A Fully Model-Theoretic Semantics for Model-Preference Default Systems

Fabrizio Sebastiani
Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
Via S. Maria, 46 - 56126 Pisa (Italy)
E-mail : fabrizio@icnucevm.cnuce.cnr.it

Abstract

Propositional systems of default inference based on the dyadic relation of *preference* between models have recently been proposed by Selman and Kautz to provide a computationally tractable mechanism for the generation of vivid knowledge bases. In this paper we argue that the formalism proposed, albeit endowed with a semantic flavour, is not a model-theoretic (or denotational) semantics, as no ontology which is independent of the existence of the knowledge representation language is postulated. Consistently with good model-theoretic practice we carry on to postulate a language-independent ontology and to use it in the subsequent definition of a fully model-theoretic semantics for model-preference default systems. This semantics is instrumental in providing guidelines for the development of algorithms that reason on model-preference default systems, and for comparing this with other "preferential" non-monotonic formalisms recently proposed in the literature.

1 Introduction

Default inference plays an important role in everyday practical reasoning[1]. Agents, be they natural or artificial, typically face situations in which they have to act and make decisions on the basis of a body of knowledge that is far from being an exhaustive description of the domain of discourse; their lack of such a description is a direct consequence of the limited throughput of their channels of communication with the external world (e.g. the visual apparatus), of the fact that the processes involved in the acquisition of knowledge (both from external sources—e.g. books—and internal ones—e.g. speculative reasoning) are computationally demanding and time consuming, and, above all, of the fact that the relevant information simply may not be available to the agent.

Nevertheless, action and decision-making is often so complex to require more than the knowledge the agents actually possess; this forces them to overcome the limited coverage of their knowledge bases (KBs) by making "default" assumptions which are then brought to bear in the reasoning task. As the name implies, "assumptions" are items of knowledge endowed with an epistemic status that is far from being solid: that is, they can be invalidated by further reasoning or by the subsequent acquisition of empirical data. These phenomena are well-known in cognitive science, and their lack of resemblance with deductive patterns of reasoning has sometimes been taken to imply that a great deal of human reasoning does not conform to the canons of "logic" and hence is not amenable to formalization [4].

Doubtless, the overall effectiveness of human action in the face of incomplete information testifies to the effectiveness of this modality of reasoning. In fact, humans are much quicker at creating surrogates of missing knowledge than at actually acquiring that knowledge in a more reliable way, either through reasoning or empirical investigation, and have the ability to generate *plausible* surrogates, surrogates that in most occasions turn out to be accurate predictions of the actual reality. Once these surrogates have been created, humans are also much quicker at reasoning on the resulting exhaustive, albeit "epistemically weaker", description of the domain of discourse than they would be if they had to rely on the smaller part of this description that they trust as accurate *tout court*. These observations are at the heart of the recent interest that the KR community has shown in *vivid knowledge bases* [2, 8, 9], i.e. exhaustive descriptions of the domain of discourse consisting of collections of atomic formulae[2]. Reasoning on these KBs, which

[1] Sections 1 and 2 draw from material first presented in [16].

[2] In formally introducing vivid KBs Levesque [8] actually situates his discussion in the framework of the first order predicate calculus; hence, for him a vivid KB is "a collection of ground, function-free atomic sentences, inequalities between all different constants (. . .), universally quantified sentences

may be considered as "analogues" of the domain being represented, is easily shown to be efficient.

It is precisely in the face of the above-mentioned empirical considerations that the bad computational properties of current formalisms that address default reasoning (such as the ones based on Circumscription [10, 11] or on Autoepistemic Logic [5, 12]) are particularly disturbing: arguably, a formalism for default reasoning not only should characterize the class of conclusions that agents draw in the presence of incomplete information, but should also possess radically better computational properties than formalisms accounting for the reasoning tasks at which humans are notoriously inefficient (such as e.g. classical logic in the case of deductive reasoning).

These considerations have lead researchers to look with special interest at formalizations of default reasoning that emphasize computational tractability. In their recent paper "The complexity of model-preference default theories" [17], Selman and Kautz describe $\mathcal{DH}_a^+$, a tractable system for performing inferences on acyclic theories of Horn defaults; in this system a vivid, complete KB may be obtained in polynomial time from an acyclic theory of Horn defaults. This tractability result accounts for what both intuition and empirical evidence suggest, namely, that in order to overcome the computational problems associated with practical reasoning and obtain KBs upon which subsequent reasoning can be carried out efficiently, humans must use a method that is itself efficient. The framework described in [17], quite similarly to other recent proposals [1, 15, 18], has the added appeal of possessing a model-theoretic flavour. However, such a framework is nevertheless a mix of different notions some of which pertain to the semantic level and some of which to the syntactic one, and does not completely live up to its promise of offering a "true semantic characterization of default inference".

In this paper we give an account of model-preference default systems that, while licensing the same set of inferences of the system described in [17], is also a fully model-theoretic (or denotational) semantics. Giving one's language a semantics that conforms to model-theoretic practice has several benefits: model-theory is a "standard" that is widely understood by the scientific community, and because of this a semantics designed along its lines provides both a valuable tool for comparison (and possibly integration) with other formalisms specified in the same style, and a machine-independent specification which implementations of the formalism must necessarily comply with.

The paper is organized as follows. In order to make it self-contained, in Section 2 we give a brief overview of $\mathcal{D}^+$, the most general system described in [17][3]. This overview is not completely faithful to the original proposal of [17] in that it incorporates the modifications that we have suggested in [16] in order to make the above mentioned systems behave correctly in the presence of both categorical information and default rules. In Section 3 we spell out in detail the proposed model theory. Quite obviously, it will not be possible to give any formal proof of equivalence of the two approaches, as such a proof would be possible only if the two approaches were formally specified in a completely denotational way: it is precisely a claim of this paper that the approach of [17] is not so. In order to allow the reader to gain a better understanding of this semantics, in Section 4 we will work out the semantics of a sample "heterogeneous default theory" (i.e. a $\mathcal{D}^+$ KB). Section 5 concludes.

2 An overview of Selman and Kautz's system $\mathcal{D}^+$

Roughly speaking, the idea around which the systems of [17] revolve is that the import of a default $d \equiv \ \alpha \rightarrow q$ is to make a model (i.e. a complete specification of what the world is like) where both α and q are true be *preferred* to another model where α is true but q is not. By combining the effects of the preferences due to the individual defaults, a set of defaults identifies a set of "maximally preferred" models; these models, isomorphic as they are to vivid KBs, are meant to represent possible ways in which the agent may "flesh out" his body of categorical (or certain) knowledge by the addition of defeasible knowledge. For instance, according to a set of defaults such as $\{a \rightarrow b,\ b \rightarrow c\}$, the model where a, b and c are all true would be a maximally preferred model. However, (some of) the systems described in [17] also account for the fact that a more specific default should override a less specific one, and they do so by "inhibiting", where a contradiction would occur, the preference induced by the less specific default; this is meant to prevent a set of defaults such as $\{a \rightarrow b,\ b \rightarrow c,\ ab \rightarrow \neg c,\ a\neg b \rightarrow \neg c\}$ to license maximally preferred models where a and c are both true.

The first thing we need to do in order to introduce $\mathcal{D}^+$ in formal detail is to describe what the language for representing knowledge in $\mathcal{D}^+$ is. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a finite set of propositional letters, and L be the language of formulae built up from P and the connectives $\neg$, $\wedge$ and $\vee$ in the standard way. We define a *default* d to be an expression of the form $\alpha \rightarrow q$, where q is a literal (i.e. a propositional letter p in P or its negation $\neg p$) and α is a set of literals[4]. We will also use the standard definition

expressing closed world assumptions (...) over the domain and over each predicate, and the axioms of equality". As our discussion will be situated in the framework of the propositional calculus, we will take this definition of vivid KB instead.

[3] Other systems discussed in [17], such as $\mathcal{DH}^+$ and $\mathcal{DH}_a^+$, are restrictions of $\mathcal{D}^+$ to the case of Horn defaults and to the case of acyclic sets of Horn defaults, respectively; the model theory that is illustrated in this paper applies straightforwardly to these more restricted systems.

[4] For notational convenience we will omit braces in ante-

of a *model* for L as a function $M : P \mapsto \{\texttt{True}, \texttt{False}\}$; accordingly, we will say that M satisfies a *propositional theory* (i.e. a set of formulae) T of L (written as $M \models T$) iff M assigns `True` to each formula in T, formulae being evaluated with respect to M in the standard manner.

The above-mentioned specificity ordering between defaults is captured by stipulating that, given a set of defaults (or *default theory*) D, a default $d \equiv \alpha \rightarrow q$ in D is *blocked* at a model M iff there exists a default d' in D such that $d' \equiv \alpha \cup \beta \rightarrow \neg q$ and $M \models \alpha \cup \beta$. A default $d \equiv \alpha \rightarrow q$ is then said to be *applicable* to a model M iff $M \models \alpha$ and d is not blocked at M. If d is applicable at M, the model $d(M)$ is defined as the model which is identical to M with the possible exception of the truth assignment to the propositional letter occurring in q, which is assigned a truth value such that $d(M) \models q$. Naturally enough, a preference ordering induced on models by a default theory D may at this point be defined. Given a default theory D and a propositional theory T, the relation "$\leq +$" is defined to hold between two models M and M' which both satisfy T (written $M \leq + M'$) iff there exists d in D such that d is applicable at M and such that $d(M) = M'$. The relation "$\leq$" is then defined as the transitive closure of "$\leq +$"[5]. Finally, we will say that a model M is *maximally preferred* (or *maximal*) with respect to a propositional theory T and a default theory D iff for all models M' such that $M' \models T$ either $M' \leq M$ is the case or $M \leq M'$ is not the case. We understand the task of reasoning in $\mathcal{D}^+$ as that of finding an arbitrary model which is maximal with respect to a given propositional theory T and a given default theory D. We will illustrate the way $\mathcal{D}^+$ works by means of an example.

Example 1 *Let* $P = \{a, b, c, d\}, T = \{d\}, D = \{a \rightarrow b, b \rightarrow c, ab \rightarrow \neg c, a\neg b \rightarrow \neg c, a\neg c \rightarrow \neg d\}$*. The models* `¬abcd`*,* `¬a¬bcd`*,* `ab¬cd` *and* `¬a¬b¬cd` *are all and the only maximal models. Note that if* $b \rightarrow c$ *had not been blocked at* `ab¬cd`*, then* `abcd` *would have been maximal too, contrary to intuitions.*

3 A fully denotational semantics for $\mathcal{D}^+$

In this section we give a description of model-preference default systems that overcomes the shortcomings hinted at in the introduction and completely satisfies the requirements of the denotational, or model-theoretic, approach to semantics. In this spirit, we will divide our specification into three parts:

1. a specification of the language;
2. a specification of the ontology, i.e. of the entities that exist in the domain of discourse;
3. a specification of the semantics, i.e. of the mapping from elements and phenomena belonging to the linguistic level to elements and phenomena belonging to the ontological level.

We recall that it is crucial to any model-theoretic endeavour that task 1 be accomplished without any commitment to the existence of an ontological level, while task 2 be accomplished without any commitment to the existence of a linguistic level. It is part 3 which finally establishes the link between language and reality.

3.1 Syntax

As the language we adopt is obviously no different from the one adopted in [17], this paragraph will be devoted to spelling out the language that was informally described above in a way that is consistent with the typographical conventions that we will adopt in the rest of the paper[6]. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a finite set of propositional letters, and L be the language of formulae built up from P and the connectives $\neg$, $\wedge$ and $\vee$ in the standard way.

Definition 1 *A* literal q_i *is either a propositional letter* p_i *in* P *or its negation* $\neg p_i$*. A* propositional theory T *is a set of formulae of* L*.*

We will let $q, q_1, q_2, \ldots$ be metavariables ranging over literals, $T, T_1, T_2, \ldots$ be metavariables ranging over propositional theories and $\alpha, \alpha_1, \alpha_2 \ldots$ be metavariables ranging over sets of literals.

Definition 2 *A* default d *is an expression of the form* $\alpha \rightarrow q$*. A* default theory D *is a set of defaults.*

We will let $d, d_1, d_2, \ldots$ be metavariables ranging over defaults, and $D, D_1, D_2, \ldots$ be metavariables ranging over default theories.

Definition 3 *A* heterogeneous default theory *(or* HD-theory*, for short)* H *is a pair* $\langle T, D \rangle$ *where* T *is a propositional theory and* D *is a default theory.*

We will let $H, H_1, H_2, \ldots$ be metavariables ranging over HD-theories.

3.2 Ontology

In this section we will give a description of the ontology on which our language will be interpreted, i.e. of the entities that are postulated to exist in the domain of discourse, of the relationships among them and of the phenomena that involve them. That a description of an

cedents of defaults. Hence we will write e.g. $ab \rightarrow \neg c$ instead of $\{a, b\} \rightarrow \neg c$.

[5] Selman and Kautz [17] define "$\leq$" to be the *reflexive* transitive closure of "$\leq +$"; that this is redundant may be seen by inspecting the way "$\leq$" is used in the definition of maximal model. Also, in the definition of "$\leq +$" (and hence of "$\leq$") [17] does not require that the two models satisfy T; that this further condition should be enforced in order to implement a correct behaviour in the presence of both categorical knowledge and defeasible knowledge is argued in [16].

[6] Throughout this paper symbols printed in boldface will denote entities belonging to the ontology.

ontology be free from any reference to the existence of a language L is a requirement that should necessarily be met in any model-theoretic account of L itself; meeting this requirement has been one of the leading motivations behind our modifications to the account of [17].

Definition 4 *An* n-interpretation $\mathbf{m_n}$ *is a n-uple* $\langle \mathbf{b_1}, \ldots, \mathbf{b_n} \rangle$*, where each* $\mathbf{b_i}$ *may be either* $\mathbf{T}$ *or* $\mathbf{F}$.

Henceforth, we will take the freedom to drop the prefix "n-" and simply speak of *interpretations* when no confusion about their cardinality could arise. We will let $\mathbf{m}, \mathbf{m_1}, \mathbf{m_2}, \ldots$ be metavariables ranging over interpretations and $\mathbf{M}, \mathbf{M_1}, \mathbf{M_2}, \ldots$ be metavariables ranging over sets of interpretations[7].

We may think of an interpretation as a set of truth conditions specifying whether some facts of interest hold or not: for example, if $n = 2$ and the domain of discourse we are focussing on has to do with Opus being or not a penguin and Binkley being or not a football player, the interpretation $\langle \mathbf{T}, \mathbf{F} \rangle$ may be thought of as the state of affairs in which the fact informally described by the natural language sentence "Opus is a penguin" holds while the one described by "Binkley is a football player" does not. Note that we have used these sentences just in order to convey an intuitive feeling for what interpretations are: things do no need to be talked of (or represented) in order to exist and, as the absence of any reference to syntactic entities in the preceding definition shows, the existence of interpretations is independent from the existence of a language that uses them as referents[8].

Definition 5 *A* local-preference function $\mathbf{d}$ *on a set of interpretations* $\mathbf{M}$ *is a partial function from* $\mathbf{M}$ *to* $\mathbf{M}$*. When* $\mathbf{m}$ *belongs to the domain of* $\mathbf{d}$ *we will also say that* $\mathbf{d}(\mathbf{m})$ *is* locally preferrable *to* $\mathbf{m}$.

We will let $\mathbf{d}, \mathbf{d_1}, \mathbf{d_2}, \ldots$ be metavariables ranging over local-preference functions and $\mathbf{D}, \mathbf{D_1}, \mathbf{D_2}, \ldots$ be metavariables ranging over sets of such functions.

A local-preference function is the ontological entity which, in Section 3.3, will be put in correspondence with a default (i.e. will be the semantics of a default). Note that the notion of blocking as from [17] has not entered yet: this is because we want local-preference functions to be the "ontological cornerstones" on which to actually build this notion. In fact, blocking will be defined on local-preference functions, and "global preference" (see below) will be defined more or less as local preference *modulo* blocking. Interestingly enough, local preference functions do not have neat correspondents in other well-known model theories in philosophical logic. Although at first sight they might resemble the "accessibility relations" of possible worlds semantics for modal logic [6, 7], the differences are more striking than the similarities: note for instance that an accessibility relation is an integral part of a single interpretation of a modal language (i.e. is a component of a Kripke structure), while a local-preference function is "external" to interpretations of our propositional language (actually, it is a dyadic relationship between them).

Definition 6 *A* model-preference default structure *(or simply* structure*)* $\mathbf{S}$ *is a pair* $\langle \mathbf{M}, \mathbf{D} \rangle$*, where* $\mathbf{M}$ *is a set of n-interpretations and* $\mathbf{D}$ *is a set of local-preference functions on* $\Pi(n)$*, the set of all n-interpretations.*

A structure is the ontological setting on which an HD-theory $H = \langle T, D \rangle$ will be interpreted. Note that the local-preference functions may be defined also on interpretations that do not belong to $\mathbf{M}$; we will need this (as will become apparent in Section 3.3) in order to define the semantics of D independently of the semantics of T.

Definition 7 *Given a structure* $\mathbf{S} = \langle \mathbf{M}, \mathbf{D} \rangle$*, two local-preference functions* $\mathbf{d_i} \in \mathbf{D}$ *and* $\mathbf{d_j} \in \mathbf{D}$ *and an interpretation* $\mathbf{m} \in \mathbf{M}$*, we will say that* $\mathbf{d_i}$ *is* blocked *by* $\mathbf{d_j}$ *at* $\mathbf{m}$ *iff*

- $\mathbf{m}$ *belongs to both the domain of* $\mathbf{d_i}$ *and the domain of* $\mathbf{d_j}$*, and*
- *the domain of* $\mathbf{d_j}$ *is a subset of the domain of* $\mathbf{d_i}$*, and*
- *the ranges of* $\mathbf{d_i}$ *and* $\mathbf{d_j}$ *are disjoint.*

The notion of blocking given here corresponds, in spirit, to the one given in [17]. However, unlike in our work, Selman and Kautz define the blocking of a default (*viz.* of a syntactic object) at a model (*viz.* an object of the ontology). In our definition, instead, all entities that are involved in the phenomenon of blocking are of an ontological nature; quite naturally, they are precisely the real-world entities that, in the description of the semantics given in Section 3.3, we will make correspond to the linguistic objects that had been used in the definition of [17]. In particular, that $\mathbf{m}$ belongs to both the domain of $\mathbf{d_i}$ and the domain of $\mathbf{d_j}$ corresponds to the original requirement that $\mathbf{m}$ satisfies the antecedents of both d_i and d_j[9]; that the domain of $\mathbf{d_j}$ is a subset of the domain of $\mathbf{d_i}$ corresponds to the original requirement that the antecedent of d_j be a *super*set of the antecedent of d_i, and that the ranges of $\mathbf{d_i}$ and $\mathbf{d_j}$ are disjoint is our rendition of the consequents of d_i and d_j being one the negation of the other.

Definition 8 *Given a structure* $\mathbf{S} = \langle \mathbf{M}, \mathbf{D} \rangle$ *and two interpretations* $\mathbf{m} \in \mathbf{M}$ *and* $\mathbf{m}' \in \mathbf{M}$*, we will say that*

[7] Also, we will use $\mathbf{b_1 b_2 \ldots b_n}$ as shorthand for $\langle \mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b_n} \rangle$.

[8] In adopting this terminology we take a slight detour from the one adopted in [17], and reserve the word *model* (of T) for an interpretation that satisfies a propositional theory T. Consistently with model-theoretic terminology, the term "interpretation" will then have an *ontological* connotation, while the term "model" will have a *semantic* one.

[9] Here we take d_i and d_j to be the defaults that, in the original definition, play the role that in our definition is played by the two local-preference functions $\mathbf{d_i}$ and $\mathbf{d_j}$.

$\mathbf{m'}$ *is* simply globally preferrable *to* $\mathbf{m}$ *(written* $\mathbf{m'} \geq + \mathbf{m}$*) if there exists a local-preference function* $\mathbf{d_i} \in \mathbf{D}$ *such that* $\mathbf{d_i}(\mathbf{m}) = \mathbf{m'}$ *and there exists no* $\mathbf{d_j} \in \mathbf{D}$ *such that* $\mathbf{d_i}$ *is blocked by* $\mathbf{d_j}$ *at* $\mathbf{m}$*. Given a structure* $\mathbf{S} = \langle \mathbf{M}, \mathbf{D} \rangle$*,* $\mathbf{m'}$ *is* globally preferrable *to* $\mathbf{m}$ *(written* $\mathbf{m'} \geq \mathbf{m}$*) iff there are interpretations* $\mathbf{m_1}, \ldots, \mathbf{m_n}$ *with* $\mathbf{m} = \mathbf{m_1}$ *and* $\mathbf{m'} = \mathbf{m_n}$ *such that* $\mathbf{m_{i+1}} \geq + \mathbf{m_i}$ *for all* $i = 1, \ldots, n-1$*.*

Hence, "$\geq +$" is just the union of all local-preference functions *modulo* blocking, and "$\geq$" is its transitive closure; "$\geq$" is similar in spirit to the strict partial orders discussed by Shoham [18] in the context of his "preferential logics", although "$\geq$" itself is not a strict partial order (in fact, it is neither irreflexive nor antisymmetric). Note that, according to the definition of "$\geq +$" (and, by consequence, to the definition of "$\geq$" too), the two interpretations are required to belong to $\mathbf{M}$: in fact, the way these definitions will be used in giving the semantics of an HD-theory will require this condition in order to neglect elements of "$\geq +$" and "$\geq$" that involve interpretations not satisfying T.

Definition 9 *Given a structure* $\mathbf{S} = \langle \mathbf{M}, \mathbf{D} \rangle$ *and an interpretation* $\mathbf{m} \in \mathbf{M}$*, we will say that* $\mathbf{m}$ *is* maximally preferrable *(or* maximal*) iff for all* $\mathbf{m'} \in \mathbf{M}$ *either* $\mathbf{m} \geq \mathbf{m'}$ *is the case or* $\mathbf{m'} \geq \mathbf{m}$ *is not the case.*

3.3 Semantics

After having specified the linguistic and the ontological level, we are ready to specify the semantics of the language, consisting of a mapping of the former level into the latter. This semantics will be *extensional*, as each linguistic expression will be mapped into an object of the ontology, which we will call its *extension* or *meaning*.

An interesting feature of this semantics is also its being strictly *compositional*; this means that the meaning of a complex linguistic object will be solely a function of the meaning of its components. This will allow the meaning of a top-level construct of our representation language (i.e. the meaning of an HD-theory) to be analyzed in a tree-like fashion, as acquiring its meaning from the meanings of its immediate sub-components and, in turn, from the meanings of its ultimate constituents (i.e. defaults and atomic propositions).

Definition 10 *We define the* extension of a propositional letter p_i *as the set* $\mathbf{M}$ *of interpretations* $\mathbf{m} = \langle \mathbf{b_1}, \ldots, \mathbf{b_n} \rangle$ *such that* $\mathbf{b_i} = \mathbf{T}$*. The* extension of a formula α *of* L *and the* extension of a propositional theory T *are then defined in the obvious way. When* $\mathbf{m}$ *is in the extension of* α *we will also say that* α *is* true *at* $\mathbf{m}$ *(in symbols:* $\mathbf{m} \models \alpha$*) and that* $\mathbf{m}$ *is a* model *of* α*.*

The preceding definitions are standard in the extensional semantics of propositional logic. Where our semantics comes in is in dealing with the extension of a default.

Definition 11 *We define the* extension of a default $d \equiv \alpha \rightarrow q_i$ *as the local-preference function* $\mathbf{d}$ *whose domain is the extension of* α *and which maps each interpretation* $\mathbf{m_j}$ *contained therein into the interpretation which is identical to* $\mathbf{m_j}$ *with the possible exception of its i-th element; this element is such that* $\mathbf{d}(\mathbf{m_j}) \models q_i$*. When a pair of interpretations* $\langle \mathbf{m}, \mathbf{m'} \rangle$*, with* $\mathbf{m'} = \mathbf{d}(\mathbf{m})$*, is in the extension of* d*, we will also say that* d *is* applicable *at* $\langle \mathbf{m}, \mathbf{m'} \rangle$*. We define the* extension of a default theory D *as the set of the extensions of the elements of* D*.*

This is roughly the analogue of the definition of "applicability" in [17]. Note that our notion of applicability is a semantic one (i.e. it has, conceptually, the same status as the notion of "truth"), as it relies on the existence of a linguistic type (namely, defaults) and of an ontological type (namely, pairs of interpretations) and puts them into correspondence, while in the original account of [17] the definition takes on a somehow hybrid status. Note that, in keeping with the compositionality requirement, we do not relativise the extension of a default theory D to that of a propositional theory T, as this relativisation is properly handled by the definition of maximality which will come into play in the following definition. This definition will also correctly handle the case when a "blocking" occurs.

Let us then define what the semantics of an HD-theory (a "knowledge base" of our representation language) is, a definition that establishes the connection among all previous work.

Definition 12 *We define the* extension of an HD-theory $H = \langle T, D \rangle$ *as the set of interpretations that are maximally preferrable with respect to the structure* $\mathbf{S} = \langle \mathbf{M}, \mathbf{D} \rangle$*, where* $\mathbf{M}$ *is the extension of* T *and* $\mathbf{D}$ *is the extension of* D*.*

Note that $\mathbf{D}$ may well involve interpretations that do not belong to $\mathbf{M}$ (i.e. that are not models of T): it is the requirement that the extension of H only contain maximally preferrable interpretations (which, in turn, brings in the requirement that simple global preference be a relation between models of T) that eventually rules out these interpretations.

A brief comment is worth regarding the meaning that this semantics attributes to a default theory. As the extension of a propositional theory T and of an HD-theory $H = \langle T, D \rangle$ is precisely a set of interpretations, their semantics is immediate and intuitive, as it relates to the semantic notion of truth, supposedly the "universal" notion with which all semantic accounts have to come to terms with. Things are quite different for a default theory D, as its extension is a set of pairs of interpretations, and this implies that the notion of truth, being monadic, is not applicable; we need instead a semantic dyadic notion, and this is precisely what we have called "applicability". This, unfortunately, does not say much, as the notion of applicability does not have the same allegedly universal status as truth; the net effect is that the semantics of a default d and of a default theory D are somehow "second-class citizens" in this semantics, as their role is only subordinate to the determination of what the extension of an HD-theory is. However, this is perfectly reasonable, because the import of a default (or of a set of them) cannot really be determined in isolation of the body of categorical knowledge that the agent possesses. It is categorical knowledge that provides the appropriate context for the determination of what default knowledge really amounts to, and it is precisely in this context (i.e. as a component of an HD-theory) that a default theory becomes a first-class citizen.

Finally, we make explicit the connection between this semantics and the task of model-preference default reasoning.

Definition 13 *A propositional theory TH consisting of literals only is an* intended theory *wrt an HD-theory $H = \langle T, D\rangle$ if the extension of TH is a set containing a single interpretation $\mathbf{m}$, and if $\mathbf{m}$ belongs to the extension of H.*

Intended theories, having a singleton as their extension, are vivid KBs; therefore, we will understand the task of a model-preference default system as that of taking an HD-theory $H = \langle T, D\rangle$ as input and returning a propositional theory TH which is intended wrt H.

4 An example

In order to give the reader a feel for what this semantic account amounts to, we will take the sample HD-theory described in Section 2 and work out what its extension according to our semantics is. Due to the compositional nature of our semantics we will be able to proceed in a completely top-down fashion.

Example 2 *The sample HD-theory we examine is $H = \langle T = \{d\}, D = \{a \rightarrow b, b \rightarrow c, ab \rightarrow \neg c, a\neg b \rightarrow \neg c, a\neg c \rightarrow \neg d\}\rangle$. Let us recall that, by Definition 12, the extension of H is the set of interpretations that are maximally preferrable with respect to the structure $\mathbf{S} = \langle \mathbf{M}, \mathbf{D}\rangle$, where $\mathbf{M}$ is the extension of T and $\mathbf{D}$ is the extension of D. By Definition 10 $\mathbf{M}$ is the set of all interpretations $\mathbf{b_1 b_2 b_3 T}$, where each of the $\mathbf{b_i}$'s is either $\mathbf{T}$ or $\mathbf{F}$. By Definition 11 $\mathbf{D}$ is the set of the extensions of the elements of D, each of these extensions being a local-preference function (i.e. a set of pairs of interpretations); for example, the extension of the default $a \rightarrow b$ stands for the set $\{\langle \mathbf{TFFT}, \mathbf{TTFT}\rangle, \langle \mathbf{TFTT}, \mathbf{TTTT}\rangle, \langle \mathbf{TTTF}, \mathbf{TTFF}\rangle, \langle \mathbf{TFTF}, \mathbf{TFFF}\rangle\}$.*

Having individuated the structure $\mathbf{S} = \langle \mathbf{M}, \mathbf{D}\rangle$, we now need to find out the set of interpretations that are maximally preferrable with respect to it; in order to do this, we have to compute the global preference relation on $\mathbf{S}$ and, in turn, the simply global preference one. By Definition 8, the latter is composed by the pairs $\langle \mathbf{m}, \mathbf{d_i}(\mathbf{m})\rangle$ such that $\mathbf{d_i}$ belongs to $\mathbf{D}$, both $\mathbf{m}$ and $\mathbf{d_i}(\mathbf{m})$ belong to $\mathbf{M}$ and such that there exists no $\mathbf{d_j} \in \mathbf{D}$ such that $\mathbf{d_i}$ is blocked by $\mathbf{d_j}$ at $\mathbf{m}$. By working upwards through Definitions 8 and 9 it is now easy to check that the set of interpretations that are maximally preferrable with respect to the structure $\mathbf{S} = \langle \mathbf{M}, \mathbf{D}\rangle$ is $\{\mathbf{FTTT}, \mathbf{FFTT}, \mathbf{TTFT}$ and $\mathbf{FFFT}\}$, which corresponds to the set of maximal models that had been individuated in Section 2 along the lines of the specification given in [17].

5 Conclusion

We have provided a semantics for model-preference default systems that is fully denotational and compositional. Such a semantics provides both a tool for comparison (and possibly integration) with other formalisms specified in model-theoretic terms, and a machine-independent specification of the model-preference formalism with which its implementations must necessarily comply. We plan to use this semantics as a framework for singling out the similarities and differences between the model-preference and other "preferential" formalisms (such as the ones described in [1, 14, 15, 18]) that have recently been proposed in the literature.

While for the purposes of this paper we have taken $\mathcal{D}^+$ as a frame of reference, the semantics we have described applies straightforwardly to $\mathcal{DH}^+$, $\mathcal{D}_a^+$ and $\mathcal{D}$, the other systems discussed in [17]. In fact, $\mathcal{DH}^+$ is obtained just by a restriction to the case of Horn defaults, $\mathcal{D}_a^+$ by restricting default theories to be acyclic, and $\mathcal{D}$ by dropping the notion of blocking; the semantics is thus identical in the case of $\mathcal{DH}^+$ and $\mathcal{D}_a^+$, while the condition on blocking has to be dropped from the definition of simple global preference to yield a semantics for $\mathcal{D}$.

Acknowledgements

This research was mainly carried out while the author was visiting the Department of Computer Science, University of Toronto. Thanks to John Mylopoulos for encouraging the work described here and for providing the peaceful research environment where this research was carried out. Thanks also to David Lauzon, Gerhard Lakemeyer, Carlo Meghini, Francesco Orilia and Dale Schurmans for reading an earlier version of this paper and providing insightful comments.

References

[1] Allen L. Brown and Yoav Shoham. New results on semantical nonmonotonic reasoning. In Michael Reinfrank, Johan De Kleer, Matthew L. Ginsberg, and Erik Sandewall, editors, *Nonmonotonic reasoning*, pages 19–26, Springer, Heidelberg, BRD, 1989.

[2] David W. Etherington, Alexander Borgida, Ronald J. Brachman, and Henry A. Kautz. Vivid knowledge and tractable reasoning: preliminary report. In *Proceedings of IJCAI-89, 10th International Joint Conference on Artificial Intelligence*, pages 1146–1152, Detroit, MI, 1989.

[3] Matthew L. Ginsberg, editor. *Readings in nonmonotonic reasoning*. Morgan Kaufmann, Los Altos, CA, 1987.

[4] Philip N. Johnson-Laird. *Mental models*. Harvard University Press, Cambridge, MA, 1983.

[5] Kurt Konolige. On the relation between default theories and autoepistemic logic. In *Proceedings of IJCAI-87, 10th International Joint Conference on Artificial Intelligence*, pages 394–401, Milan, Italy, 1987. [a] An extended version appears as "On the relation between default logic and autoepistemic theories" on *Artificial Intelligence 35*, pp. 343–382.

[6] Saul A. Kripke. Semantical analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9:67–96, 1963.

[7] Saul A. Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94, 1963. [a] Appears also in Linsky, Leonard (ed.), *Reference and modality*, Oxford, GB: Oxford University Press, 1971, pp. 63–72.

[8] Hector J. Levesque. Logic and the complexity of reasoning. *Journal of Philosophical Logic*, 17:355–389, 1988.

[9] Hector J. Levesque. Making believers out of computers. *Artificial Intelligence*, 30:81–108, 1986.

[10] John McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986. [a] Appears also in [3], pp. 153–166.

[11] John McCarthy. Circumscription - a form of nonmonotonic reasoning. *Artificial Intelligence*, 13:81–108, 1980. [a] Appears also in [3], pp. 145–152.

[12] Robert C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25:75–94, 1985. [a] Appears also in [3], pp. 127–136.

[13] Michael Reinfrank, Johan De Kleer, Matthew L. Ginsberg, and Erik Sandewall, editors. *Nonmonotonic reasoning*. Springer, Heidelberg, BRD, 1989.

[14] Piotr Rychlik. The generalized theory of model preference (preliminary report). In *Proceedings of AAAI-90, 8th Conference of the American Association for Artificial Intelligence*, pages 615–620, Boston, MA, 1990.

[15] Erik Sandewall. The semantics of non-monotonic entailment defined using partial interpretations. In Michael Reinfrank, Johan De Kleer, Matthew L. Ginsberg, and Erik Sandewall, editors, *Nonmonotonic Reasoning*, pages 27–41, Springer, 1989.

[16] Fabrizio Sebastiani. On heterogeneous model-preference default theories. In *Proceedings of CSCSI/SCEIO-90, 8th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, pages 84–91, Ottawa, Ontario, 1990.

[17] Bart Selman and Henry A. Kautz. The complexity of model-preference default theories. In *Proceedings of CSCSI/SCEIO-88, 7th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, pages 102–109, Edmonton, Alberta, 1988. [a] Appears also in [13], pp. 115–130. [b] An extended version appears as "Model-preference default theories" in *Artificial Intelligence 45*, pp. 287–322.

[18] Yoav Shoham. A semantical approach to nonmonotonic logics. In *Proceedings of IJCAI-87, 10th International Joint Conference on Artificial Intelligence*, pages 388–392, Milan, Italy, 1987. [a] Appears also in [3], pp. 227–250.