INTENDED MODELS, CIRCUMSCRIPTION and COMMONSENSE REASONING

Wtodek W. Zadrozny

IBM TJ.Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

ABSTRACT

We describe a new method of formalizing commonsense reasoning : Instead of minimizing extensions of predicates we formalize common sense as "truth in intended models", and we give a mathematical formulation of the proposed method by defining a class of intended models based on preferences in interpretations.

We propose to use problem independent natural language constraints to exclude interpretations that contradict common sense. For this reason we augment the usual, two-part formal structures consisting of a *metalevel* and an *object level,* with a third level - *a referential level.* We show that such a model is an adequate tool for generating intended interpretations, and also for problems that cannot be satisfactorily solved by circumscription.

We argue that the criticism of Hanks and McDermott (1986) does not apply to the formalization of commonsense reasoning by intended models. Namely, it is not necessary to know the consequences of the intended theory to find the right interpretation, neither needed are complex and problem specific policy axioms, which exclude some possible interpretations.

Section 1. Introduction : Commonsense Reasoning in AI

We will use the term "commonsense reasoning" in a technical sense, meaning "the drawing of (some) sound inferences that are not logicaUy valid". This means we will not talk about "real" understanding, or of "intelligence of the degree conventionally expected of adult persons in practical affairs".

There were a few attempts to formalize commonsense reasoning by extending the set of inference rules of classical logic: nonmonotonic logic of D.McDermott and J.Doyle (1980), default reasoning of R.Reiter (1980), and circumscription of J.McCarthy (1980,1986).

The idea behind circumscription (and - similarly - behind the other formalisms) can be expressed as follows:

- commonsense reasoning means ruling out possibilities not corresponding to normal situations,
- typicality is the absence of abnormality,
- formalization of common sense can be achieved by minimizing a class of abnormal objects,
- this minimization can be formally expressed as a (second order) logical formula.

All this sounds reasonable, but:

(a) the number (of types) of examples that can be handled by different circumscriptions is about the same as the number of circumscriptions themselves (predicate, formula, pointwise, variable, prioritized, protected).

(b) the paradox of Hanks and McDermott (1986) shows, the <u>impossibility</u> of expressing all commonsense in one type of circumscription ; moreover we have to take into account the <u>inconvenience</u> of describing all commonsense inferences that can be handled by circumscription as resulting from minimization.

(c) certain types of commonsense inferences do not involve minimization at all (e.g. the first example in the next section), or even require an *Open* World Assumption, (cf. Zadrozny, 1987b). And there is little hope that circumscription can be extended to cover such cases. Hence typicality is not always a simple absence of abnormality; we will argue in this paper, it can be better characterized as resulting from selection of a most plausible interpretation from a range of (not equally reasonable) possibilities.

On the other hand the above formalisms represent well at least one aspect of commonsense reasoning - its nonmonotonicity. Therefore we will show how this phenomenon can be captured in the alternative framework we propose.

We will not discuss the frame problem. Whether the formalism we propose can contribute to an analysis of this issue is an open question. It can , if there is indeed a homomorphism between the qualification problem and the frame problem.

Section 2. Intended models and principles of reference

2.1. Natural language and common sense reasoning

Example (cf.Haugeland, 1985, p. 195):

We bought the boys apples, because they were so hungry.

We bought the boys apples, because they were so cheap.

"Native speakers, however, understand such sentences so quickly that they usually don't even notice any ambiguity. One question is: How do they do that ?

Obviously, it's just common sense."

There is little hope, we think, to find a natural relationship between this example and circumscription. But, almost certainly, non-logical inferences show up in finding references of the pronouns.

The example obviously suggests a connection between natural language (NL) and commonsense reasoning. We intend to prove that there is a formalization of commonsense reasoning based on natural language. Of course, NL has something to do with common sense; we have however to find a formal and computational content of this truism. Since, as J.McCarthy (1986) pointed out, NL lacks an independent reasoning process, logic has to play a role in reasoning.

We need then a method of linking language with deduction, and we introduce it in this paper: NL (or NL programs, like grammars and dictionaries on-line) should be treated as a referential model for a logical formalization. This means that information contained in grammars, dictionaries (Webster's, Longman Dictionary of Contemporary English,...) etc. should be used to constrain possible interpretations of *logical* predicates.

But we will not discuss transformations of parsing trees into logical formulae. Neither, are we presenting a formal semantics or a theory of meaning of English sentences. This is not our aim. Rather, a program with (some) common sense should be able to make inferences not warranted by classical logic, and we propose to use linguistic constraints to ensure their soundness.

Notice that this is a way to reject certain possibilities, which , as J.Doyle (1985) has pointed out, is an attribute of commonsense inferences, whereas in logical reasoning impossibilities are eliminated.

Our approach differs from circumscription because we do not "compile from a slightly higher level non-monotonic language into mathematical logic " (McCarthy, 1986), but we force logical inferences to conform to the references of predicates, instead.

2.2. Principles of reference

We now introduce the notion of intended interpretation:

DEFINITION. Assume that a theory T is a formal description of a situation an agent is to reason about. An *intended theory of* T is the set of all formulae true in all intended models of T :

$$Ith(T) = \{ \ \phi : \ T \models_{IM} \ \phi \ \}.$$

We identify *the commonsense consequences of* T with $Ith(T)$, for which the set of intended models satisfies the following two *principles of reference* :

> *I. Predicates (of the theory* T *) do not admit arbitrary interpretations: There exists a separate logical level - the referential level - which constrains possible interpretations of the predicates.*
>
> *II. Natural language is the most important referential level for commonsense reasoning.*

These principles express our conviction that there is a certain body of knowledge underlying the common sense, and this knowledge creates restrictions on possible interpretations. Natural language is the most significant part of this body, but certainly not unique. We can illustrate this with a simple example explaining the notion of referential level, and how language can be used as such.

Example.

Consider two hypothetical expert systems, one dealing with a zoo management and the other one about a hospital management. If any of them would give information

$* \quad is \quad dead(t) \quad \& \quad is\_\_\_alive(t)$

where the individual t is either Tweety the Ostrich or F.J. Strauss, we would rightly suspect that something is wrong with this expert system.

This would not be the case if we used predicates $p\_\_17(X)$ and $q\_\_23(X)$ instead of $is\_\_dead(X)$ and $is\_\_alive(X)$. Logic does not exclude models in which ☆ is true, if only $is\_\_dead(X) \vee is\_\_alive(X)$ follows from the domain specific knowledge.

Of course , in this case we can simply add a constraint

$☆☆ \quad \neg(is\_\_dead(X) \ \& \ is\_\_alive(X))$

and such an absurd (*) is prevented. However a better solution is to make use of the fact that ** is true in all the models where "dead" and "alive" mean dead and alive, and add a rule saying that two predicates about the same object should not be true at the same time if they are antonyms. The knowledge which words are antonyms does not change. It is problem independent; for that reason - we argue - it should be treated as a referential level, constraining possible interpretations, i.e. preventing *it.* It follows also that a program does not have to know the "real" meanings of "dead" and "alive" in order to make sound conclusions.

We hope the example renders our intuitions. A formal definition of intended models and referential models will appear in Section 5.

2.3. Levels of reasoning

From the logical standpoint we have then three levels of reasoning :

* METALEVEL : specifies type of intended models,

like : minimize "abnormal(X)" , use CWA , use mental models for syllogisms (Johnson-Laird, 1980,1983).

* OBJECT LEVEL : describes given situation/problem,

• REFERENTIAL LEVEL : provides situation independent interpretation(s) of symbols.

There are known AI programs, which can be treated as examples of the division of levels of reasoning into the three levels, although they do not deal with commonsense deductions. For instance:

♦ H.Gelernter's (1963) geometry theorem prover

METALEVEL : "Like the human mathematician, the geometry machine makes use of the potent heuristic properties of a diagram to help it distinguish the true from the false sequences". Implementation: "heuristie computer".

OBJECT LEVEL : Axioms + theorems in a well defined formal system - geometry .

REFERENTIAL LEVEL : diagrams.

> The most notable feature of the program, however, was an additional part of the problem statement used to avoid attempting proofs by blind syntactic manipulation alone. (...) Whenever a subgoal was generated, it was checked for consistency with the diagram. If false in the diagram, the subgoal could not be possibly a theorem and therefore could be *pruned* from the search tree.
> (Handbook of AI, vol.1., p. 121, 1981).

♦ B.V.Funt's(1980) WHISPER

Other examples of referential levels may be : a procedure for a primitive robot preventing it from falling from a table, and other procedures for using mental models (Johnson-Laird, 1983); a procedure for interpreting predicates *left(X_fY)* and *right(X, Y)* , (in a dictionary "left" is defined as opposite to "right"). Notice that, in cases like these, the truth of predicates on the referential level may be established by non-logical means, for instance by a numerical procedure.

We should point out that a referential level does not exist in CWA, Clark's completion or circumscription. We think this is precisely why circumscription fails to capture certain aspects of commonsense reasoning.

J. McCarthy (1986) uses meta theory to obtain the right class of models for his theories of abnormality. A meta theory provides constraints coming from knowledge about knowledge ; referential models constrain possible interpretations in a completely different way : reference means that names of predicates do matter, we cannot change names and preserve meaning, grammatical form reflects semantics.

Dictionaries and grammar will be used by us as a referential level for commonsense reasoning.

Section 3. A comparison of circumscription and intended interpretations

In this section we want to examine two solutions to the problem of reasoning about Tweety the Canary :

* We know that Tweety is a canary. We ask : Can Tweety fly ?

3.1. Circumscription

J. McCarthy(1986) introduces three kinds of abnormalities and then formalizes the problem . as follows:

$$
\begin{array}{lll}
\neg ab\_1(x) & \rightarrow & \neg flies(x) \\
bird(x) & \rightarrow & ab\_1(x) \\
bird(x) & \rightarrow & flies(x) \vee ab\_2(x) \\
canary(x) & \rightarrow & bird(x) \vee ab\_3(x) \\
canary(Tweety). & &
\end{array}
$$

These sentences mean:

Usually tilings don't fly.

Birds are not usual things, in this aspect.

Birds usually fly, unless something prevents it.

The normal meaning of 'canary* is 'a kind of bird'.

Tweety is a canary.

In the above logical theory it is not possible to prove that Tweety flies. However when the abnormality predicates are minimized in a right order, the remaining minimal model contains the right conclusion. The correct sequence in this case is : ab_3 , ab_2 , ab_1 .

In this case circumscription works. Although the question remains : what is the source of the priorities. Must it always be a human ?

3.2. An alternative solution - an interaction of an object level and a referential level

We show how the same conclusion can be achieved as a result of interaction of an object level and a referential level within the framework we propose.

The *object level theory* is, in this case, very simple:

♦ *canary(Tweety)*, i.e. "Tweety is a canary".

+flies(Tweety) , i.e. "Can Tweety fly ?"

We assume that the information about birds and flying belongs to the referential level. In this case we take the

* We write three abnormality predicates instead of one, with two arguments, to facilitate the exposition.

Longman Dictionary of Contemporary English. Thus the *referential level* contains among thousands of facts the following information:

- bird is a creature with wings and feathers,
- feather is one of the many parts of the covering which grows on a bird's body, (...)
- wing is one of the 2 feathered limbs by which a bird flies, (...)
- canary is a type of small yellow bird usually kept as a pet.

This information is sufficient to deduce that Tweety can fly, provided we have a method of parsing the dictionary definitions and transforming them into logical formulae. Such tools begin to appear, and some of them have been used in a similar context: K.Jensen and J.- L. Binot (1987) suggest that the disambiguation of prepositional phrases, like these in the sentences

I ate a fish with a fork.

I ate a fish with bones.

is possible using on-line dictionaries and grammar. They have built tools to parse dictionary entries, and to extract semantically relevant information from these definitions and example sentences.

In our example, a dictionary has been used as a referential model in finding a plausible interpretation. In the next section we describe how an on-line grammar can suggest a non-classical interpretation of negation in the solution of a paradox of circumscription. In Section 5 we reexamine the above example in order to show how the nonmonotonicity of commonsense deductions can be expressed using intended models based on preferences. We explain there how to deal with situations when Tweety is a chicken, or a canary without wings.

Section 4. The paradox of Hanks and McDermott

S.Hanks and D.McDermott (1986) recently criticized circumscription and other formal methods of nonmonotonic reasoning as inherently incapable of representing a certain, commonsense case of temporal reasoning. One of their arguments is based on the facts that these methods cannot choose the right conclusion in an instance of only one type of abnormality.

This is the set of axioms that results in the paradox.

1. $T(alive, s_0)$,
2. $T(loaded, result(load, S))$,
3. $T(loaded, S) \rightarrow ab(alive, shoot, S)$ & $T(dead, result(shoot, S))$,
4. $T(F, S)$ & $\neg ab(F, E, S) \rightarrow T(F, result(E, S))$.

In English, these axioms say:

E1. At the moment S0 a person is alive.

E2. A gun is loaded, as a result of loading (in any situation).

E3. People die as a result of being shot with a loaded gun, moreover being shot with a loaded gun is abnormal with respect to staying alive.

E4. A fact does not change, unless it is abnormal with respect to an event.

And this is the sequence of actions:

s_0 , $s_1 = result(load, s_0)$, $s_2 = result(wait, s_1)$,
$s_3 = result(shoot, s_2)$.

Possible interpretations:

1. Without circumscription we have $T(alive, s_0)$, $T(loaded, result(load, S))$, and nothing else can be proven.

2. With circumscription: $T(alive, s_0)$,
$T(loaded, result(load, S))$, **and either (a) or (b):**

(a) $ab(alive, shoot, s_2)$, $T(dead, s_3)$;

(b) $ab(loaded, wait, s_1)$, $T(alive, s_3)$.

Hence there are two models minimizing abnormality. But circumscription fails to choose the correct one, in which the person dies as the result of being shot.

This forms a basis for the critique of circumscription by Hanks and McDermott:

1. Even a single default rule can have multiple extensions, and circumscription (or default reasoning) is not able to choose the right one.

2. There is no benefit from new, more complex versions of circumscription.

> "The problem is that the original idea behind circumscription, that a simple, problem independent extension to a first order theory would minimize predicates in just the right way has been lost along the way."

3.

> (...) it can be very hard to characterize the consequences of the circumscription axioms for a reasonably large and complex theory, and when the consequences are understood they may not be what we intended. The upshot is that no one really wants to know what follows from circumscription axioms; they usually wind up as hopefully harmless decoration to the actual theory.

A SOLUTION OF THE PARADOX

To find a solution of the paradox let us compare first the following sentences:

> Facts do not change every moment, unless something is unusual.
>
> Either something is unusual or facts don't change.

It is clear that these sentences do not have the same meaning. Hence the logical formulae

$$T(F, S) \;\&\; \neg ab(F, E, S) \;\rightarrow\; T(F, result(E, S))$$

$$T(F, S) \;\rightarrow\; ab(F, E, S) \vee T(F, result(E, S))$$

which correspond to these sentences should not be treated as equivalent either. (But in classical logic, they are.)

A computational grammar of English - the PLNLP English Grammar of K.Jensen et al.(1986), for instance - can be used to pick up the grammatical differences between "or" and "unless". For example, the first sentence is "a complex declarative with an embedded subordinate clause", while the second one is "a compound declarative". These grammatical differences can be then reflected in semantics:

> Since it is not known that "loaded" is abnormal with respect to waiting, one can assume *-,ab(loaded, wait, S)* . Then *ab{alive^hoot)* , and *T(dead$_y$ s^)* can be deduced.

Thus the natural interpretation of "unless" is the negation under the closed world assumption. * And this way we obtain the correct extension in the case of the problem of Hanks and McDermott.

This solution is based on the claim that grammatical differences express semantical ones. It is obviously a plausible assertion in the context of commonsense reasoning. Notice that the interpretation of negation in *-,ab(F.E,S)* as a negation under CWA , is not the main issue here, although this is the way the unique model is chosen. The crux of the approach we propose is to use the linguistic constraints to decide what is the right interpretation of this negation. Thus the formulation of the rule in English suggests both: its logical form and the interpretation of symbols appearing in this logical form.Of course, it often happens that the formulation of a rule in English does not have an obvious and unambiguous translation into a logical formula. But then, if the English form of the rule is also given, as it is for instance the case with many expert systems, it might help in interpreting logical symbols.

Section 5 . Mathematics of intended interpretations - elements of a theory

We define first the notion of a referential model and the logical structure of models of commonsense reasoning we consider. Then we formally describe a particular class of referential models - the ones expressing preferences in interpretations. Also, we reexamine the example of Tweety the Canary, to account for the nonmonotonicity of commonsense reasoning. Finally, we make some remarks about interaction of a referential level and a metalevel in the context of the problem of Hanks and McDermott.

5.1. The formal structures

A *referential model* can be any model that constrains interpretations of predicates of a theory, i.e. there is a formula P(c) whose truth is undecided by a theory T , but which is either true or false in the model. For instance only $x, x \;\&\; \neg w \rightarrow y \vee z$ can be provable in an object theory but a referential model may allow to conclude $\neg w$ and -y.

DEFINITION. Let T be a theory. n is a *referential model* for T , if there is a form $P(\bar{c})$ uch that T does not prove P(?) or $\neg P(\bar{c})$, but $\| P(\bar{c}) \|_{\mathcal{N}} \in \{ \mathit{True}, \mathit{False} \}$.

($\| \phi \|_{\mathcal{N}}$ stands for a logical value of the formula <£ in the models).

We already explained in Section 2.3 that we assume the existence of three levels of reasoning: a metalevel, an object level, and a referential level. Thus we do not discuss formal structures of the form $(\mathbf{M}, \mathbf{T}, \vdash_{M})$, where M is a collection of metarules (e.g. $\mathbf{M} = \{$ "formula circumscription" }), T is a collection of object theories to which M are applicable, and $\vdash_{M}$ is a provability relation that possibly extends the classical provability by making use of the rules in M. Instead, we want to analyze the quadruples $(\mathbf{M}, \mathbf{T}, \mathbf{R}, \vdash_{R+M})$, where R is a referential level, and $\vdash_{R+M}$ uses R, and possibly M . In particular, we are interested in the case (described below), when $\mathbf{R} = \{ (t, < ,) : t \in \mathit{Terms} \}$ contains partial orderings that express preferences in interpretations of terms.

5.2. Preferred interpretations

R. Jackendoff (1983), discusses the importance of preference rule systems for the theory of word meaning. He also points to the ubiquity of preference rule systems (cf. also Rock, 1983). Jackendoffs insights justify our claim that *preferred interpretation* is a right formalism for the reasoning which uses language as its referential model. We show then in Section 5.3 how preferences can be used to capture the non-monotonicity of commonsense reasoning.

It is now time to introduce definitions that are needed to formalize the notion of an intended theory .

DEFINITIONS. MODELS AND INTENDED THEORIES

♦ We assume we have a *lexicon* (a set of symbols) $\mathcal{L}$, and a *grammar* $\mathcal{G}$ which decides which sequences of elements of *SB* constitute valid terms and formulae.

* Since we do not exclude indefinite conclusions, the CWA here is the generalized CWA (Minker,1982), i.e. the theory of all minimal models of a given theory.

For instance, a computer grammar (like the already mentioned PEG) and a list of English words can be such a pair $< \mathcal{G}, \mathcal{L} >$. Then terms would be words, phrases (NP's , VP's, PFs etc.), and sentences.

♦ The *formulae* can be identified with indicative sentences.

♦ A *theory* is a collection of formulae.

♦ In particular, *logical formulae* can be treated as a subclass of formulae. Similarly for *logical theories* .

♦ With each term we associate a list of its possible interpretations - other terms - (partially) ordered by a relation of preference. In particular, with some terms there will be associated sequences of theories ordered by preferences:

t_0 , $(T^0_0, T^0_1, \ldots, T^0_{k_0})$; ... ; t_m $(T^m_0, T^m_1, \ldots, T^m_{k_m})$;
and if $j < j'$ *then* $T^i_j < T^i_{j'}$,
i.e. T^i_j **is more preferred than** $T^i_{j'}$.

(Since the partial order will be often a linear order, so, in order not to complicate the notation, we assume it).

We can also assume that the *empty interpretation* $\emptyset$ **is** least preferred, for all terms.

♦ A *referential model based on preferences* is a collection of such pairs.

♦ An *intended interpretation of a term* consisting of a sequence of subterms (possibly one element) is a union of their most preferable interpreutions. We require however such an interpretation (i.e. this union) to be consistent . More formally, let

$$\Pi(t) = \prod_{i \le m} (T^i_0, T^i_1, \ldots, T^i_{n_i}),$$

$$\tilde{\Pi}(t) = \{ \pi \in \Pi(t) : \cup \pi \text{ is a consistent theory } \}.$$

Let $<$ be the partial order induced on $\tilde{\Pi}(t)$ by the orderings of associated interpretations.

♦ The *partial models* PM(t) of a the sequence of terms are the the most likely theories of t given by $(\tilde{\Pi}, <)$:

$$\mathrm{PM}(t) = \{ m : (\exists \pi)[m = \cup \pi \ \& \ \pi \text{ is a minimal element of } (\tilde{\Pi}(t), <)] \}.$$

The partial models pick up from the referential level the most obvious information about /. This immediate information may be insufficient to decide the truth of the formulae of /. For instance, if

$$\mathrm{PM}(t) = \mathrm{PM}(\mathit{bird}(T). \leftarrow \mathit{flies}(T))$$
$$= \{\mathit{bird}(x) \rightarrow \mathit{has}(x,\mathit{wings})\} \cup \{ t \},$$

but only PM (PM (t)) **contains the formula** $\mathit{has}(x,\mathit{wings}) \leftarrow \mathit{flies}(x)$, **then the iteration of PM's is needed.**

♦ **Let** $\mathrm{PM}_0(t) = \mathrm{PM}(t)$,
$\mathrm{PM}_{n+1}(t) = \mathrm{PM}(\mathrm{PM}_n(t))$,
$\mathrm{PM}_\omega(t) = \cup \ \{\mathrm{PM}_n(t) : n < \infty \}$.

PM(t) is a collection of many models that interpret t. It is infinite even if all PM„(/) are one element. Clearly, we are interested in these elements of PM(t) which contain maximum of information.

♦ We define the *intended models* of t

$$\mathrm{IM}(t) = \{ m \in \mathrm{PM}_\omega(t) : m \text{ is maximal under } \subset \} .$$

♦ The *intended theory* of t is defined, as before, as the set of formulae which are true in all intended models.

5.3. *Nonmonotonic reasoning and changed preferences*

We illustrate now these notions with an example which also points to the source of nonmonotonicity in our formalization of commonsense reasoning. Namely, changed preferences, not very strong inference rules like circumscription, produce modifications in intended theories. We will use a very simple, Prolog-like notation to make the inferences and changes transparent.

Tweety the Canary - revisited.

♦ A fragment of the referential model:

bird
1. $\mathit{bird}(x) \rightarrow \mathit{has}(x,\mathit{wings})$.
$\mathit{bird}(x) \rightarrow \mathit{has}(x,\mathit{feathers})$.
2. $\mathit{bird}(x) \rightarrow \mathit{person}(x)$.
3. $\mathit{bird}(x) \rightarrow \mathit{woman}(x)$.
4. $\mathit{early\_\_bird}(x) \rightarrow \mathit{arrives\_\_early}(x) \vee$
$\vee \ \mathit{gets\_\_up\_\_early}(x)$.

wing
1. $\mathit{has}(x,\mathit{wings}) \& \mathit{bird}(x) \rightarrow \mathit{flies}(x)$.
2. $\mathit{wing}(x) \rightarrow \mathit{part\_\_of}(x, \mathit{airplane}) \&$

fly
1. $\mathit{flies}(x) \rightarrow \mathit{moves\_\_thru\_\_air}(x,\mathit{self})$ &
& $\mathit{has}(x,\mathit{wings})$.
2. $\mathit{flies}(x) \rightarrow \mathit{moves\_\_thru\_\_air}(x,\mathit{machine})$.
3. $\mathit{flies}(x) \rightarrow \mathit{controls}(x,\mathit{aircraft})$
& $\mathit{guides}(x,\mathit{aircraft})$.

canary
1. $\mathit{canary}(x) \rightarrow \mathit{bird}(x)$.
2. $\mathit{canary}(x) \rightarrow \mathit{woman}(x) \ \& \ \mathit{sings}(x)$.

Note: In this example, the logical formulae are based on dictionary definitions (Longman,1978) , but the information on the referential level may come from other sources too, thus a "bird" may also contain *bird* 1. $\mathit{bird}(x) \rightarrow \mathit{has}(x,\mathit{wings})$.

...

10. $\mathit{bird}(x) \wedge \neg \mathit{flies}(x) \rightarrow \mathit{penguin}(x)$.

11. $\mathit{bird}(x) \wedge \neg \mathit{flies}(x) \rightarrow \mathit{ostrich}(x)$.

• The object theory says that "Tweety is a canary", and asks "Can Tweety fly ?" :

$\mathit{canary}(\mathit{Tweety}), \qquad \leftarrow \mathit{flies}(\mathit{Tweety})$.

• The first partial model is obtained by using the rules (canary 1) , (fly 1) :

$t = \mathit{canary}(\mathit{Tweety}) + \leftarrow \mathit{flies}(\mathit{Tweety})$

$\mathrm{PM}_0(t) = \{m\}$, and

$m = \{ \mathit{canary}(\mathit{Tweety}) + \mathit{canary}(x) \rightarrow \mathit{bird}(x) + \mathit{flies}(x) \rightarrow \mathit{moves\_\_thru\_\_air}(x,\mathit{self}) \& \mathit{has}(x,\mathit{wings}) + \leftarrow \mathit{flies}(\mathit{Tweety}) \}$

(+ is a catenation symbol. We can assume that + is associative and commutative).

• The partial model after the iteration, when additionally the entries (bird 1) ,(wing 1) are used :

$\mathrm{PM}_1(t) = \{ m' \}$, where

$m' = \{ \mathit{canary}(\mathit{Tweety}) + \mathit{canary}(x) \rightarrow \mathit{bird}(x) + \mathit{flies}(x) \rightarrow \mathit{moves\_\_thru\_\_air}(x,\mathit{self}) \& \mathit{has}(x,\mathit{wings}) + \mathit{bird}(x) \rightarrow \mathit{has}(x, \mathit{wings}) + \mathit{has}(x,\mathit{wings}) \& \mathit{bird}(x) \rightarrow \mathit{flies}(x) + \ldots - \text{theories about } \textbf{move}, \textbf{air}, \ldots + \leftarrow \mathit{flies}(\mathit{Tweety}) \}$

• The conclusion: *flies(Tweety)*. (Notice that no further iterations were necessary) .

♦ A changed object theory will produce different partial models. Let us consider the theory "Tweety is a canary, but he has no wings", and the same question : " Can Tweety fly ?".

$\mathit{canary}(\mathit{Tweety}), \quad \neg \mathit{has}(\mathit{Tweety},\mathit{wings}), \quad \leftarrow \mathit{flies}(\mathit{Tweety})$.

• The new partial models PM(t) are obtained using rules (wing 1) (flies 1) (canary 1) , and then (bird 2), since (bird 1) is inconsistent with the previous three.

$t = \mathit{canary}(\mathit{Tweety}) + \neg \mathit{has}(\mathit{Tweety},\mathit{wings}) + \leftarrow \mathit{flies}(\mathit{Tweety})$

$\mathrm{PM}_0(t) = \{ \{ \mathit{canary}(\mathit{Tweety}) + \neg \mathit{has}(\mathit{Tweety},\mathit{wings}) + \mathit{flies}(x) \rightarrow \mathit{moves\_\_thru\_\_air}(x,\mathit{self}) \& \mathit{has}(x,\mathit{wings}) + \mathit{has}(x,\mathit{wings}) \& \mathit{bird}(x) \rightarrow \mathit{flies}(x) + \mathit{canary}(x) \rightarrow \mathit{bird}(x) + \leftarrow \mathit{flies}(\mathit{Tweety}) \} \}$

$\mathrm{PM}_1(t) = \{ \{ \mathit{canary}(\mathit{Tweety}) + \neg \mathit{has}(\mathit{Tweety},\mathit{wings}) + \mathit{flies}(x) \rightarrow \mathit{moves\_\_thru\_\_air}(x,\mathit{self}) \& \mathit{has}(x,\mathit{wings}) + \mathit{canary}(x) \rightarrow \mathit{bird}(x) + \mathit{bird}(x) \rightarrow \mathit{person}(x) + \mathit{has}(x,\mathit{wings}) \& \mathit{bird}(x) \rightarrow \mathit{flies}(x) + \ldots - \textit{theories about } \textbf{move}, \textbf{air}, \ldots + \leftarrow \mathit{flies}(\mathit{Tweety}) \} \}$

♦ The new partial models do not support the previous conclusion *flies(Tweety)* If the formula $\mathit{person}(x) \rightarrow \neg \mathit{flies}(x)$ appears in the intended models *IM(t)* , then *--flies(Tweety)* can be derived. Even if this does not happen, we can observe the nonmonotonic change of the theory : a theory does not prove all theorems of its subtheory.

This example illustrates how changes In the object theory produce different interpretations , although the same set of preferences serves as a referential model. Our model well explains the nonmonotonicity of commonsense reasoning, and seems to be more natural than the models based on circumscription or default logics. Of course, it is inevitable that conclusions based on dictionary knowledge will be quite often incorrect. But neither will they violate linguistic constraints. And we believe that such violations would yield unsound interpretations.

What we presented above is just a basis for a theory of referential models grounded on preferences. Algebraic and logical properties of such models are open to investigation; basic logical results about these models are presented in Zadrozny, 1987a.

5.4. Relationship between the preferred interpretations and default logics

The notion of a referential model based on preferences has some of the spirit of prioritized circumscription (McCarthy, 1986) or default logic (Reiter,1980). There are nonetheless important differences:

1. The logic of default reasoning has only one level, and possibly a metalevel. We have three levels, one of which is problem independent.

2. Default logic does not permit any ordering on defaults. Circumscription does, but such an ordering is a part of a metalevel, and its origins are mysterious. We can account for the preferences on the referential level. This "ubiquity of the preference ruies" (Jackendoff,1983; Rock,1983) gives also some psychological plausibility to the proposed model.

3. Defaults represent variants of possible interactions between rules of an object theory; preferences represent priorities on interpretations, which are unchanged across different object theories (or knowledge bases), and belong to the referential level

4. Prioritized circumscription deals with orderings on different kinds of abnormalities; priorities in our framework are about interpretation of any terms. Thus prioritized circumscription has nothing to say about examples like the one from the beginning of Section 2, while our model can capture also some aspects of pronominal reference and of the disambiguation of prepositional phrases (as presented in K.Jensen & J.- L.Binot, 1987).

5.5. Interaction of a referential level and a metalevel

A referential level should provide situation independent interpretations of symbols. In the above example these interpretations were interacting with an object theory. But our analysis of the "Yale shooting problem" in Section 4 indicates that the referential level may contain specifications of relationships between terms and and their metalevel interpretations. For instance, the term "unless" may be associated on the referential level with the metalevel formula "if $t = {}'\Phi \textbf{ unless } \Psi{}' \textbf{ then } \Psi$ should be interpreted under CWA ".

Section 6. Conclusions

The most significant novelty presented here is the separation of the referential level from the object level and the metalevel, and the suggestion that natural language is the most important reference for common sense.

We've proven that priorities in interpretation of predicates and constants on the logical level of reference can be the source of nonmonotonicity in reasoning. We have shown that NL in the form of (on-line) dictionaries, grammars etc, can be taken as the referential level for commonsense reasoning. Moreover, the three-level logical framework can be applied to semantics of natural languages (Zadrozny and Jensen, 1987).

Since more types of problems can be solved using the division into the three logical levels than by minimization alone, this approach is superior to circumscription. But it is clearly an open problem what are the limits of the proposed method, how much of common sense can be captured by referring to language, and how one could measure the common sense.

Acknowledgements. I want to thank J.L.Lassez, K.Jensen, J.L.Binot, K.Skandalis and T.Wetter for useful discussions. Part of this paper was written in October 1986 during my stay in Germany, which was supported by the Alexander von Humboldt Stiftung.

REFERENCES

1. AJBarr, E.A.Feigenbaum (Eds.), *Handbook of AI volL,* William Kaufmann, Inc. ,1981.
2. J.Doyle , *Circumscription and Implicit Definability,* J. of Automated Reasoning, 1 , 1985, pp. 391 - 405.
3. B.V.Funt, *Problem Solving with Diagrammatic Representations,* AI Journal, vol. 13,No.3 ,1980.
4. H.Gelernter, *Realization of a geometry-theorem proving machine,* in : E.A.Feigenbaum, and J.Feldman (Eds), *Computers and Thought*, McGraw-Hill, NY , 1963.
5. S. Hanks and D. McDermott , *Default Reasoning, Nonmonotonic Logics, and the Frame Problem,* Proc. AAAI - 86 , pp. 328 - 333.
6. J.Haugeland, *Artificial Intelligence : The Very Idea.* MIT Press, 1985.
7. R.Jackendoff, *Semantics and Cognition,* MIT Press, 1983.
8. K.JensenJ-L.Binot, *Disambiguating Prepositional Phrase Attachments by Using On-line Dictionary Definitions* ,Proc. IJCAI-87, 1987.
9. K.Jensen, G.E.Heidorn, S.D.Richardson, N.Haas, *PLNLP, PEG, and CRITIQUE : Three Contributions to Computing in the Humanities* , IBM Research Report RC 11841, 1986.
10. P.N.Johnson-Laird, *Mental Models in Cognitive Science,* Cognitive Science,4,1980.
11. P.N.Johnson-Laird, *Mental Models,* Cambridge University Press, 1983.
12. *Longman Dictionary of Contemporary English,* Longman Group Ltd., London, 1978.
13. J. McCarthy , *Circumscription - A Form of Non-Monotonic Reasoning,* AI Journal , 13, 1980, pp. 27-39 .
14. J. McCarthy , *Applications of Circumscription to Formalization of Common-Sense Knowledge,* AI Journal, 28, 1986, pp. 89-116.
15. D. V. McDermott and J.Doyle , *Non-Monotonic Logic I,* AI Journal, 13, 1980, pp. 41 - 72 .
16. J.Minker,On *indefinite data bases and Closed World Assumption",* Proc. 6-th Conference on Automated Deduction,Springer, 1982.
17. R. Reiter, *A Logic For Default Reasoning,* AI Journal, 13, 1980, pp. 81 -132.
18. I.Rock, *The Logic of Perception,* MIT Press, 1983.
19. W.Zadrozny, *A theory of default reasoning,* Proc. AAAI-87, 1987 (a).
20. W.Zadrozny, *Minimization and common-sense reasoning,* submitted for publication, 1987 (b).
21. W.Zadrozny and K.Jensen, *Semantics of paragraphs,* in preparation, 1987 .