqcup au10 \sqcup au17 \sqsubseteq \bot \geq 1$, i.e., imposing that the three AUs have value 0, can be added, assuming that they are not compatible with *happiness*. In this case, the properties $\mathbf{T}(happiness) \sqsubseteq au1 \sqcup au6 \sqcup au12 \sqcup au14 \geq 3/5$ and $\mathbf{T}(happiness) \sqsubseteq au6 \sqcup au12 \geq 2/5$ (see section 8.2) can indeed be proved in the example (even though hours of computation are needed).

Model checking and entailment are complementary also in the sense that the limited set of stimuli used for model checking is expected to be a good sample of the real world, while entailment considers all possible stimuli in the discretized input space, i.e. (unless constraints on inputs are used, as described above), it uniformly explores the input space, even though not the whole space of real numbers. Depending on the purpose of verification, a user may be satisfied with the fact that a formula is verified to hold by model checking, even though counterexamples could be found by entailment.

9. Conclusions and related work

The paper investigates the relationships between a logic of commonsense reasoning in knowledge representation and multilayer perceptrons. It develops a fuzzy semantics for weighted knowledge bases with typicality, in which, differently from previous work [33,34], the typicality operator has a non-crisp interpretation. For the logic $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ we have considered three different closure constructions, thus defining a faithful, a coherent and a φ-coherent semantics and studied the properties of defeasible entailment, proving that the logic satisfies the KLM properties of a preferential consequence relation [6,3] for some choices of fuzzy combination functions. We have also considered a finitely many-valued version of the φ-coherent semantics, the φ_n-coherent semantics [38], and proven that it is indeed an approximation of the fuzzy φ-coherent semantics. ASP based proof methods for the φ_n-coherent entailment [39] have been exploited in our experimentation.

We have seen that a (fuzzy) multi-preferential interpretation of a trained network can be built from a domain containing a set of input stimuli, and using the activity level of neurons for the stimuli. We have proven that such an interpretation is a model of the conditional knowledge base which can be associated to the network, corresponding to a set of weighted defeasible inclusions in a fuzzy description logic. The logical interpretation of a multilayer network can be used in the verification of properties of the network based on a model checking approach and an entailment-based approach, as experimented on networks recognizing emotions from facial features.

Our semantics builds on fuzzy Description Logics [26,62,63], and we have used fuzzy concepts within a multi-preferential semantics based on semantic closure constructions which have been developed in the line of Lehmann's semantics for lexicographic closure [31] and of Kern-Isberner's c-representations [9,32]. A fuzzy extension of preferential logics has been first studied by Casini and Straccia [64] for Gödel logic, based on the Rational closure construction.

The idea of having different preference relations, associated to different typicality operators, has been first explored by Gil [65] to define a multipreference formulation of the description logic $\mathcal{ALC} + \mathbf{T}_{min}$, a typicality DL with a minimal model preferential semantics. A multi-preferential extension of the rational closure for $\mathcal{ALC}$ and some refinements has been developed by Gliozzi et al. [66,21]. The concept-wise multipreference semantics (introduced first in the two-valued case for ranked DL knowledge bases [23]) follows a different route concerning both the definition of preferences, which are associated with concepts, and the way of combining them. In particular, as we have seen in Section 3, in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ the fuzzy interpretation of concepts induces a preference relation over domain elements for each concept, based on the fuzzy combination functions. An extension of DLs with multiple preferences has also been developed by Britz and Varzinczak [67,68] to define defeasible role quantifiers and defeasible role inclusions, by associating multiple preference relations with roles. A related semantics with multiple preferences has also been proposed in a first-order logic setting by Delgrande and Rantsaudis [49].

When the preferences associated to concepts are induced by the fuzzy interpretation of concepts, the fuzzy combination functions also provide a notion of *preference combination*. A related problem of commonsense concept combination has been addressed in a probabilistic extension of the typicality description logic $\mathcal{ALC} + \mathbf{T}_{\mathsf{R}}$ by Lieto and Pozzato [69]. In the two valued case, alternative notions of preference combinations have been considered to define a global preference relation $<$ from the preferences with respect to single aspects. For instance, the multi-preferential semantics for ranked $\mathcal{EL}^{+}_{\bot}$ knowledge bases [23] exploits one of the strategies studied in Brewka's framework of basic preference descriptions [53], while an algebraic framework for preference combination in Multi-Relational Contextual Hierarchies has been developed by Bozzato et al. [70].

In the two valued case, in description logics *threshold concepts* have been introduced by Baader et al. [71]. Graded membership functions m in the semantics assign to a domain element d and a concept C a membership degree $m(d, C)$ in $[0, 1]$. The logic is two-valued and the interpretation of concepts and roles is crisp. For instance, a threshold concept $C_{>0.8}$ is interpreted as the set of domain elements having a membership degree in C greater than 0.8. Weighted Threshold Operators have as well been introduced in description logics by Porello et al. [72]. They are n-ary operators $W^{t}(C_1 : w_1, \ldots, C_n : w_n)$, where the C_i are concepts and the $w_i \in \mathbb{R}$ are weights, which compute a weighted sum of their arguments and verify whether it reaches a certain threshold t. They are also called perceptron connectives. In [73] an operator $W^{max}(C_1 : w_1, \ldots, C_n : w_n)$ is also introduced, which selects the set of entities that maximally satisfy a combination of concepts $C_1, \ldots, C_m$. It is proven that the operator can be defined in terms of the universal modality (in a monotonic DL). While the logic $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ is monotonic, the notions of faithful, coherent and φ-coherent entailment

are nonmonotonic, and cannot be encoded in a monotonic description logic (and this is also true for the two-valued case, under the faithful semantics [74]).

Our semantics, which stems from the combination of (many-valued and fuzzy) DLs semantics [27,28,30], and the semantics of preferential logics of commonsense reasoning [1–9], has also some relations with Freund's ordered models for *concept representation* [75]. Under some respects, our approach can be regarded as a simplification of the ordered model approach (in a many-valued case), as we regard features as concepts and we consider a single (rather than two) preference relation $<_C$ for a concept C, which is used for evaluating the degree of typicality of domain elements with respect to C, and which is induced by the degree of membership of domain elements in C. Under these assumptions, simple multi-preferential structures can be defined and, as we have seen, can be used for providing a semantic interpretation to multilayer networks. A two-valued version of the concept-wise multi-preferential semantics has also been considered, e.g., for ranked $\mathcal{EL}^+_\bot$ knowledge bases [23], for weighted DL knowledge bases [33,74], and for SOMs [76]. Freund's assumption that the features can be weighted on a finite scale is mitigated in our semantics, by assuming that preferences are well-founded (as usual in the KLM approach [6]). However, as we have seen, restricting to finite values provides an approximation of the fuzzy case.

The correspondence between neural network models and fuzzy systems has been first investigated by Kosko in his seminal work [77]. In his view, "at each instant the n-vector of neuronal outputs defines a fuzzy unit or a fit vector. Each fit value indicates the degree to which the neuron or element belongs to the n-dimensional fuzzy set." In our approach, in a fuzzy interpretation of a multilayer network, each concept (representing a learned category, or simply a unit) is regarded as a fuzzy set over a domain (i.e., a set of input stimuli) which is the usual way of viewing concepts in fuzzy description logics [26,62,63], and we have used fuzzy concepts within a multi-preferential semantics based on some semantic closure constructions. The problem of learning fuzzy rules has been as well investigated in the context of fuzzy description logics [78,79] based on different machine learning approaches.

Much work has been devoted to the combination of neural networks and symbolic reasoning (e.g., the work by d'Avila Garcez et al. [80–82] and Setzu et al. [83]), as well as to the definition of new computational models [84–87], and to extensions of logic programming languages with neural predicates [88,89]. Among the earliest systems combining logical reasoning and neural learning are the Knowledge-Based Artificial Neural Network (KBANN) [90], the Connectionist Inductive Learning and Logic Programming (CILP) [91] systems, and Penalty Logic [92], a non-monotonic reasoning formalism used to establish a correspondence with symmetric connectionist networks. The relationships between normal logic programs and connectionist network have been investigated by Garcez and Gabbay [91,80] and by Hitzler et al. [93]. None of these approaches provides a semantics of neural networks in terms of concept-wise multi-preferential interpretations with typicality.

The work presented in this paper opens to the possibility of adopting conditional logics as a basis for neuro-symbolic integration, e.g., by learning the weights of a conditional knowledge base from empirical data, and combining the defeasible inclusions extracted from a neural network with other defeasible or strict inclusions for inference.

Using a multi-preferential logic for the verification of typicality properties of a neural network by model-checking is a general (*model agnostic*) approach. It can be used for SOMs, as in [76], by exploiting a notion of *distance* of a stimulus from a category to define a preferential structure, as well as for MLPs, by exploiting units activity to build a fuzzy preferential interpretation. Given the simplicity of the approach, a similar construction can be adapted to other neural network models and learning approaches.

Both the model-checking approach and the entailment-based approach are *global* approaches to explanation for neural networks (see, e.g., [83] for the notions of local and global approaches), as they consider the behavior of the network over a set Δ of input stimuli. Indeed, the evaluation of typicality inclusions considers all the individuals in the domain to establish preference relations among them, with respect to different aspects. However, properties of single individuals can as well be verified (by instance checking, in DL terminology). Whether this approach can as well be considered for counterfactual reasoning has still to be investigated.

The model-checking approach does not require to consider the activity of all units, but only of the units involved in the property to be verified. In the entailment-based approach, on the other hand, all units and network parameters are considered, which limits the scalability of the approach, consistently with the complexity results. Whether it is possible to extend the logical encoding of MLPs as weighted KBs to other neural network models is a subject for future investigation. The development of a temporal extension of this formalism to capture the transient behavior of MLPs is also an interesting direction to extend this work.

Declaration of competing interest

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.

Data availability

Data will be made available on request.

Acknowledgements

This work was partially supported by MUR and GNCS-INdAM. Mario Alviano was partially supported by Italian Ministry of Research (MUR) under PNRR project FAIR "Future AI Research", CUP H23C22000860006 and by the LAIA lab (part of the SILA labs) under project Tech4You, CUP H23C22000370006.

Appendix A. Proofs of Propositions 2 and 3

Proposition 2. *Under the choice of combination functions as in Gödel logic, any $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ satisfies the postulates (REFL′), (LLE′), (RW′), (AND′), (OR′) and (CM′).*

Proof. Let $I = \langle \Delta, \cdot^I \rangle$ be an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation, where the t-norm, s-norm, implication function and negation functions are as in Gödel logic. To prove that I satisfies the properties $(REFL'), (LLE'), (RW'), (AND')$, (OR') and (CM'), when the t-norm, s-norm and implication function are as in Gödel logic, while for negation we adopt standard involutive negation, we proceed by cases.

−(**REFL′**) to prove that $\mathbf{T}(C) \sqsubseteq C \geq 1$ is satisfied in I, we have to prove that $\inf_{x\in\Delta}(\mathbf{T}(C))^I(x) \rhd C^I(x) \geq 1$.
Let us prove that for all $x \in \Delta$, $(\mathbf{T}(C))^I(x) \rhd C^I(x) \geq 1$.
We consider two cases: $(\mathbf{T}(C))^I(x) = 0$ (i.e., x is not a typical C-element) and $(\mathbf{T}(C))^I(x) > 0$ (i.e., x is a typical C-element).
If $(\mathbf{T}(C))^I(x) = 0$, $(\mathbf{T}(C))^I(x) \rhd C^I(x) = 0 \rhd C^I(x) = 1$, and the thesis holds trivially.
If $(\mathbf{T}(C))^I(x) > 0$, by definition $(\mathbf{T}(C))^I(x) = C^I(x)$. Again, $(\mathbf{T}(C))^I(x) \rhd C^I(x) = 1$, and the thesis holds.

−(**LLE′**) Assume that $\models A \equiv B$, i.e., axioms $A \sqsubseteq B \geq 1$, $B \sqsubseteq A \geq 1$ are valid in fuzzy $\mathcal{ALC}$ and that $\mathbf{T}(A) \sqsubseteq C \geq k$ is satisfied in I. We prove that $\mathbf{T}(B) \sqsubseteq C \geq 1$ is satisfied in I, that is $(\mathbf{T}(B) \sqsubseteq C)^I \geq 1$.
From the validity of $A \sqsubseteq B \geq 1$ and $B \sqsubseteq A \geq 1$, $\inf_{x\in\Delta} A^I(x) \rhd B^I(x) \geq 1$ and $\inf_{x\in\Delta} B^I(x) \rhd A^I(x) \geq 1$. Hence,

$$\text{for all } x \in \Delta,\ A^I(x) \rhd B^I(x) \geq 1 \text{ and } B^I(x) \rhd A^I(x) \geq 1 \tag{A.1}$$

This implies that: for all $x \in \Delta$, $A^I(x) \leq B^I(x)$ and $B^I(x) \leq A^I(x)$, i.e., $A^I(x) = B^I(x)$ for all $x \in \Delta$. Therefore, the preference relations $<_A$ and $<_B$ must be the same and also $A^I_{>0} = B^I_{>0}$. Hence, $\mathbf{T}(A)^I(x) = \mathbf{T}(B)^I(x)$ for all $x \in \Delta$, and from $(\mathbf{T}(A) \sqsubseteq C)^I \geq 1$, it follows that $(\mathbf{T}(B) \sqsubseteq C)^I \geq 1$, that is, $\mathbf{T}(B) \sqsubseteq C \geq 1$ is satisfied in I.

−(**RW′**) Assume that axiom $C \sqsubseteq D \geq 1$ is valid in fuzzy $\mathcal{ALC}$. Hence, it holds that $\inf_{x\in\Delta} C^I(x) \rhd D^I(x) \geq 1$ and for all $x \in \Delta$, $C^I(x) \rhd D^I(x) \geq 1$.
As we have seen above, this implies that: for all $x \in \Delta$, $C^I(x) \leq D^I(x)$.
Let us assume that $\mathbf{T}(A) \sqsubseteq C \geq 1$ is satisfied in I, i.e., $\inf_{x\in\Delta}(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$ and, for all $x \in \Delta$, $(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$.
By monotonicity of $\rhd$, $1 \leq (\mathbf{T}(A))^I(x) \rhd C^I(x) \leq (\mathbf{T}(A))^I(x) \rhd D^I(x)$. Hence, for all $x \in \Delta$, $(\mathbf{T}(A))^I(x) \rhd D^I(x) \geq 1$, so that $\mathbf{T}(A) \sqsubseteq D \geq 1$ is satisfied in I.

−(**AND′**) Let us assume that $\mathbf{T}(A) \sqsubseteq C \geq 1$ and $\mathbf{T}(A) \sqsubseteq D \geq 1$ are satisfied in I, i.e., $\inf_{x\in\Delta}(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$ and $\inf_{x\in\Delta}(\mathbf{T}(A))^I(x) \rhd D^I(x) \geq 1$. Then, for all $x \in \Delta$, $(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$ and $(\mathbf{T}(A))^I(x) \rhd D^I(x) \geq 1$.
We prove that $x \in \Delta$, $(\mathbf{T}(A))^I(x) \rhd (C \sqcap D)^I(x) \geq 1$, from which $(\mathbf{T}(A) \sqsubseteq C \sqcap D)^I \geq 1$ follows.
If $(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$ holds, then (a) $(\mathbf{T}(A))^I(x) \leq C^I(x)$ or (b) $C^I(x) \geq 1$. Note that, if (b) holds, (a) must hold as well. Hence, if $(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$ holds, $(\mathbf{T}(A))^I(x) \leq C^I(x)$ also holds.
Similarly, from $(\mathbf{T}(A))^I(x) \rhd D^I(x) \geq 1$, it follows that $(\mathbf{T}(A))^I(x) \leq D^I(x)$ holds.
Therefore, for any $x \in \Delta$, both $(\mathbf{T}(A))^I(x) \leq C^I(x)$ and $(\mathbf{T}(A))^I(x) \leq D^I(x)$ hold. It follows that $(\mathbf{T}(A))^I(x) \leq min\{C^I(x), D^I(x)\} = (C \sqcap D)^I(x)$ holds and, hence, $(\mathbf{T}(A))^I(x) \rhd (C \sqcap D)^I(x) \geq 1$ holds.

−(**OR′**) Assume $\mathbf{T}(A) \sqsubseteq C \geq 1$ and that $\mathbf{T}(B) \sqsubseteq C \geq 1$ are satisfied in I. Then, $\inf_{x\in\Delta}(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$ and $\inf_{x\in\Delta}(\mathbf{T}(B))^I(x) \rhd C^I(x) \geq 1$. Hence, for all $x \in \Delta$, $(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$ and $(\mathbf{T}(B))^I(x) \rhd C^I(x) \geq 1$.
To prove that $\mathbf{T}(A \sqcup B) \sqsubseteq C \geq 1$, we prove that, for all $x \in \Delta$, $(\mathbf{T}(A \sqcup B))^I(x) \rhd C^I(x) \geq 1$.
If $(\mathbf{T}(A \sqcup B)^I(x) = 0$, the thesis follows trivially.
If $(\mathbf{T}(A \sqcup B))^I(x) > 0$, x is a typical $A \sqcup B$-element. Then there is no $y \in \Delta$ such that $(A \sqcup B)^I(y) > (A \sqcup B)^I(x)$.
It can be proven that, when x is a typical $A \sqcup B$-element, x is also a typical A-element or a typical B-element.
Given that $(\mathbf{T}(A \sqcup B))^I(x) = (A \sqcup B)^I(x) = max\{A^I(x), B^I(x)\}$, let us assume $max\{A^I(x), B^I(x)\} = B^I(x)$. Then, for all $y \in \Delta$, $max\{A^I(y), B^I(y)\} \leq B^I(x)$, and $B^I(y) \leq B^I(x)$. Hence, there is no $y \in \Delta$ such that $B^I(y) > B^I(x)$, and x is a typical B element. Furthermore, $(\mathbf{T}(B))^I(x) = B^I(x) = (\mathbf{T}(A \sqcup B))^I(x))$.
From the hypothesis, we know that $(\mathbf{T}(B))^I(x) \rhd C^I(x) \geq 1$; hence, $(\mathbf{T}(B))^I(x) \leq C^I(x)$. It follows that $(\mathbf{T}(A \sqcup B))^I(x) = (\mathbf{T}(B))^I(x) \leq C^I(x)$, and then $(\mathbf{T}(A \sqcup B))^I(x) \rhd C^I(x) \geq 1$.
The case where $max\{A^I(x), B^I(x)\} = A(x)$ is similar.

−(**CM′**) Assume $\mathbf{T}(A) \sqsubseteq D \geq 1$ and that $\mathbf{T}(A) \sqsubseteq C \geq 1$ are satisfied in I. Then, $\inf_{x\in\Delta}(\mathbf{T}(A))^I(x) \rhd D^I(x) \geq 1$ and $\inf_{x\in\Delta}(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$. Hence, for all $x \in \Delta$, $(\mathbf{T}(A))^I(x) \rhd D^I(x) \geq 1$ and $(\mathbf{T}(A))^I(x) \rhd C^I(x) \geq 1$.
To prove that $\mathbf{T}(A \sqcap D) \sqsubseteq C \geq 1$ is satisfied in I, we prove that, for all $x \in \Delta$, $(\mathbf{T}(A \sqcap D))^I(x) \rhd C^I(x) \geq 1$.
If $(\mathbf{T}(A \sqcap D))^I(x) = 0$ the thesis holds trivially.
If $(\mathbf{T}(A \sqcap D))^I(x) > 0$, x is a typical $A \sqcap D$-element. Then, $(\mathbf{T}(A \sqcap D))^I(x) = (A \sqcap D)^I(x) = min\{A^I(x), D^I(x)\} > 0$. Also, $A^I(x) > 0$ and $D^I(x) > 0$.
We prove that x is a typical A-element. By contradiction, if x is not a typical A-element, there is a $y \in \Delta$ such that y is a typical A-element and $A^I(y) > A^I(x)$. As $\mathbf{T}(A) \sqsubseteq D \geq 1$, $(\mathbf{T}(A))^I(y) \rhd D^I(y) \geq 1$, and then $(\mathbf{T}(A))^I(y) \leq D^I(y)$. But $(\mathbf{T}(A))^I(y) = A^I(y)$ (as y is a typical A-element), hence $A^I(y) \leq D^I(y)$.

Therefore, $min\{A^I(y), D^I(y)\} = A^I(y) > A^I(x) \geq min\{A^I(x), D^I(x)\}$. Then, $(A \sqcap D)^I(y) > (A \sqcap D)^I(x)$, which contradicts the hypothesis that x is a typical $A \sqcap D$-element. Therefore, x must be a typical A-element.

As $\mathbf{T}(A) \sqsubseteq C \geq 1$ and $\mathbf{T}(A) \sqsubseteq D \geq 1$, $(\mathbf{T}(A))^I(x) \leq C^I(x)$ and $(\mathbf{T}(A))^I(x) \leq D^I(x)$ hold. Furthermore, as $(\mathbf{T}(A))^I(x) = A^I(x)$, $A^I(x) \leq C^I(x)$ and $A^I(x) \leq D^I(x)$ hold. Then, $(A \sqcap D)^I(x) = min\{A^I(x), D^I(x)\} = A^I(x)$. Thus, $(\mathbf{T}(A \sqcap D))^I(x) = (A \sqcap D)^I(x) = A^I(x) \leq C^I(x)$, and the thesis follows. □

Proposition 3. *For the choice of combination functions as in Gödel logic, (RM′) does not hold in $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ (and the same with standard involutive negation.*

Proof. Consider a KB K such that the ABox $\mathcal{A}$ contains the following assertions:

$$A(a) \leq 0.8,\ A(a) \geq 0.8\ B(a) \leq 0.3,\ B(a) \geq 0.3,\ C(a) \leq 0.9,\ C(a) \geq 0.9;$$

$$A(b) \leq 0.5,\ A(b) \geq 0.5,\ B(b) \leq 0.6,\ B(b) \geq 0.6,\ C(b) \leq 0.4,\ C(b) \geq 0.4$$

and the TBox $\mathcal{T}$ contain the axiom $\mathbf{T}(A) \sqsubseteq C \geq 1$.

Clearly K entails $\mathbf{T}(A) \sqsubseteq C \geq 1$. We show that K does not entail $\mathbf{T}(A) \sqsubseteq \neg B \geq 1$. We define an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ which is a model of K, but falsifies $\mathbf{T}(A) \sqsubseteq \neg B \geq 1$.

Let $I = \langle \Delta, \cdot^I \rangle$ be such that $\Delta = \{x, z\}$ and, for concept names A, B, C,

$$A^I(x) = 0.8,\ B^I(x) = 0.3,\ C^I(x) = 0.9$$

$$A^I(z) = 0.5,\ B^I(z) = 0.6,\ C^I(z) = 0.4$$

Hence, x is a typical A element, and the only one. $\mathbf{T}(A)^I(x) = 0.8$ and $\mathbf{T}(A)^I(x) \rhd C^I(x) = 1$. Hence, $\mathbf{T}(A) \sqsubseteq C \geq 1$ is satisfied in I. Clearly, all the assertions in ABox $\mathcal{A}$ are also satisfied in I, by letting $a^I = x$ and $b^I = z$. I is an $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ model of K.

$\mathbf{T}(A) \sqsubseteq \neg B \geq 1$ is not satisfied in I, as $\mathbf{T}(A)^I(x) = 0.8$ and $(\neg B)^I(x) = 0$ (using the negation function in Gödel logic), and $\mathbf{T}(A)^I(x) \rhd (\neg B)^I(x) = 0$. Therefore, $\mathbf{T}(A) \sqsubseteq \neg B \geq 1$ is not entailed from K.[4]

By (RM') we would conclude that $\mathbf{T}(A \sqcap B) \sqsubseteq C \geq 1$ should be entailed from K, but this is not true, as the model I of K falsifies $\mathbf{T}(A \sqcap B) \sqsubseteq C \geq 1$. In fact z is the only typical $\mathbf{T}(A \sqcap B)$ element in I and $\mathbf{T}(A \sqcap B)^I(z) = 0.5$. However, $\mathbf{T}(A \sqcap B)^I(z) \rhd C^I(z) = 0.4 < 1$. □

References

[1] J. Delgrande, A first-order conditional logic for prototypical properties, Artif. Intell. 33 (1) (1987) 105–130.
[2] D. Makinson, General theory of cumulative inference, in: Non-Monotonic Reasoning, 2nd International Workshop, Grassau, FRG, June 13–15, 1988, Proceedings, 1988, pp. 1–18.
[3] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 1988.
[4] S. Kraus, D. Lehmann, M. Magidor, Nonmonotonic reasoning, preferential models and cumulative logics, Artif. Intell. 44 (1–2) (1990) 167–207.
[5] J. Pearl, System Z: a natural ordering of defaults with tractable applications to nonmonotonic reasoning, in: TARK'90, Pacific Grove, CA, USA, 1990, pp. 121–135.
[6] D. Lehmann, M. Magidor, What does a conditional knowledge base entail?, Artif. Intell. 55 (1) (1992) 1–60.
[7] S. Benferhat, C. Cayrol, D. Dubois, J. Lang, H. Prade, Inconsistency management and prioritized syntax-based entailment, in: Proc. IJCAI'93, Chambéry, 1993, pp. 640–647.
[8] R. Booth, J.B. Paris, A note on the rational closure of knowledge bases with both positive and negative knowledge, J. Log. Lang. Inf. 7 (2) (1998) 165–190, https://doi.org/10.1023/A:1008261123028.
[9] G. Kern-Isberner, Conditionals in Nonmonotonic Reasoning and Belief Revision - Considering Conditionals as Agents, LNCS, vol. 2087, Springer, 2001.
[10] D. Lewis, Counterfactuals, Basil Blackwell Ltd, 1973.
[11] D. Nute, Topics in Conditional Logic, Reidel, Dordrecht, 1980.
[12] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider, The Description Logic Handbook - Theory, Implementation, and Applications, Cambridge, 2007.
[13] L. Giordano, V. Gliozzi, N. Olivetti, G.L. Pozzato, Preferential description logics, in: LPAR 2007, in: LNAI, vol. 4790, Springer, Yerevan, Armenia, 2007, pp. 257–272.
[14] K. Britz, J. Heidema, T. Meyer, Semantic preferential subsumption, in: G. Brewka, J. Lang (Eds.), KR 2008, AAAI Press, Sidney, Australia, 2008, pp. 476–484.
[15] L. Giordano, V. Gliozzi, N. Olivetti, G.L. Pozzato, ALC+T: a preferential extension of Description Logics, Fundam. Inform. 96 (2009) 1–32.
[16] G. Casini, U. Straccia, Rational closure for defeasible description logics, in: T. Janhunen, I. Niemelä (Eds.), JELIA 2010, in: LNCS, vol. 6341, Springer, Helsinki, 2010, pp. 77–90.
[17] G. Casini, T. Meyer, I.J. Varzinczak, K. Moodley, Nonmonotonic reasoning in description logics: rational closure for the ABox, in: 26th International Workshop on Description Logics (DL 2013), in: CEUR Workshop Proceedings, vol. 1014, 2013, pp. 600–615.
[18] L. Giordano, V. Gliozzi, N. Olivetti, G.L. Pozzato, Semantic characterization of rational closure: from propositional logic to description logics, Artif. Intell. 226 (2015) 1–33.
[19] M. Pensel, A. Turhan, Reasoning in the defeasible description logic $EL_\perp$ - computing standard inferences under rational and relevant semantics, Int. J. Approx. Reason. 103 (2018) 28–70.
[20] G. Casini, U. Straccia, T. Meyer, A polynomial time subsumption algorithm for nominal safe ELO⊥ under rational closure, Inf. Sci. 501 (2019) 588–620.
[21] L. Giordano, V. Gliozzi, A reconstruction of multipreference closure, Artif. Intell. 290 (2021).
[22] G. Casini, T.A. Meyer, I. Varzinczak, Contextual conditional reasoning, in: AAAI-21, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 6254–6261.

[4] Similarly, using standard involutive negation, $(\neg B)^I(x) = 1 - 0.3 = 0.7$, $\mathbf{T}(A)^I(x) \rhd (\neg B)^I(x) < 1$, and $\mathbf{T}(A) \sqsubseteq \neg B \geq 1$ is not satisfied in I as well, and $\mathbf{T}(A) \sqsubseteq \neg B \geq 1$ is not entailed from K.

[23] L. Giordano, D. Theseider Dupré, An ASP approach for reasoning in a concept-aware multipreferential lightweight DL, Theory Pract. Log. Program. 10 (5) (2020) 751–766.
[24] L. Giordano, D. Theseider Dupré, A framework for a modular multi-concept lexicographic closure semantics, in: Proc. 18th Int. Workshop on Non-Monotonic Reasoning, NMR2020, September 12th - 14th, 2020, 2020.
[25] T. Kohonen, M. Schroeder, T. Huang (Eds.), Self-Organizing Maps, third edition, Springer Series in Information Sciences, Springer, 2001.
[26] U. Straccia, Towards a fuzzy description logic for the semantic web (preliminary report), in: ESWC 2005, Heraklion, Crete, May 29 - June 1, 2005, in: LNCS, vol. 3532, Springer, 2005, pp. 167–181.
[27] G. Stoilos, G.B. Stamou, V. Tzouvaras, J.Z. Pan, I. Horrocks, Fuzzy OWL: uncertainty and the semantic web, in: OWLED*05 Workshop on OWL Galway, Ireland, Nov 11-12, 2005, in: CEUR Workshop Proc., vol. 188, 2005.
[28] T. Lukasiewicz, U. Straccia, Description logic programs under probabilistic uncertainty and fuzzy vagueness, Int. J. Approx. Reason. 50 (6) (2009) 837–853.
[29] A. García-Cerdaña, E. Armengol, F. Esteva, Fuzzy description logics and t-norm based fuzzy logics, Int. J. Approx. Reason. 51 (6) (2010) 632–655, https://doi.org/10.1016/j.ijar.2010.01.001.
[30] S. Borgwardt, R. Peñaloza, Undecidability of fuzzy description logics, in: G. Brewka, T. Eiter, S.A. McIlraith (Eds.), Proc. KR 2012, Rome, Italy, June 10–14, 2012, AAAI Press, 2012.
[31] D.J. Lehmann, Another perspective on default reasoning, Ann. Math. Artif. Intell. 15 (1) (1995) 61–82.
[32] G. Kern-Isberner, C. Eichhorn, Structural inference from conditional knowledge bases, Stud. Log. 102 (4) (2014) 751–769.
[33] L. Giordano, D. Theseider Dupré, Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model, in: Proc. JELIA 2021, May 17-20, in: LNCS, vol. 12678, Springer, 2021, pp. 225–242.
[34] L. Giordano, On the KLM properties of a fuzzy DL with Typicality, in: Proc. ECSQARU 2021, Prague, Sept. 21-24, 2021, in: LNCS, vol. 12897, Springer, 2021, pp. 557–571.
[35] L. Giordano, From weighted conditionals of multilayer perceptrons to a gradual argumentation semantics, in: Proc. 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021, Milan, Italy, Nov. 29, in: CEUR Workshop Proceedings, vol. 3086, CEUR-WS.org, 2021, http://ceur-ws.org/Vol-3086/paper8.pdf.
[36] S. Haykin, Neural Networks - a Comprehensive Foundation, Pearson, 1999.
[37] M. Cerami, U. Straccia, On the (un)decidability of fuzzy description logics under Łukasiewicz t-norm, Inf. Sci. 227 (2013) 1–21, https://doi.org/10.1016/j.ins.2012.11.019.
[38] L. Giordano, D. Theseider Dupré, An ASP approach for reasoning on neural networks under a finitely many-valued semantics for weighted conditional knowledge bases, Theory Pract. Log. Program. 22 (4) (2022) 589–605, https://doi.org/10.1017/S1471068422000163.
[39] M. Alviano, L. Giordano, D. Theseider Dupré, Complexity and scalability of defeasible reasoning in many-valued weighted knowledge bases, CoRR, arXiv: 2303.04534 [abs], 2023, https://doi.org/10.48550/arXiv.2303.04534.
[40] F. Bartoli, M. Botta, R. Esposito, L. Giordano, D. Theseider Dupré, An ASP approach for reasoning about the conditional properties of neural networks: an experiment in the recognition of basic emotions, in: Datalog 2.0, in: CEUR Workshop Proceedings, vol. 3203, CEUR-WS.org, 2022, pp. 54–67, http://ceur-ws.org/Vol-3203/paper4.pdf.
[41] F. Bartoli, M. Botta, R. Esposito, L. Giordano, V. Gliozzi, D. Theseider Dupré, From common sense reasonig to neural network models: a conditional and multi-preferential approach for explainability and neuro-symbolic integration, in: C. Beierle, M. Ragni, F. Stolzenburg, K. Sauerwald, M. Thimm (Eds.), Proceedings of the 8th Workshop on Formal and Cognitive Reasoning, in: CEUR Workshop Proceedings, vol. 3242, CEUR-WS.org, 2022, pp. 66–78.
[42] P. Cintula, P. Hájek, C. Noguera (Eds.), Handbook of Mathematical Fuzzy Logic, vol. 37-38, College Publications, 2011.
[43] F. Bobillo, U. Straccia, Reasoning with the finitely many-valued Łukasiewicz fuzzy Description Logic SROIQ, Inf. Sci. 181 (4) (2011) 758–778, https://doi.org/10.1016/j.ins.2010.10.020.
[44] F. Bobillo, M. Delgado, J. Gómez-Romero, U. Straccia, Joining Gödel and Zadeh fuzzy logics in fuzzy description logics, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 20 (4) (2012) 475–508, https://doi.org/10.1142/S0218488512500249.
[45] S. Borgwardt, R. Peñaloza, The complexity of lattice-based fuzzy description logics, J. Data Semant. 2 (1) (2013) 1–19.
[46] F. Bobillo, U. Straccia, Reasoning within fuzzy OWL 2 EL revisited, Fuzzy Sets Syst. 351 (2018) 1–40.
[47] L. Giordano, V. Gliozzi, Encoding a preferential extension of the description logic SROIQ into SROIQ, in: Proc. ISMIS 2015, in: LNCS, vol. 9384, Springer, 2015, pp. 248–258.
[48] R. Booth, G. Casini, T. Meyer, I. Varzinczak, On rational entailment for propositional typicality logic, Artif. Intell. 277 (2019).
[49] J. Delgrande, C. Rantsoudis, A preference-based approach for representing defaults in first-order logic, in: Proc. 18th Int. Workshop on Non-Monotonic Reasoning, NMR, 2020.
[50] P.A. Bonatti, L. Sauro, On the logical properties of the nonmonotonic description logic DL^N, Artif. Intell. 248 (2017) 85–111.
[51] G. Casini, U. Straccia, Lexicographic closure for defeasible description logics, in: Proc. of Australasian Ontology Workshop, vol. 969, 2012, pp. 28–39.
[52] P.A. Bonatti, C. Lutz, F. Wolter, The complexity of circumscription in DLs, J. Artif. Intell. Res. 35 (2009) 717–773.
[53] G. Brewka, A rank based description language for qualitative preferences, in: 6th Europ. Conf. on Artificial Intelligence, ECAI'2004, Valencia, Spain, August 22-27, 2004, pp. 303–307.
[54] E. Weydert, System JLZ - rational default reasoning by minimal ranking constructions, J. Appl. Log. 1 (3–4) (2003) 273–308.
[55] K. Britz, G. Casini, T. Meyer, K. Moodley, U. Sattler, I. Varzinczak, Principles of KLM-style defeasible description logics, ACM Trans. Comput. Log. 22 (1) (2021) 1:1–1:46.
[56] P. McLeod, K. Plunkett, E. Rolls (Eds.), Introduction to Connectionist Modelling of Cognitive Processes, Oxford University Press, 1998.
[57] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System, Research Nexus, 2002.
[58] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 2584–2593.
[59] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L. Morency, Openface 2.0: facial behavior analysis toolkit, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, IEEE Computer Society, 2018, pp. 59–66.
[60] B. Waller, J.C. Jr., A. Burrows, Selection for universal facial emotion, Emotion 8 (3) (2008) 435–439.
[61] L. Zadeh, Probability measures of fuzzy events, J. Math. Anal. Appl. 23 (1968) 421–427.
[62] T. Lukasiewicz, U. Straccia, Managing uncertainty and vagueness in description logics for the semantic web, J. Web Semant. 6 (4) (2008) 291–308.
[63] F. Bobillo, U. Straccia, The fuzzy ontology reasoner fuzzyDL, Knowl.-Based Syst. 95 (2016) 12–34.
[64] G. Casini, U. Straccia, Towards rational closure for fuzzy logic: the case of propositional Gödel logic, in: Proc. LPAR-19, Stellenbosch, South Africa, December 14-19, 2013, in: LNCS, vol. 8312, Springer, 2013, pp. 213–227.
[65] O.F. Gil, On the non-monotonic description logic ALC+T_{min}, CoRR, arXiv:1404.6566 [abs], 2014.
[66] V. Gliozzi, Reasoning about multiple aspects in rational closure for DLs, in: Proc. AI*IA 2016 - XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 - December 1, 2016, 2016, pp. 392–405.
[67] K. Britz, I.J. Varzinczak, Rationality and context in defeasible subsumption, in: Proc. 10th Int. Symp. on Found. of Information and Knowledge Systems, FoIKS 2018, Budapest, May 14-18, 2018, 2018, pp. 114–132.
[68] A. Britz, I. Varzinczak, Contextual rational closure for defeasible ALC (extended abstract), in: Proc. 32nd International Workshop on Description Logics, Oslo, Norway, June 18–21, 2019, 2019.

[69] A. Lieto, G. Pozzato, A description logic of typicality for conceptual combination, in: Proc. ISMIS 2018, Cyprus, October 29-31, 2018, in: LNCS, vol. 11177, Springer, 2018, pp. 189–199.
[70] L. Bozzato, T. Eiter, R. Kiesel, Reasoning on multi-relational contextual hierarchies via answer set programming with algebraic measures, Theory Pract. Log. Program. 21 (2021) 593–609.
[71] F. Baader, G. Brewka, O.F. Gil, Adding threshold concepts to the description logic EL, in: Frontiers of Combining Systems - 10th International Symposium, FroCoS 2015, Wroclaw, Poland, September 21-24, 2015. Proceedings, in: Lecture Notes in Computer Science, vol. 9322, Springer, 2015, pp. 33–48.
[72] P. Galliani, G. Righetti, O. Kutz, D. Porello, N. Troquard, Perceptron connectives in knowledge representation, in: Knowledge Engineering and Knowledge Management - 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16-20, 2020, Proceedings, in: Lecture Notes in Computer Science, vol. 12387, Springer, 2020, pp. 183–193.
[73] D. Porello, O. Kutz, G. Righetti, N. Troquard, P. Galliani, C. Masolo, A toothful of concepts: towards a theory of weighted concept combination, in: Proceedings of the 32nd International Workshop on Description Logics, Oslo, Norway, June 18-21, 2019, in: CEUR Workshop Proceedings, vol. 2373, CEUR-WS.org, 2019.
[74] L. Giordano, D. Theseider Dupré, Weighted conditional $EL^{\perp}$ knowledge bases with integer weights: an ASP approach, in: Proc. 37th Int. Conf. on Logic Programming, ICLP 2021 (Technical Communications), Porto, Sept. 20-27, 2021, in: EPTCS, vol. 345, 2021, pp. 70–76.
[75] M. Freund, Ordered models for concept representation, J. Log. Comput. 30 (6) (2020).
[76] L. Giordano, V. Gliozzi, D. Theseider Dupré, A conditional, a fuzzy and a probabilistic interpretation of self-organizing maps, J. Log. Comput. 32 (2) (2022) 178–205.
[77] B. Kosko, Neural Networks and Fuzzy Systems: a Dynamical Systems Approach to Machine Intelligence, Prentice Hall, 1992.
[78] F.A. Lisi, U. Straccia, Learning in description logics with fuzzy concrete domains, Fundam. Inform. 140 (3–4) (2015) 373–391.
[79] U. Straccia, M. Mucci, pFOIL-DL: learning (fuzzy) EL concept descriptions from crisp OWL data using a probabilistic ensemble estimation, in: Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, April 13–17, 2015, ACM, 2015, pp. 345–352.
[80] A.S. d'Avila Garcez, K. Broda, D.M. Gabbay, Symbolic knowledge extraction from trained neural networks: a sound approach, Artif. Intell. 125 (1–2) (2001) 155–207.
[81] A.S. d'Avila Garcez, L.C. Lamb, D.M. Gabbay, Neural-Symbolic Cognitive Reasoning, Cognitive Technologies, Springer, 2009.
[82] A.S. d'Avila Garcez, M. Gori, L.C. Lamb, L. Serafini, M. Spranger, S.N. Tran, Neural-symbolic computing: an effective methodology for principled integration of machine learning and reasoning, FLAP 6 (4) (2019) 611–632.
[83] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, GlocalX - from local to global explanations of black box AI models, Artif. Intell. 294 (2021) 103457, https://doi.org/10.1016/j.artint.2021.103457.
[84] L.C. Lamb, A.S. d'Avila Garcez, M. Gori, M.O.R. Prates, P.H.C. Avelar, M.Y. Vardi, Graph neural networks meet neural-symbolic computing: a survey and perspective, in: C. Bessiere (Ed.), Proc. IJCAI 2020, ijcai.org, 2020, pp. 4877–4884.
[85] L. Serafini, A.S. d'Avila Garcez, Learning and reasoning with logic tensor networks, in: XVth Int. Conf. of the Italian Association for Artificial Intelligence, AI*IA 2016, Genova, Italy, Nov 29 - Dec 1, in: LNCS, vol. 10037, Springer, 2016, pp. 334–348.
[86] P. Hohenecker, T. Lukasiewicz, Ontology reasoning with deep neural networks, J. Artif. Intell. Res. 68 (2020) 503–540.
[87] D. Le-Phuoc, T. Eiter, A. Le-Tuan, A scalable reasoning and learning approach for neural-symbolic stream fusion, in: AAAI 2021, February 2-9, AAAI Press, 2021, pp. 4996–5005.
[88] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L.D. Raedt, Deepproblog: neural probabilistic logic programming, in: NeurIPS 2018, 3-8 December 2018, Montréal, Canada, 2018, pp. 3753–3763.
[89] Z. Yang, A. Ishay, J. Lee, Neurasp: embracing neural networks into answer set programming, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, 2020, pp. 1755–1762.
[90] G.G. Towell, J.W. Shavlik, Knowledge-based artificial neural networks, Artif. Intell. 70 (1–2) (1994) 119–165.
[91] A.S. d'Avila Garcez, G. Zaverucha, The connectionist inductive learning and logic programming system, Appl. Intell. 11 (1) (1999) 59–77.
[92] G. Pinkas, Reasoning, nonmonotonicity and learning in connectionist networks that capture propositional knowledge, Artif. Intell. 77 (2) (1995) 203–247.
[93] P. Hitzler, S. Hölldobler, A.K. Seda, Logic programs and connectionist networks, J. Appl. Log. 2 (3) (2004) 245–272.