Defeasible Reasoning with Prototype Descriptions: A New Preference Order

Gabriele Sacco[1,2], Loris Bozzato[3] and Oliver Kutz[2]

[1] *Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy*
[2] *Free University of Bozen-Bolzano, Piazza Domenicani 3, 39100, Bolzano, Italy*
[3] *DiSTA - Università dell'Insubria, Via O. Rossi 9, 21100 Varese, Italy*

Abstract

The representation of defeasible information in Description Logics is a well-known issue and many formal approaches have been proposed. However, in these proposals, little attention has been devoted to studying their capabilities in capturing the interpretation of typicality and exceptions from an ontological and cognitive point of view. In this regard, we are developing a model of defeasible knowledge for description logics based on combining ideas from prototype theory, weighted description logic (aka 'tooth logic'), and earlier work on justifiable exceptions. This machinery is then used to determine exceptions in case of conflicting axioms. In this paper, we analyse this formalisation with respect to some interesting cases where the defeasible properties to which we may have exceptions are also present as features in prototype descriptions. The analysis will suggest that a new preference order, which considers what happens inside the models, may be best suited and we outline how this new preference order can be defined.

Keywords

Non-monotonic logic, Typicality, Description Logics, Exceptions

1. Introduction

The study of defeasible reasoning and its modelling through non-monotonic logics has a long story in Artificial Intelligence [1, 2]. More recently, approaches for representing non-monotonicity have been proposed also in the context of Descriptions Logics (DL) [3, 4]. Following this direction, in [5] we started developing a formal approach for defeasible reasoning in DLs which satisfies some desiderata extracted from the studies on *generics* [6] and which moreover takes into account some insights we could identify within the philosophical and cognitive research on phenomena related to exceptions and defeasibility.

In the present paper, we build on that formalism by proposing a more elaborate preference relation on models which is able to address some shortcomings of the previous formulation. Therefore, we firstly describe the approach of [5] in section 2, proposing also a new terminology for some key definitions in order to simplify the presentation. Then, section 3 treats the new preference by firstly describing a problematic case for the previous one, then giving the formal definitions, and finally by showing that through those new definitions we are able to reach the desired conclusion in the outlined problematic case.

2. DLs with Weighted Prototypes

In our approach, we distinguish two parts in the knowledge bases: the actual DL knowledge base, which represents the knowledge of interest and can contain defeasible axioms and information about features of individuals, and a separate set containing prototype definitions.

In this section we outline the syntax and semantics of such enriched KBs.

DL 2024: 37th International Workshop on Description Logics, June 18–21, 2024, Bergen, Norway

gsacco@fbk.eu (G. Sacco); loris.bozzato@uninsubria.it (L. Bozzato); oliver.kutz@unibz.it (O. Kutz)
0000-0001-5613-5068 (G. Sacco); 0000-0003-1757-9859 (L. Bozzato); 0000-0003-1517-7354 (O. Kutz)

2.1. Syntax: Features, Prototype Definitions, Prototype Knowledge Bases

The following definitions are independent from the DL language used for representing the main knowledge base: we consider a fixed concept language $\mathcal{L}_\Sigma$ based on a DL signature Σ with disjoint and non-empty sets NC of *concept names*, NR of *role names*, and NI of *individual names*. We identify a subset of the concept names as denoting *prototype names* by assuming a subset $\mathrm{NP} \subseteq \mathrm{NC}$. For simplicity, we call *general concepts* the concepts composed only of concepts in $\mathrm{NC} \setminus \mathrm{NP}$.

The features associated with prototypes together with the degree of their importance are given in *prototype definitions.*

Definition 1 (Positive prototype definition). *Let $P \in \mathrm{NP}$ be a prototype name, let $C_1, \ldots, C_m$ be general concepts of $\mathcal{L}_\Sigma$ and let $\overline{w} = (w_1, \ldots, w_m) \in \mathbb{Q}^m$ be a weight vector of rational numbers, where for every $i \in \{1, \ldots, m\}$ we have $w_i > 0$. Then, the expression*

$$P(C_1 : w_1, \ldots, C_m : w_m)$$

is called a (positive) prototype definition *for P.*

Note that this definition of prototypes is similar to the definition of concepts by the tooth operator defined in [7]. Intuitively, the weights associated to the features can be then combined to compute a score denoting the degree of typicality of an instance w.r.t. the prototype: for the current definition, weights are assumed to be positive and features are independent.

In the present work we are agnostic with respect to where the weights come from. However, in future we plan to study sources for getting these weights as for example, learning models.

Moreover, we use here rational weights, which is sufficient for practical purposes. Real numbers could be allowed as well, but this would not substantially change the formal setup; this is also the case for the related perceptron logic [8].

Another important remark is that, since some features could be mutually exclusive (e.g. the color of an apple can be red or green, but not both), prototype definitions should not be seen as denoting a “perfect individual”.

A final point on prototype definitions is that to allow for a direct comparison across scores of different prototypes, these need to be normalised to a common value interval, possibly with a scoring function that does not depend on the number of features defining different prototypes. In an initial proposal, we simply constrained the weights of features to be in the $[0, 1]$ interval and to prescribe further that they would add up to 1, i.e. prototypes were simply assumed to be given as *positive* and *normalised* [9]: in the following sections, we provide instead a more general proposal for normalising prototype scores. With respect to [9], in the current definition of prototype definition we allow any general concept to characterise the weighted features.

In the knowledge part of the KB, we can use prototype names in DL axioms to describe properties of the members of such classes. Here we consider the case in which prototype names are only used as primitive concepts on the left-hand side of concept inclusions.

Definition 2 (Prototype axiom). *A concept inclusion of the type $P \sqsubseteq D$ is a* prototype axiom *of $\mathcal{L}_\Sigma$ if $P \in \mathrm{NP}$ and D is a general concept of $\mathcal{L}_\Sigma$.*

Intuitively, these axioms are not absolute and can be “overridden” by prototype instances (cf. *defeasible axioms* in [10]), also depending on the “degree of membership” of the individual to the given prototype (i.e., the satisfaction of its features).

As noted above, we consider knowledge bases which can contain prototype axioms and which are enriched with an accessory KB, the PBox $\mathcal{P}$ providing prototype definitions. Formally:

Definition 3 (Prototyped Knowledge Base, PKB). *A* prototyped knowledge base*, PKB for short, in language $\mathcal{L}_\Sigma$ is a triple $\mathfrak{K} = \langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle$ where:*

- *$\mathcal{T} = T_P \uplus T_C$ is a DL TBox consisting of concept inclusion axioms of the form $C \sqsubseteq D$, and partitioned into the disjoint sets T_P of prototype axioms and T_C of general concept inclusions based on general concepts;*
- *$\mathcal{A} = A_P \uplus A_C$ is a set of ABox assertions partitioned into the disjoint sets A_P of prototype assertions (of the form $P(a)$ with $P \in \mathrm{NP}$ and $a \in \mathrm{NI}$) and A_C of assertions for general concepts and roles;*
- *$\mathcal{P}$ is a set of prototype definitions, exactly one for each prototype name $P \in \mathrm{NP}$ appearing in the prototype TBox T_P.*

Remark. Note that a PKB $\langle \mathcal{T}, \mathcal{A}, \emptyset \rangle$ can be seen as a standard DL knowledge base.

Example 1. *Consider the following* prototyped knowledge base *$\mathcal{K} = \langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle$:*

$$\mathcal{T} = \{\, \mathtt{Dog} \sqsubseteq \mathtt{Trusted},\ \mathtt{Wolf} \sqsubseteq \neg\mathtt{Trusted},\ \mathtt{Dog} \sqsubseteq \mathtt{hasLegs},\ \mathtt{Wolf} \sqsubseteq \mathtt{hasLegs} \,\},$$

$$\begin{aligned}\mathcal{A} = \{\, & \mathtt{Dog(balto)},\ \mathtt{Wolf(balto)},\ \mathtt{Dog(pluto)},\ \mathtt{Wolf(alberto)},\ \mathtt{Dog(cerberus)},\\ & \mathtt{livesInWoods(balto)},\ \mathtt{hasLegs(balto)},\ \mathtt{isTamed(balto)},\\ & \mathtt{hasCollar(pluto)},\ \mathtt{hasLegs(pluto)},\ \mathtt{isTamed(pluto)},\\ & \mathtt{hasLegs(alberto)},\ \mathtt{Hunts(alberto)}\\ & \neg\mathtt{Trusted(cerberus)} \},\end{aligned}$$

$$\begin{aligned}\mathcal{P} = \{\, & \mathtt{Wolf(livesInWoods:10,\ hasLegs:4,\ livesInPack:8,\ Hunts:11)},\\ & \mathtt{Dog(hasCollar:33,\ livesInHouse:22,\ hasLegs:11,\ isTamed:44)} \,\}\end{aligned}$$

Below we give a semantics for this kind of PKB which will entail and justify the conclusion that `balto` *is a trusted dog which is a wolf, without being inconsistent, and that* `cerberus` *is an exceptional dog with respect to the property of dogs of being trusted. Note that in the case of the instances* `pluto` *and* `alberto` *no contradiction arises, thus we want that the axioms in $\mathcal{T}$ are applied to them normally. Observe that the conflict regarding* `balto` *has the same structure of the so called* Nixon diamond. ◇

2.2. Semantics: Justified Models

The semantics of PKBs is based on standard interpretations for the underlying DL $\mathcal{L}_\Sigma$. In fact, interpretations of a PKB are DL interpretations for its knowledge base part.

Definition 4 (Interpretation). *An* interpretation *$\mathcal{I}$ is pair $\langle \Delta^\mathcal{I}, \cdot^\mathcal{I} \rangle$ for signature Σ with a non-empty domain, $\Delta^\mathcal{I}$, $a^\mathcal{I} \in \Delta^\mathcal{I}$ for every $a \in \mathrm{NI}$, $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$ for every $A \in \mathrm{NC}$, $R^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}$ for every $R \in \mathrm{NR}$, and where the extension of complex concepts is defined recursively as usual for language $\mathcal{L}_\Sigma$.*

We are not giving a DL interpretation to the prototype definition expressions in $\mathcal{P}$, however, we need to introduce additional semantic structure to manage exceptions to prototype axioms in T_P, exploiting the prototype definition expressions in $\mathcal{P}$. We consider the notion of axiom instantiation as defined in [10]: intuitively, for an axiom $\alpha \in \mathcal{L}_\Sigma$ the *instantiation* of α with $e \in \mathrm{NI}$, written $\alpha(e)$, is the specialization of α to e.[1] In other words, we "apply" the axiom to the individual e, e.g. if the axiom means that dogs are trusted and we instantiate it to Balto, we are saying that if Balto is a dog, then it is trusted.

Definition 5 (Exception assumptions and clashing sets). *An* exception assumption *is a pair $\langle \alpha, e \rangle$ where $\alpha \in T_P$ is a prototype axiom, $e \in \mathcal{A}$ is an individual name and such that $\alpha(e)$ is an axiom instantiation of α.*

A clashing set *for $\langle \alpha, e \rangle$ is a satisfiable set $S_{\langle \alpha, e \rangle}$ of ABox assertions s.t. $S_{\langle \alpha, e \rangle} \cup \{\alpha(e)\}$ is unsatisfiable.*

[1] As in [10], $\alpha(e)$ can be formally specified via the FO-translation of α.

Intuitively, an exception assumption $\langle P \sqsubseteq D, e\rangle$ states that we assume that e is an exception to the prototype axiom $P \sqsubseteq D$ in a given interpretation. Then, the fact that a clashing set $S_{\langle P \sqsubseteq D, e\rangle}$ for $\langle P \sqsubseteq D, e\rangle$ is derived by such an interpretation gives a "justification" of the validity of the assumption of overriding in terms of ABox assertions. This intuition is reflected in the definition of models: we first extend interpretations with a set of clashing assumptions.

Definition 6 ($\boldsymbol{\chi}$-interpretation). *A* $\boldsymbol{\chi}$-interpretation *is a structure* $\mathcal{I}_{\boldsymbol{\chi}} = \langle \mathcal{I}, \chi\rangle$ *where* $\mathcal{I}$ *is an interpretation and* χ *is a set of exception assumptions.*

Then, $\boldsymbol{\chi}$-models for a PKB $\mathfrak{K}$ are those $\boldsymbol{\chi}$-interpretations that verify "strict" axioms in T_C and defeasibly apply prototype axioms in T_P (excluding the exceptional instances in χ).

Definition 7 ($\boldsymbol{\chi}$-model). *Given a PKB* $\mathfrak{K}$*, a* $\boldsymbol{\chi}$*-interpretation* $\mathcal{I}_{\boldsymbol{\chi}} = \langle \mathcal{I}, \chi\rangle$ *is a* $\boldsymbol{\chi}$-model *for* $\mathfrak{K}$ *(denoted* $\mathcal{I}_{\boldsymbol{\chi}} \models \mathfrak{K}$*), if the following holds:*

(i) for every $\alpha \in T_C \cup \mathcal{A}$ *of* $\mathcal{L}_\Sigma$, $\mathcal{I} \models \alpha$;

(ii) for every $\alpha = P \sqsubseteq D \in T_P$*, if* $\langle \alpha, d\rangle \notin \chi$*, then* $\mathcal{I} \models \alpha(d)$.

Two DL interpretations $\mathcal{I}_1$ and $\mathcal{I}_2$ are NI-*congruent*, if $c^{\mathcal{I}_1} = c^{\mathcal{I}_2}$ holds for every $c \in$ NI. This extends to $\boldsymbol{\chi}$-interpretations $\mathcal{I}_{\boldsymbol{\chi}} = \langle \mathcal{I}, \chi\rangle$ by considering interpretations $\mathcal{I}$.

Intuitively, we say that a $\boldsymbol{\chi}$-interpretation is justified if all of its exception assumptions have a clashing set that is verified by the interpretation.

Definition 8 (Justifications). *We say that* $\langle \alpha, e\rangle \in \chi$ *is* justified *for a* $\boldsymbol{\chi}$*-model* $\mathcal{I}_{\boldsymbol{\chi}}$*, if some clashing set* $S_{\langle \alpha, e\rangle}$ *exists such that, for every* $\mathcal{I}'_{\boldsymbol{\chi}} = \langle \mathcal{I}', \chi\rangle$ *of* $\mathfrak{K}$ *that is* NI*-congruent with* $\mathcal{I}_{\boldsymbol{\chi}}$*, it holds that* $\mathcal{I}' \models S_{\langle \alpha, e\rangle}$.

A $\boldsymbol{\chi}$*-model* $\mathcal{I}_{\boldsymbol{\chi}}$ *of a PKB* $\mathfrak{K}$ *is* justified*, if every* $\langle \alpha, e\rangle \in \chi$ *is justified in* $\mathfrak{K}$.

We define the consequence from justified $\boldsymbol{\chi}$-models: $\mathfrak{K} \models_J \alpha$ if $\mathcal{I}_{\boldsymbol{\chi}} \models \alpha$ for every justified $\boldsymbol{\chi}$-model $\mathcal{I}_{\boldsymbol{\chi}}$ of $\mathfrak{K}$.

Remark. Note that there can be more than one justified model, in particular for different valid combinations of exception assumptions and justifications. As will be shown in examples, this allows reasoning by cases: scores defined over prototype definitions' values allow to define a preference over such cases.

2.3. Semantics: Prototype Score and Preference

The main intuition of prototype definitions is that each individual which is an instance of a prototype is associated with a score which denotes the "degree of typicality" of the individual with respect to the concept described by the prototype. As in [7], such a degree is computed from the prototype features that are satisfied by the instances and their score. Ideally, the prototype score of an individual allows us to determine preferences over models: for an individual, axioms on prototypes with higher score are preferred to the ones on lower scoring prototypes; thus the measure needs to be comparable across different prototypes.

Given the set of prototypes $P \in \mathcal{P}$, a family of *prototype score functions* $\{f_P\}_{P \in \mathcal{P}}$ is composed by functions $f_P : \mathrm{NI} \rightarrow \mathbb{R}$ for each prototype name $P \in \mathcal{P}$ such that every function of the family has range in a fixed interval $[x, ..., y] \in \mathbb{R}$.

Ideally, these families of functions can then be used to define preferences over models: different preference criteria can be defined, in particular, by using the results of score functions on the exceptional individuals in the exception assumptions' sets χ of $\boldsymbol{\chi}$-interpretations.

In general, we define a *preference* on exception assumption sets as a partial order $\chi_1 > \chi_2$ on the sets χ for $\mathfrak{K}$. Given two $\boldsymbol{\chi}$-interpretations $\mathcal{I}^1_{\boldsymbol{\chi}} = \langle \mathcal{I}^1, \chi_1\rangle$ and $\mathcal{I}^2_{\boldsymbol{\chi}} = \langle \mathcal{I}^2, \chi_2\rangle$, we say that $\mathcal{I}^1_{\boldsymbol{\chi}}$ is *preferred* to $\mathcal{I}^2_{\boldsymbol{\chi}}$ (denoted $\mathcal{I}^1_{\boldsymbol{\chi}} > \mathcal{I}^2_{\boldsymbol{\chi}}$) if $\chi_1 > \chi_2$.

Finally, we define the notion of PKB model as a minimal justified model for the PKB.

Definition 9 (PKB model). *An interpretation $\mathcal{I}$ is a* PKB model *of $\mathfrak{K}$ (denoted, $\mathcal{I} \models \mathfrak{K}$) if*

- *$\mathfrak{K}$ has some justified $\boldsymbol{\chi}$-model $\mathcal{I}_\chi = \langle \mathcal{I}, \chi \rangle$.*
- *there exists no justified $\mathcal{I}'_{\boldsymbol{\chi}} = \langle \mathcal{I}', \chi' \rangle$ that is preferred to $\mathcal{I}_{\boldsymbol{\chi}}$.*

The consequence from PKB models of $\mathfrak{K}$ (denoted $\mathfrak{K} \models \alpha$) characterizes the "preferred" consequences of the PKB, on the basis of the degree of typicality of instances.

2.4. Semantics: Model Independent Prototype Score

A simple score function can be defined by considering the features that are inferable from the KB (in all justified models):

Definition 10 ((Model independent) Prototype score). *Given a prototype definition $P(C_1 : w_1, ..., C_m : w_m)$, we define the* (model independent) score function $score_P : \mathrm{NI} \to \mathbb{R}$ *for prototype P as:*

$$score_P(a) = \sum_{\mathfrak{K} \models_J C_i(a)} w_i$$

This measure, however, depends on the value interval over which the prototype weights have been defined: in order to compare the score of an individual with scores relative to other prototypes, this value needs to be normalized. We do so by computing the maximum score max_P and minimum score min_P for all prototypes.

The maximum score max_P denotes the score of the maximum value of $score_P$ obtainable from the weights of consistent subset of features of P. Formally, given a prototype definition $P(C_1 : w_1, ..., C_m : w_m)$, let $\boldsymbol{S}_P$ be the set of sets $S_P \subseteq \{C_1, \ldots, C_n\}$ s.t. $S_P \cup \mathfrak{K}$ is consistent. Then:

$$max_P = \max(\sum_{C_i \in S_P} w_i \mid S_P \in \boldsymbol{S}_P)$$

The minimal score min_P denotes the sum of the weights for “unavoidable” features, namely those that are strictly implied by the membership to the prototype concept. Formally:

$$min_P = \sum_{\mathfrak{K} \models_J P \sqsubseteq C_i} w_i$$

A normalized score function $nscore_P$ can be derived from $score_P$ as:

$$nscore_P(a) = \frac{score_P(a) \; - \; min_P}{max_P \; - \; min_P}$$

Consider that with such normalisation we obtain a family of prototype score functions with range in the same interval $[0, 1]$, allowing for comparison of prototype scores on the same individual.

We can then define a simple preference using the model independent score defined above: we prefer justified $\boldsymbol{\chi}$-models where the *exceptions* appear on elements of the *less* scoring prototypes.

Definition 11 (Preference SP). *$\chi_1 > \chi_2$ if, for every $\langle P \sqsubseteq D, e\rangle \in \chi_1 \setminus \chi_2$ such that there exists a $\langle Q \sqsubseteq E, e\rangle \in \chi_2 \setminus \chi_1$ with $\mathfrak{K} \cup \{D(e), E(e)\}$ unsatisfiable, it holds that $nscore_P(e) < nscore_Q(e)$.*

The intuition behind the condition $\mathfrak{K} \cup \{D(e), E(e)\}$ unsatisfiable is that we want to make the comparison between the clashing assumptions that are directly in conflict.

Example 2. *Considering the PKB reported in the example above, assume to have two PKB interpretations $\mathcal{I}^1$ and $\mathcal{I}^2$ associated respectively with the following two sets of clashing assumptions*

$$\chi_1 = \{\langle \texttt{Wolf} \sqsubseteq \neg\texttt{Trusted}, \texttt{balto}\rangle, \langle \texttt{Dog} \sqsubseteq \texttt{Trusted}, \texttt{cerberus}\rangle\} \text{ and}$$

$$\chi_2 = \{\langle \texttt{Dog} \sqsubseteq \texttt{Trusted}, \texttt{balto}\rangle, \langle \texttt{Dog} \sqsubseteq \texttt{Trusted}, \texttt{cerberus}\rangle\}.$$

We have now two $\boldsymbol{\chi}$*-interpretations corresponding to* $\langle \mathcal{I}^1, \chi_1\rangle$ *and* $\langle \mathcal{I}^2, \chi_2\rangle$. *Assuming that they are also* $\boldsymbol{\chi}$-models, *we can check if the two are also* justified. *Since, the clashing assumptions have the following clashing sets, respectively* $\{\texttt{Wolf(balto)}, \texttt{Trusted(balto)}, \texttt{Dog(cerberus)}, \neg\texttt{Trusted(cerberus)}\}$ *for the clashing assumptions in* χ_1 *and* $\{\texttt{Dog(balto)}, \neg\texttt{Trusted(balto)}, \texttt{Dog(cerberus)}, \neg\texttt{Trusted(cerberus)}\}$ *for those in* χ_2*, they are both justified.*

In order to decide which model is preferred, we need to compute the prototype scores for `balto` *and for* `cerberus`*: we have* $score_{Wolf}(balto) = 14$, $score_{Dog}(balto) = 55$, $score_{Dog}(cerberus) = 11$. *Then we need to normalise them:*

$$nscore_{Dog}(balto) = \frac{score_{Dog}(balto) \ - \ min_{Dog}}{max_{Dog} \ - \ min_{Dog}} \approx 0,4$$

$$nscore_{Wolf}(balto) = \frac{score_{Wolf}(balto) \ - \ min_{Wolf}}{max_{Wolf} \ - \ min_{Wolf}} \approx 0,3$$

$$nscore_{Dog}(cerberus) = \frac{score_{Dog}(cerberus) \ - \ min_{Dog}}{max_{Dog} \ - \ min_{Dog}} = 0$$

Consequently $score_{Wolf}(balto) < score_{Dog}(balto)$ *and, since* $\langle \texttt{Dog} \sqsubseteq \texttt{Trusted}, \texttt{cerberus}\rangle$ *is present in both* χ_1 *and* χ_2 *so it does not influence the preference order, then we can conclude that* $\chi_1 > \chi_2$. *This means that the preferred model, i.e. the PKB model, is* $\mathcal{I}^1$ *where* `balto` *is an exception to* $\texttt{Wolf} \sqsubseteq \neg\texttt{Trusted}$ *and* `cerberus` *is an exception to* $\texttt{Dog} \sqsubseteq \texttt{Trusted}$. *Consequently, it holds that* $\mathfrak{K} \models \texttt{Trusted(balto)}$ *and* $\mathfrak{K} \models \neg\texttt{Trusted(cerberus)}$.

Moreover, we can note that for `pluto` *and* `alberto` *we can standardly infer* `Trusted(pluto)` *and* $\neg$`Trusted(alberto)`. *The reason is that the clashing assumptions are referred to specific individuals, and since there are no contradicting assertions for* `pluto` *and* `alberto`*, there are no clashing sets that justify their assumptions as exceptions. Therefore, axioms in* $\mathcal{T}$ *apply to them standardly.* ◇

Remark. For simplicity, we are assuming independence of scores across the features: for example, the weight of a feature *hasWhiteTail* is not dependent on the weight of a more general *hasTail*. Dependence across features and their impact on the evaluation of weights is indeed an interesting extension to our work and we plan to provide a characterization in our future work.

3. A New Preference Order

The preference relation defined above may be too coarse-grained with respect to some cases. For instance, consider the following example which is a modified version of the example considered above:

Example 3. *For the concepts* `Dog` *and* `Wolf` *we use here the letters* D *and* W *respectively and for* `Trusted` *we use* $\boldsymbol{T}$. *Moreover,* `balto` *is here simplified to* b.
Now, we can imagine to have two new prototype axioms talking of house animals (HA) *and wild animals* (WA)*, where the first are considered docile* (DC)*, while the second not docile. So, we now have the four prototype axioms*

$$D \sqsubseteq \boldsymbol{T},\ W \sqsubseteq \neg\boldsymbol{T},\ HA \sqsubseteq DC \text{ and } WA \sqsubseteq \neg DC;$$

the individual causing the conflict because it is an instance of D, W, DA *and* WA*:*

$$D(b), W(b), HA(b) \text{ and } WA(b);$$

and the four prototype descriptions, with the third one that includes $\boldsymbol{T}$ *among its features:*

$$D(A:2,\ B:8),\ W(C:4,\ H:6), HA(\boldsymbol{T}:3,\ E:7) \text{ and } WA(F:8,\ G:2).$$

where the features A, B, C, H, E, F, G do not have a specific meaning. Moreover, we know that $B(b)$, $C(b)$, $E(b)$ and $F(b)$.

From this KB we can see that we have four sets of justified exception assumptions, that is the sets of exception assumptions which have an associated clashing set:

$$\chi_1 = \{\langle D \sqsubseteq \boldsymbol{T}\rangle,\ b >,\ \langle HA \sqsubseteq DC,\ b\rangle\}$$

$$\chi_2 = \{\langle D \sqsubseteq \boldsymbol{T},\ b\rangle,\ \langle WA \sqsubseteq \neg DC,\ b\rangle\}$$

$$\chi_3 = \{\langle W \sqsubseteq \neg\boldsymbol{T},\ b\rangle,\ \langle HA \sqsubseteq DC,\ b\rangle\}$$

$$\chi_4 = \{\langle W \sqsubseteq \neg\boldsymbol{T},\ b\rangle,\ \langle WA \sqsubseteq \neg DC,\ b\rangle\}$$

We can now compute the normalised typicality scores, which are respectively:

$$nscore_D(b) = 0,8;\ nscore_W(b) = 0,4;$$
$$nscore_{HA}(b) = 0,7;\ nscore_{WA}(b) = 0,8.$$

Consequently, the order we have on sets of exception assumptions is:

$$\begin{array}{cc} \multicolumn{2}{c}{\chi_3} \\ \hline \chi_4 & \chi_1 \\ \hline \multicolumn{2}{c}{\chi_2} \end{array}$$

from which, $\mathcal{I}_{\chi_3}$ results to be the preferred model.

We have two key observations regarding this example: first, $\boldsymbol{T}(b)$ is not considered in the computation of the scores because $\mathfrak{K} \not\models \boldsymbol{T}(b)$. In fact, in $\mathcal{I}_{\boldsymbol{\chi}_1}$ and $\mathcal{I}_{\boldsymbol{\chi}_2}$ we have $\neg\boldsymbol{T}(b)$.
Second, the fact that the preferred model is the one where $\boldsymbol{T}(b)$ and $\neg DC(b)$ hold is counter-intuitive. The reason is that if we can conclude $\boldsymbol{T}(b)$ in a model, it should mean that we can add the weight associated with that feature in the prototype description of HA. Consequently, b would result as a more typical HA than WA, and so we would like to conclude $DC(b)$. Consequently, the desired interpretation would be $\mathcal{I}_{\boldsymbol{\chi}_4}$. ◇

The problem derives from the use of consequence to define the score of individuals: while this assures a uniform score across the models, this score does not consider the satisfaction of features in the single interpretations. Thus, to deal with such cases, we can define a new preference order, which considers what holds inside the models and consequently can be called *model-dependent.*

The first step is to change the definition of prototype score in order to have a different score for each model:

Definition 12 ((Model dependent) Prototype score). *Given a prototype definition $P(C_1 : w_1, ..., C_m : w_m)$, we define the* score function $score^P_{\mathcal{I}_{\boldsymbol{\chi}_j}} : \mathrm{NI} \to \mathbb{R}$ *for prototype P and a justified $\boldsymbol{\chi}$ model $\mathcal{I}_{\boldsymbol{\chi}_j}$ as:*

$$score^P_{\mathcal{I}_{\boldsymbol{\chi}_j}}(a) = \sum_{\mathcal{I}_{\boldsymbol{\chi}_j} \models C_i(a)} w_i$$

We leave the other steps of the computation of the typicality score as they are, such that we will have a family of normalised score functions which now are relative to the $\boldsymbol{\chi}$-models they are in, making the score an *intra-interpretations* score. The idea is to measure the typicality of the individual according to the hypothetical situation we are considering, that is according to the hypotheses regarding what is exceptional and especially to what it is exceptional.
In fact, remember that a $\boldsymbol{\chi}$-model is a DL interpretation with an associated set of hypothetical exceptions. This would precisely address the problem arising in the case above, since if we are supposing that b is exceptional with respect to $W \sqsubseteq \neg\mathbf{T}$, we should assume $\mathbf{T}(b)$. However, now the problem is how to compare the scores, which are dependent from interpretations, in a consistent and meaningful way with respect to the role that the comparison has in our system. This role is that of ordering those

interpretations with the goal to find out which hypothetical exceptions are reasonably actual exceptions. In the first formalisation, we relied on comparing typicality scores that were independent from the different interpretations and hypotheses and so were, so to say, the outcome of the strict knowledge that is certain in all the interpretations. Therefore, the typicality scores could be compared consistently since they depended on the same knowledge.

Endorsing this intuition, also in the new preference order we can compare the scores that do not change across the interpretations. The reason is that, since they do not change it means that they are independent from the particular interpretation and so we are in the same position as in the model-independent preference. From a formal perspective, this means that the computation of the typicality scores do not depend on other prototype axioms. Then, we can order the interpretations using the existing preference function slightly adjusted to consider only the scores that do not change across the models, individuating in this way the set of the local preferred models, that is those that are preferred with respect of only the strict knowledge. At this point, we can apply recursively the previous step, that is comparing the scores that do not change across these locally preferred models and order them according to the preference function we have, individuating the set of the new locally preferred models. We continue with this method iteratively until we remain with the set of globally preferred models.

Now we can give a more precise definition of this new preference order. Firstly, we need a definition of the scores we would consider:

Definition 13 (Stable score). *Given a set $M = \{\mathcal{I}_{\boldsymbol{\chi}} \mid \mathcal{I}_{\boldsymbol{\chi}}$ is a justified $\boldsymbol{\chi}$-model$\}$, $score_P(a)$ is a stable score iff $\forall \mathcal{I}^i_{\boldsymbol{\chi}}, \mathcal{I}^j_{\boldsymbol{\chi}} \in M \ score^P_{\mathcal{I}^i_{\boldsymbol{\chi}}}(a) = score^P_{\mathcal{I}^j_{\boldsymbol{\chi}}}(a)$*

Now, we can define the new preference mechanism by modifying the previous one with the addition of the constraint that we are comparing only the scores that are stable across all the justified $\boldsymbol{\chi}$-models.

Definition 14 (Local Preference). *$\chi_1 > \chi_2$ if, for every $\langle P \sqsubseteq D, e\rangle \in \chi_1 \setminus \chi_2$ such that there exists a $\langle Q \sqsubseteq E, e\rangle \in \chi_2 \setminus \chi_1$ with $\mathfrak{K} \cup \{D(e), E(e)\}$ unsatisfiable and such that $score_P(e)$ and $score_Q(e)$ are stable scores, it holds that $nscore_P(e) < nscore_Q(e)$.*

As before, a preferred justified $\boldsymbol{\chi}$-model is a justified $\boldsymbol{\chi}$-model that has no justified $\boldsymbol{\chi}$-model which is preferred to it: formally, a justified $\boldsymbol{\chi}$-model $\mathcal{I}_{\boldsymbol{\chi}} = \langle \mathcal{I}, \chi\rangle$ is a *locally preferred justified $\boldsymbol{\chi}$-model* of $\mathfrak{K}$ if there exists no justified $\mathcal{I}_{\boldsymbol{\chi}'} = \langle \mathcal{I}', \chi'\rangle$ that is preferred to $\mathcal{I}_{\boldsymbol{\chi}}$. Therefore, we can define the set of these preferred models:

Definition 15 (Set of locally preferred justified $\boldsymbol{\chi}$-models). *We denote the set of locally preferred justified $\boldsymbol{\chi}$-models of a set of justified $\boldsymbol{\chi}$-models M of $\mathfrak{K}$ with $\mu(M)$.*

Note that $\mu(M) \subseteq M$ and so $\mu(M) \in \mathcal{P}(M)$, where $\mathcal{P}(M)$ stands for the power-set of M.

Now we are ready to define the global preference between the models, thanks to an iterative application of the local preference:

Definition 16 (Global Preference). *Given the set $\mathcal{M}$ of all the justified $\boldsymbol{\chi}$-models of $\mathfrak{K}$, consider the sequence of sets of models M_0, ..., M_n where $M_i \subseteq \mathcal{M}$ such that (i) $M_0 = \mathcal{M}$; (ii) $M_{i+1} = \mu(M_i)$; (iii) M_n is the fixed point such that $\mu(M_i) = M_i$.*

A justified $\boldsymbol{\chi}$-model $\mathcal{I}_{\boldsymbol{\chi}}$ is a globally preferred model *of $\mathfrak{K}$ iff $\mathcal{I}_{\boldsymbol{\chi}} \in M_n$*

Proposition 1. *The global preference construction above has a fixed point M_n.*

Proof. Assume that there is no fixed point. This can happen in two ways: either (i) there are infinitely many M_i, or (ii) there is a loop such that $M_{i+k} = M_i$ where $k > 1$.
Consider situation (i): we can notice that $M_0 \supseteq M_i \supseteq M_{i+1}$ and so on *ad infinitum* since we never produce new justified $\boldsymbol{\chi}$-models, but we select among the elements of the ith-set those that are preferred and we use them to build the new ith+1-set. However, this selection depends only on the $\boldsymbol{\chi}$s of the justified $\boldsymbol{\chi}$-models, that we recall are sets of exception assumptions. Since the latter are defined on

axioms and individual names in $\mathfrak{K}$, the exceptional assumptions are finite and consequently also the χs. Therefore, there cannot be infinitely many justified $\boldsymbol{\chi}$-models.

Now, consider situation (ii). A loop would have a form like this: $\mu(M_i) = M_{i+1}$; $\mu(M_{i+1}) = M_{i+2} = M_{i-1}$ and $\mu(M_{i-1}) = M_i$.

Since $\forall i(M_i \supseteq M_{i+1})$, $M_i \supseteq M_{i+1} \supseteq M_{i-1} \supseteq M_i$. But this means that $M_i = M_{i+1}$ and this is inconsistent with the assumption. □

By considering the globally preferred models as those preferred *tout court* for the Definition 9 of the PKB models, we now have a new preference order which allows to reach the desired conclusion in cases like those in Example 3 above. To illustrate that this is the case and how the new method works, we can apply the new preference to the example above:

Example 4. *Consider the knowledge base presented in Example 3, we have the same four exception assumptions* $\boldsymbol{\chi}_1$, $\boldsymbol{\chi}_2$, $\boldsymbol{\chi}_3$ *and* $\boldsymbol{\chi}_4$. *So, now we can start applying the new preference order: we have the elements of our set* $M_0 = \mathcal{M} = \{\mathcal{I}_{\boldsymbol{\chi}_1}, \mathcal{I}_{\boldsymbol{\chi}_2}, \mathcal{I}_{\boldsymbol{\chi}_3}, \mathcal{I}_{\boldsymbol{\chi}_4}\}$.
Then, we can compute the normalised typicality scores, but now each justified $\boldsymbol{\chi}$*-model will have its set of typicality scores thanks to Definition 12, which are respectively:*

	$\mathcal{I}_{\boldsymbol{\chi}_1}$	$\mathcal{I}_{\boldsymbol{\chi}_2}$	$\mathcal{I}_{\boldsymbol{\chi}_3}$	$\mathcal{I}_{\boldsymbol{\chi}_4}$
$nscore_D(b)$	$0,8$	$0,8$	$0,8$	$0,8$
$nscore_W(b)$	$0,4$	$0,4$	$0,4$	$0,4$
$nscore_{HA}(b)$	$0,7$	$0,7$	***1***	***1***
$nscore_{WA}(b)$	$0,8$	$0,8$	$0,8$	$0,8$

Now we can apply the new definition of preference, which will compare only the stable scores. In this case the stable score are $nscore_D(b)$ *and* $nscore_W(b)$. *Note that* $\boldsymbol{\chi}_3$ *and* $\boldsymbol{\chi}_4$ *assume that Balto is exceptional with respect to wolves being not trusted and the stable score with respect to the prototype* W *is smaller than that of the prototype* D. *Therefore, the locally preferred models are* $\mathcal{I}_{\boldsymbol{\chi}_3}$ *and* $\mathcal{I}_{\boldsymbol{\chi}_4}$, *or, in other words,* $\mu(M_0) = M_1 = \{\mathcal{I}_{\boldsymbol{\chi}_3}, \mathcal{I}_{\boldsymbol{\chi}_4}\}$

In the next step, we have to select the locally preferred justified $\boldsymbol{\chi}$*-models, but in the new set* M_1. *So, now, we compare also* $nscore_{HA}(b)$ *and* $nscore_{WA}(b)$ *which are stable normalised scores in* M_1 *and we have* $\chi_4 > \chi_3$. *Therefore,* $\mu(M_1) = M_2 = \{\mathcal{I}_{\boldsymbol{\chi}_4}\}$.

Again, we search for the preferred models in M_2. *In this case, the preferred model is the only one in the set, since it is trivially true that there is no other model in the set that is preferred to it. So,* $\mu(M_2) = M_3 = \{\mathcal{I}_{\boldsymbol{\chi}_4}\} = M_2$, *which means that* M_2 *is our fixed point.*

Thus, we can conclude that $\mathcal{I}_{\boldsymbol{\chi}_4}$ *is the globally preferred justified* $\boldsymbol{\chi}$*-model and therefore the PKB model of* $\mathfrak{K}$ *as expected.*

4. Conclusions and Future Work

In this paper, we built on the work in [5] by refining the terminology, simplifying the formal setting, and by proposing a new preference mechanism that allows to overcome certain shortcomings of the previous one.

Regarding future work, we would like to further develop the formalism through a relaxation of some of the requirements found in the current version, such as the restriction that the individuals found in the exceptional assumptions must be named entities in the knowledge base. Moreover, we will study the formal properties enjoyed by the resulting logic in order to compare it with other approaches which use some notion of typicality in DLs for defeasible reasoning such as [11, 12]. A logical analysis comprehending a consistency proof and a discussion of the coherence between the knowledge base and the weights of the features is needed too, along with a study of the computational costs. Finally, we will propose an implementation of our approach in the framework of Answer Set Programming.

Acknowledgements. This work was partially supported by the GULP - Gruppo Ricercatori e Utenti Logic Programming. The support is gratefully acknowledged.

References

[1] C. Strasser, G. A. Antonelli, Non-monotonic Logic, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Summer 2019 ed., Metaphysics Research Lab, Stanford University, 2019.

[2] J. McCarthy, P. J. Hayes, Some Philosophical Problems from the Standpoint of Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987, p. 26–45.

[3] L. Giordano, V. Gliozzi, N. Olivetti, G. Pozzato, Semantic characterization of rational closure: From propositional logic to description logics, Artificial Intelligence 226 (2015) 1–33. URL: https://www.sciencedirect.com/science/article/pii/S0004370215000673. doi:https://doi.org/10.1016/j.artint.2015.05.001.

[4] K. Britz, G. Casini, T. Meyer, K. Moodley, U. Sattler, I. Varzinczak, Principles of klm-style defeasible description logics, ACM Trans. Comput. Logic 22 (2020). URL: https://doi.org/10.1145/3420258. doi:10.1145/3420258.

[5] G. Sacco, L. Bozzato, O. Kutz, Defeasible reasoning with prototype descriptions: First steps, in: Proceedings of the 36th International Workshop on Description Logics (DL 2023), volume 3515 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[6] G. Sacco, L. Bozzato, O. Kutz, Generics in defeasible reasoning. exceptionality, gradability, and content sensitivity, 2023. Accepted at 7th CAOS Workshop 'Cognition and Ontologies', 9th Joint Ontology Workshops (JOWO 2023), co-located with FOIS 2023, 19-20 July, 2023, Sherbrooke, Québec, Canada.

[7] P. Galliani, G. Righetti, O. Kutz, D. Porello, N. Troquard, Perceptron connectives in knowledge representation, in: C. M. Keet, M. Dumontier (Eds.), Knowledge Engineering and Knowledge Management, Springer International Publishing, Cham, 2020, pp. 183–193.

[8] D. Porello, O. Kutz, G. Righetti, N. Troquard, P. Galliani, C. Masolo, A toothful of concepts: Towards a theory of weighted concept combination, in: Description Logics, volume 2373 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.

[9] G. Sacco, L. Bozzato, O. Kutz, Introducing weighted prototypes in description logics for defeasible reasoning, in: A. F. Agostino Dovier (Ed.), Proceedings of the 38th Italian Conference on Computational Logic, CEUR Workshop Proceedings, Udine, Italy, 2023.

[10] L. Bozzato, T. Eiter, L. Serafini, Enhancing context knowledge repositories with justifiable exceptions, Artif. Intell. 257 (2018) 72–126.

[11] L. Giordano, D. Theseider Dupré, Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model, in: Logics in Artificial Intelligence: 17th European Conference, JELIA 2021, Virtual Event, May 17–20, 2021, Proceedings 17, Springer, 2021, pp. 225–242.

[12] K. Britz, J. Heidema, T. Meyer, Modelling object typicality in description logics, in: A. Nicholson, X. Li (Eds.), AI 2009: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 506–516.