A Commentary on the Literature of Self-Reference

by

Donald Perlis and Michael Miller

A COMMENTARY ON THE LITERATURE OF SELF-REFERENCE

Donald Perlis and Michael Miller

Computer Science Department
University of Maryland
College Park, MD

Abstract

Self-reference, far from being just a logician's and a philosopher's puzzle is ,in fact, a central feature of human language and reason. It, thus, seems natural that intelligent machine will also have to deal with the issue of self-reference. We discuss some of the formal problems, and potential solutions and applications. Portions of this essay are descriptive in nature, portions prescriptive. We are involved in the development of some of the ideas in the relevant literature, and make no apology for injecting a certain subjective note into the text, as opposed to forcing a false objectivity. We have also freely drawn on portions of essays written by one of the authors.

Support for the preparation of this document and related research from both the Systems Research Center at the University of Maryland [NSF contract #OIR-8500108] and the Martin Marietta Corporation is gratefully acknowledged.

I. General Introduction

One of the most important characteristic features of human language and reasoning is the concept of self-reference. As researchers continue to attempt to endow computers with generalized intelligence, it is becoming more and more apparent that such intelligence cannot be had without self-referential capabilities.

There are at least two roles that self-reference plays regarding intelligence. The first is concerned with language while the second is concerned with beliefs. With respect to language, self-reference allows a language to refer to components of the language itself. Similarly, a self-referential belief system is able to consider its own thoughts. Computers, as of yet have not been granted language or belief systems that are self-referential. Much work, however, has been undertaken that attempts to formalize these concepts so that we may provide computers the ability to reason in an intelligent manner.

This paper highlights the issues and progress of the research, most of which is extremely current, in the area of self-reference. We treat both the theoretical and practical aspects of these works in an effort to provide as comprehensive a view as possible.

II. Self Reference in Language

A. Introduction

The Liar Paradox is perhaps the most famous case of a self-referential statement, most compactly given in the form, ´This sentence is false.´ This and related paradoxical statements have been the focus of an enormous amount of attention in modern philosophy and logic, often with the view that extremely serious foundational issues are at play here,

showing formal languages of a sufficiently rich sort to be on a very tenuous plane. However, although no one would claim that natural language is any less complex than formal languages are, still the tendency in the former is to emphasize the complex usages of language rather than explore any inherent foundational infirmity. Thus work in knowledge representation produces increasingly complex systems (e.g., Bobrow and Winograd [1977] and Brachman [1978]) by creating whatever representational structures are desired, without any apparent trouble with systematic inconsistency. The concern with such statements as the Liar's, then, has not spread to the domain of natural language research per se.

A statement need not be as formal or deliberately concocted as the Liar to produce paradox. In fact, natural language abounds with self-reference, as was partially indicated in [Kripke 1975], with combinations of sentences such as

1) John: 'Don't trust what Bill tells you.'

2) Bill: 'Don't trust what John tells you.'

Here each individual is making a statement that refers to (possibly many) statements made by the other, with the understood intention of including the corresponding statement given above; but *that* statement then refers back to the other. The listener is being urged by each speaker to take *his* statement and *not* the other's.

Other examples of straightforward sentences that nonetheless involve a kind of self-reference or reference to a set of sentences or beliefs *containing* the given one, are:

3) I am now speaking English.

4) Any reasonably short sentence can go into

an abstract.

5) Every sentence must end in a period.

6) I'd know if I knew that!

7) Some of my beliefs may be false, but none are stupid.

8) Some things anyone says are false.

9) Some sentences are about other sentences.

All of these seem to read fairly easily and could be heard spoken in normal conversation. Furthermore, they express ideas that are sufficiently close to matters of general concern that it would be fairly restrictive of a representational system not to handle them, if it purported to be a formalization of natural language.

B. The Problem

We wish to emphasize a point that is implied in the above discussion. That is, that self-reference is necessary in any language with a concept of its own grammar. By the latter we mean simply the capacity of the language to refer to grammatical entities such as words and sentences. This is needed for instance to discuss whether John said X or Y, surely a construction present in and vital to all natural languages. But this requires some kind of device to quote utterances, i.e., to recognize that a sentence or other grammatical construct is being talked about rather than asserted. Even if the language does not contain an explicit quotation mechanism, the underlying semantics surely does. For how else would we account for the

fact that such an utterance as ´Bill is here´ informs the listeners not only about Bill but also about the *name* ´Bill´? Consider Alice, who did not know of the individual referred to: how is the sentence understood by her? Surely as evidence that someone named ´Bill´ is present or arriving. So her ´inner language,´ so to speak, has a recognition of the parts of the outer language; people know they are using words to express ideas, and this knowledge is actually part of very many instances of language use. Alice knows ´Bill is here´ has been said and that it has given her this information: she can unquote the sentence to arrive at its meaning.

Another way of putting this is that semantics is a part *of* language. People know about it and use it explicitly. We tell one another what meaning sentences have for us. We refer to a sentence ´...´ and say that it means something. Formally, we have something like

Unquote(´...´) <--> ***

where *** expresses what it means for ´...´ to hold. For example,

Unquote(´Snow is white´) <--> Snow is white.

Of course, semantics purports to do better than this, usually by providing a formal object on the right instead of repeating the same sentence unquoted! A recent and much heralded effort in this direction is Barwise and Perry [1981].

However, any such effort must sooner or later face the issue of explicating the semantics of semantics itself, i.e., such a predicate as Unquote (or Hold or True). When these predicates themselves are part of the sentence being quoted, we have self-reference in its full glory, and the Liar Paradox is a distinctly ominous possibility. (See Perlis [1981] for a formal construction of the Liar Paradox in a language with an expression of its own grammar.) It appears then that we need rules governing when certain quotations are allowed.

C. Semantics for Truth with Self-Reference

Such rules have been worked on extensively in what we may call the hierarchical approaches. These involve using a notational type for each level of reference, as in Russell [1908] and Tarski [1944]. Unfortunately they do not serve to express what is needed. For a sentence to contain another as term, requires the main sentence to be of higher order than that of the term:

´John said "I am six feet tall"/1, but he is wrong.´/2

where the indices show the respective order types. The trouble with this, apart from its cumbersomeness, is that it fails to capture such salient facts as

´A higher order sentence can contain only lower order terms.´

For we could do this only with a sentence that itself would have some order:

´A sentence/9 can contain only terms/8.´/10

Thus only a small part of what we intend is actually expressed.

We are apparently faced, then, with the need for both arbitrary referring of sentences to one another, and a truth predicate that also can apply to arbitrary sentences. Yet we must avoid the Liar Paradox if we are to have any trustworthy semantics. Can we simply banish troublesome kinds of self-reference? This is what Russell and Tarski were after, of course, but to do so we need another approach. It turns out that this is not at all trivial (viz., 75 years of foundational efforts to this end in mathematical logic), in particular as evidenced in Kripke [1975] by further examples in the John-Bill vein as above.

On the other hand, Kripke also showed that a lot can be accomplished *toward* a

semantics for truth with self-reference. He constructs a truth definition for ´grounded´ sentences, ones that can be reduced to non-self-referential ones and then finally decided true or false. However, the non-grounded ones are left open, and they include both the Liar´s and others such as 1) and 2) above. Moreover, in analogy with the hierarchical situation, we can form

10) Each true sentence (in Kripke´s scheme) is so by
virtue of others decided before it.

This in fact is how his truth definition proceeds, and is crucial to its use and understanding. Yet (10) is not so decided! To do so would require first checking that *all* true sentences so followed, and this would require (10) to be true or not!

D. Quotation and Truth

We now discuss a scheme, having much in common with Kripke´s, but borrowing a tool from Gilmore [1974], to formalize a simple and computationally cheap truth definition, which allows the statement of the fact that a given sentence may not be decided one way or another. Moreover, Gilmore´s work (which was in the foundations of set theory) can be modified to show the resulting system consistent (see Perlis [1981]).

The scheme is as follows: for definiteness we work in a first-order logic, with a truth predicate T. Then instead of the axiom schema T(´P´) <--> P for all sentences P, which leads to the Liar and other paradoxes, we postulate T(´P´) <--> P*, where the *-operator replaces each occurrence (inside P) of a subformula of the form -T(´Q´) with the subformula T(´-Q´). This apparently modest change makes all the difference, leading from an inconsistent formalization to a consistent one. It recognizes a difference between T(´-Q´) (i.e., Q

really failing) and -T(´Q´) (i.e., Q not really holding). Thus in a sense this could be regarded as akin to a three-valued logic, although in fact it is not. It is a perfectly ordinary first-order logic, in which excluded middle is valid: P v -P, and even T(´P´) v -T(´P´). The point is that T here is a predicate, and has a slightly different meaning from what we are accustomed to: intuitively it can be regarded as saying whatever its argument says *and* that that can be established *before* any judgement about T is made. We refer to this as the ´normal order´ of judging that a sentence is true. To make the judgement ´"Snow is white" is true,´ we first judge that snow is white and then pass to the observation about the sentence.

In effect we find we can consistently write

R: True(´-R´)

and

S: -True(´S´)

and even prove S (for it undeniably is not ´True´) as well as -R (for R, being True(´-R´), is also not tenable). In terms of the Liar versions above, we find both S and R violate normal order, since each makes claims about its own quoted truth; hence we have -True(´S´) and -True(´R´). (The former is even S itself!) Although it takes getting used to this new sense of truth, it seems fairly close to intuition when examined in detail as in Kripke´s paper, and for ordinary (normal) sentences, it yields exactly the naive result that quotation changes nothing about the meanings. It appears then that the formal issues surrounding quotation are on a fairly solid footing.

E. Summary

We conclude that although self-reference is central to any powerful encapsulation of natural language, this does not present insurmountable problems, and that the notion of truth that often arises in self-referential statements can be handled in a straightforward manner that is both computationally feasible and more expressive than the standard alternatives.

III. Commonsense Reasoning and Non-Monotonicity

A. Introduction

Marvin Minsky [1974] coined the phrase "non-monotonic logic" in developing an argument that tools of formal logic are inadequate to the task of representing commonsense reasoning. His argument stressed the point that much of commonsense reasoning involves the use of default rules, that is, conclusions are drawn on the basis of the absence of other information. Consider the example of concluding from the knowledge that Tweety is a bird that Tweety can fly. Such a conclusion is not necessarily correct, although it is certainly one that can be very useful in many situations. One could try to specify carefully precisely those situations in which such a conclusion is sound, but any effort to do so quickly leads to despair. The possible circumstances in which any presumed correct line of reasoning can be defeated astounds: Tweety may be an ostrich, may have a broken wing, may be chained to a perch, may be too weak, etc. Indeed, the problem is virtually the same as that of the well-known frame problem: the special conditions relevant to determining what may be the case in a complex environment defies precise specification. Although we seem to have a strong sense of certain "typical" situations (such as "typical" birds being ones which among other things can fly), it is notoriously hard to define typicality.

Minsky seems to have concluded that formal methods per se are inappropriate to cap-

ture such reasoning, whereas others have taken his ideas as a challenge by which to find more powerful formal methods. Out of this challenge has arisen a substantial field of research in non-monotonic reasoning. The terrain has by now shown itself to be a rich and varied one involving ideas from divers parts of artificial intelligence, logic, natural language, and philosophy. One theme that seems to have emerged is that a key element in commonsense reasoning dealing with uncertainty (due to the abundance of special conditions defying specification) is self-reference: the reasoning entity utilizes information about the extent of its own knowledge. Indeed, most approaches to commomsense reasoning can be viewed in terms of their approach to representing such self-reference. We will explore this in what follows.

B. The Problem

Minsky argued that, firstly, conclusions of the sort given in the Tweety example are contingent on what else is known (e.g., if it is already known that Tweety cannot fly, we refrain from concluding the opposite), and that secondly such conclusions do not obey the customary phenomenon of "monotonicity" of formal systems of logic. That is, a standard logic L has the property that if ϕ is a theorem of L and if L is augmented to L* by additional axioms, then ϕ remains a theorem of L*. Indeed the same proof of ϕ in L is a proof of ϕ in L*. However, commonsense reasoning seems to allow a "proof" (at least in the form of a tentative supposition) that Tweety can fly, given *only* that Tweety is a bird, whereas in the augmented state in which it is known also that Tweety cannot fly no such "proof" is forthcoming. In effect account seems to be taken of what the reasoner does not know, an issue already much studied in the area of databases in the context of the closed world assumption (e.g., Minker [1982], Reiter [1978], and Reiter [1980]), in which any atomic formula not explicitly present in the database is intended to be false.

It is true that any *straightforward* attempt to represent such reasoning in terms of sentences in a traditional "monotonic" logic in which the stated conclusions are theorems, will fail, for the simple reason that these logics will necessarily have the original theorem (Tweety can fly) carried over to the augmented theories by virtue of their monotonicity. Several questions then arise:

(a) are there other formal logics that can represent such reasoning?

(b) has the commonsense reasoning been fairly portrayed here or are there other factors involved that might change the assessment of the role of non-monotonicity?

(c) might not a clever use of monotonic logic allow the effect of non-monotonic deductions?

These questions have guided much of the work in contemporary commonsense reasoning research.

C. Non-Monotonic Formalisms

Two distinct formalisms emerged around 1980 attempting to capture the essence of non-monotonic reasoning by providing a new kind of logical framework. One, due to McDermott and Doyle [1980], bears simply the name of "non-monotonic logic," and the other, due to Reiter [1980a], is called "default logic." Both employ inferential tools making explicit use of information about what information the formalism itself has available to it. In both cases new syntactic and inferential constructs are developed. We discuss each of these in turn.

The non-monotonic logic of McDermott and Doyle ("NML" for short) takes as point

of departure the desire to represent axiomatically such notions as "If an animal is a bird then unless proven otherwise it can fly." To do this they introduce into the language (initially a first-order language) a modal operator M, so that if p is a formula then so is Mp (read "p is consistent"). Now in this language it is possible to write formulas that seem to express the kind of reasoning given earlier. For instance, the formula

$$(x)[\text{Bird}(x) \;\&\; \text{M Flies}(x) \dashrightarrow \text{Flies}(x)]$$

appears to convey information appropriate to concluding of typical birds that they can fly. A means is needed to characterize deductions with formulas containing the operator M, however, and this McDermott and Doyle go to some length to develop. As this is essential to their treatment, we spend some time examining it now.

At first blush, it would appear easy to state what is wanted. For if indeed the formula p is consistent (with the rest of the axioms of the particular instance of NML one wishes to utilize), and if in "typical" situations (i.e., one's in which p is consistent) the formula q happens to be true, then a rule such as "from Mp deduce q" seems appropriate. However, McDermott and Doyle have chosen M to be a part of the language itself, i.e., Mp is a formula as well as p. This means that a mechanism is needed to make it possible to prove formulas such as Mp, and this is problematic since proofs of consistency are not only notoriously hard in general but in fact are usually impossible within the same axiomatic system with respect to which consistency is sought. To deal with this problem, McDermott and Doyle extend the notion of proof to allow a kind of consistency test, at the expense of effectiveness. (In fact, all formal approaches to non-monotonic reasoning seem to run into this same issue.) Their notion of proof is as follows: If A is a first-order theory and S is a set of formulas in the language L of A, let

$$NM_A(S) = Th(A \cup As_A(S))$$

where

$$As_A(S) = \{Mq{:}q \in L \text{ and } \text{-}q \notin S\} - Th(A)$$

Here Th(A) is the usual set of first-order consequences of A, and $As_A(S)$, the so-called set of assumptions from S, consists of those formulas Mq not in Th(A) for which -q is not in S. Intuitively, an Mq that is not already proven is to be considered an assumption on the basis of S if S does not rule q out, i.e., Q is considered to be "possible." The idea is to adjoin assumptions to A and find all (usual) consequences, this producing the set $NM_A(S)$. S of course could be A itself, or even empty. However, when NM_A is formed, new formulas are thereby available for use (i.e., they are considered "proven") and these may themselves provide the basis for another round of assumptions. So S plays the role of a recursion variable, and a fixed point of $NM_A(S)$ is sought. Thus the set of theorems non-monotonically derivable from A is defined as

$$TH(A) = \cap(\{L\} \cup \{S : NM_A(S) = S\})$$

Note that any attempt to calculate TH(A) leads to consistency tests. For in iterating $NM_A(S)$ for S initially empty, we arrive immediately at the necessity of determining whether, for any given p, Mq is in Th(A). But this is in general undecidable, and amounts precisely to determining whether A+{-Mq} is inconsistent. McDermott and Doyle acknowledge this difficulty and show that in very restricted cases--essentially propositional logic--there is a remedy. (They also define a notion of model for NML; however there is some dispute as to the completeness of their definition).

McDermott [1982] tries to strengthen NML so as to overcome certain weaknesses in the original version, in particular the fact that Mp and -p were not contradictory. The new effort makes fuller use of the modal character of the language, but in the most interesting case col-

lapses into equivalence with ordinary first-order logic.

Moore [1983] re-examines the underlying goals of NML and concludes that two ideas are being conflated: typicality on the one hand, and beliefs about ones beliefs on the other. He distinguishes between concluding Tweety can fly on the basis that it is not known that Tweety cannot fly and that typically birds can fly, and concluding Tweety can fly on the basis that it is not known that Tweety cannot fly and that "I would know it if Tweety could not fly." Moore argues that the former is intended to be approximate and error-prone, while the latter (which he calls "autoepistemic reasoning") is intended to be sound. He devises a consistent logic for the latter form of reasoning.

It does appear that autoepistemic reasoning forms a part of commonsense reasoning. Our example above is not as striking as one given by Moore: "I would know it if I had an elder brother." Here one is presumably not merely stating a belief about typicality (that one typically knows ones older brothers, although that seems true enough) but rather a belief that "I" specifically do know of all "my" brothers. Admittedly this is arguable, since one can think of situations in which an older brother may be unknown, but they are not likely to be taken seriously by us, so that again a kind of typicality may be present here.

Moore points out that in autoepistemic beliefs there is a possibility of failure, i.e., the belief can be false (we may have an elder brother after all) in which case we must alter that belief, whereas in the case of typicality we may merely conclude that we are atypical regarding knowledge of brothers and yet preserve the belief that typically elder brothers are known. Still, if we do discover to our surprise that such a brother exists, it would seem likely that we would conclude immediately that we were wrong about our autoepistemic belief *but that the belief still applies to most people.* I.e., there seems a very fine and tenuous line between the two forms of beliefs. It seems possible that we may move back and forth between explicit typ-

icality beliefs in which we acknowledge their uncertainty, and more stubborn autoepistemic ones, for the *same* assertions, depending on context, and our willingness to alter our position when challenged may attest to an implicit default character even in autoepistemic cases.

It is of interest that both forms of reasoning, however, like all other non-monotonic formalisms, depend at least implicitly on a determination that in fact certain formulas are not theorems of the formalism in question. Note that in Moore's example we must somehow determine that in fact we do not know of an elder brother, before using the autoepistemic belief and modus ponens to conclude we have no such brother. Again, this self-referential or consistency aspect of the reasoning seems the most striking characteristic, and the one presenting the greatest formal difficulty.

Reiter [1980] introduces a logic for default reasoning (which we here denote as DL). In specifically singling out default reasoning, Reiter identifies his concern as that of studying typicality rather than other possible non-monotonic forms of reasoning. His formalism in fact bears close resemblance to NML, the most obvious difference being that the language is strictly first-order, with the operator M playing a role only in rules of inference rather than in axioms. Specifically, Reiter allows inference rules ("default rules") such as "from Bird(x) and M Flies(x), conclude Flies(x)" where "M Flies(x)" is intended not as an antecedent theorem to the consequent Flies(x) but instead as a condition that must be met before Flies(x) can be concluded from Bird(x). The condition is, roughly (and as in all non-monotonic formalisms) that Flies(x) be consistent with the rest of the axiomatic framework. Making this precise and showing it to be useful is the bulk of the task Reiter undertakes. He employs a hierarchy of iterations along lines similar to that of NML, also arriving at a fixed point, in determining a notion of proof for default rules. Since DL uses rules in place of the axioms of NML, it would appear that in general DL is weaker than NML. This may in fact be a reason to prefer DL to

NML, in that one of the hoped-for features of reasoning about typicality is that limitations are placed on what conclusions are drawn. However, at present the outlines of what a reasoning system *should* do regarding typicality as so vague and ill-defined that it is difficult to defend strong claims.

Reiter and Criscuolo [1981] also consider what they call interacting defaults, i.e., default rules which separately might lead to opposed conclusions, such as in "Richard Nixon is a Quaker and a Republican" where it is known, say, that typically Quakers are pacifists and Republicans are not. This appears to be a substantial difficulty for any form of non-monotonic reasoning that pretends to deal with typicality.

D. Circumscription

Circumscription is a technique devised by John McCarthy [1980] for formalizing certain notions in commonsense reasoning, and specifically in non-monotonic reasoning. It differs from NML and DL in being formalized wholly within first-order logic, and as such is perhaps best not viewed as a non-monotonic *logic* so much as simply a first-order technique for non-monotonic *reasoning*.

McCarthy describes circumscription as a "rule of conjecture" as to what objects have a given property P. A useful example exploited is the familiar "missionaries and cannibals" puzzle: Three missionaries and three cannibals must cross a river, using a boat that can hold only two persons; if the cannibals outnumber the missionaries on either bank of the river, the missionaries will be eaten. How can the crossing be arranged safely? Now, there are numerous features of interest in the puzzle. The one of concern here is that it is in fact a puzzle, i.e., the puzzler is expected to recognize certain implicit ground rules, such as that the boat does not have a leak or any other incapacity for transporting people. Moreover, there are no

additional cannibals or missionaries lurking in the background, who may upset otherwise sound plans, even though it was not specifically stated that there are *only* three cannibals and three missionaries. It is as if there is an implicit assumption that if something is not mentioned in the puzzle then it is not to be considered, i.e., the closed-world assumption. It corresponds to minimizing the number of objects having certain properties. In effect we are considering conjectures that for certain properties P, an object x does not have P unless it is required to do so. Moreover, this sort of minimizing assumption appears to be very useful even in non-puzzle situations. Circumscription provides one way to make this rather vague idea precise.

Circumscription involves the use of an axiom schema in a first-order language, intended to express the idea that certain formulas (wffs) have the smallest possible extensions consistent with certain given axioms. To illustrate, if B is a belief system* including world knowledge W and specific domain knowledge A[P] concerning a predicate P, then it may be desired to consider that P is to be minimized, in the sense that as few entities x as possible have property P as is consistent with A[P]. The world knowledge W together with A[P] and the circumscriptive schema, are used to derive conclusions in standard first-order logic, which then may be added to B (hopefully consistently and appropriately). It is this notion of consistency with a part of the belief system itself that causes conceptual as well as computational problems in non-monotonic reasoning, essentially problems of self reference. McCarthy has found a very ingenious way of finessing such self reference in the context of minimization, allowing a mechanical means of establishing the effect of consistency tests in certain cases.

As suggested above, given a predicate symbol P and a formula A[P] containing P, the minimization of P by A[P] can be thought of as saying that the P-objects consist of certain ones as needed to satisfy A[P] and no more, in the sense that any tentative set of P-objects x (such as those given by a wff Zx such that A[Z] holds) already includes all P-objects.

Circumscription expresses this by means of a schema or set of wffs, which we denote here by A[P]/P, as follows:

$$A[P]/P = \{[A[Z] \ \& \ (x)(Z(x) \dashrightarrow P(x))] \dashrightarrow (y)(P(y) \dashrightarrow Z(y)) \mid Z \text{ is a wff}\}$$

(Here A[Z] results from A[P] by replacing P by Z.)

A key example, a variation on one emphasized by McCarthy, is the following: let A[P] be a=/=b &. P(a) v P(b). Let Z1(x) be x=a and Z2(x) be x=b. Then from P(a) v P(b) we get that either Z1 or Z2 will serve for circumscription, i.e., either Z1(x) --> P(x) and hence P(x) --> Z1(x), or Z2(x) --> P(x) and hence P(x) --> Z2(x). Thus either a is the only P-object, or b is; indeed, -P(a) v -P(b) will then be provable from A[P] + A[P]/P. In fact, it then follows that there is a unique P-object; this however should not cause concern, for the intention is to explore the consequences of conjecturing the stated minimization of P.

McCarthy [1984] generalized his original notion of (predicate) circumscription to allow specified predicates other than P to vary as well as P; this decisively extends the range of applicability of circumscription. In the new formulation, called formula circumscription, the schema can be replaced by a single second-order formula, but comparison with predicate circumscription is easier when a schema or set A[P1,...,Pn]/E is retained, in the following form:

$$\{A[Z1,...,Zn] \ \& \ (x)(E[Z1,...,Zn] \dashrightarrow E) \dashrightarrow (x)(E \dashrightarrow E(Z1,...,Zn)) \mid \text{wffs } Z1,...,Zn\}$$

where E = E[P1,...,Pn] is a formula in which P1,...,Pn may appear, and E[Z1,...,Zn] is obtained from E by substituting Zi for each Pi. Here the intuitive idea is to minimize (the extension of) the formula E, by allowing variations in (the extensions of) P1,...,Pn.

As McCarthy has observed, it is the presence of the parameters P1,...,Pn that gives formula circumscription its power, and not the fact that E may be a formula. Indeed, form-

ing an extension-by-definitions of A[P] by adding the new axiom (x)(P0x<-->Ex) where P0 is a new predicate letter, one can simply circumscribe P0 with P0,...,Pn as parameters in the extension of A[P]. That is, we can just as well take E to be a single predicate letter P0, since any formula that we may wish to minimize can be made equivalent to such a P0 by means of an appropriate axiom included in A[P] itself. Thus we will employ this version of circumscription, which perhaps is best called parameter circumscription. In the sequel then, E is the predicate letter P0, and P stands for P0,P1,...,Pn, i.e., E plays the role of P0 above, unless context dictates otherwise. Then the schema A[P]/P is as above except that the parameters P0,...,Pn appear rather than simply P1,...,Pn, and the wffs Z0,...,Zn as well, again where P0 is E[P0,...,Pn] and Z0 substitutes for E[Z0,...,Zn]. To be precise, A[P]/P will be the set of wffs

$$\{[A[Z0,...,Zn] \;\&\; (x)(Z0x \dashrightarrow P0x)] \dashrightarrow (y)(P0y \dashrightarrow Z0y) \mid Z0,...,Zn \text{ are wffs}\}$$

The theory obtained from A[P] by adjoining the set A[P]/P as new axioms, will be abbreviated with the notation A[P]* whenever the P can be understood from context. I.e., A[P]* = A[P] + A[P]/P.

An example using formula or parameter circumscription, is the following "Life and Death" problem: Let A[D,L] be the axiom

$$(x)(Dx \leftrightarrow \neg Lx) \;\&\; La \;\&\; Db \;\&\; Kc \;\&\; (a \neq b \;\&\; a \neq c \;\&\; b \neq c)$$

which intended to have the interpretation that dead things (D) are those that are not living (L), and a is living, b is dead, and c is a kangaroo (K). The circumscription of D then corresponds to the notion that as few things as possible are to be considered dead. However, using mere predicate circumscription, i.e., A[D]* rather than A[D,L]*, D could not be "squeezed" down by means of an appropriate Z predicate since L, being unchanged, would force D to be its unchanging complement. Thus A[D]* would not have either Dc or Lc as

theorems. On the other hand, A[D,L]* does have -Dc, and hence Lc, as theorems. This can be seen by circumscribing with the two predicates x=b (for Z0) and x=/=b (for Z1).

Aside from giving examples, it is desirable to show in precise terms in what sense the circumscriptive schema A[P]/P does in fact minimize. For this purpose McCarthy [1980] proposed the concept of minimal model in the context of predicate circumscription. Etherington [1982] has re-defined minimal model in a manner appropriate to McCarthy´s new (formula) version of circumscription, which is presented here in slightly modified form as follows. Let M and N be models of A[P] = A[P0,P1,...,Pn] with the same domains and the same interpretations of all constant, function, and predicate symbols except possibly P0,P1,...,Pn. M P-reduces N if the extension of P0 in M is a proper subset of that in N. Then N is a P-minimal model of A[P0,...,Pn] if N is a model of A[P0,...,Pn] and no model M of A[P0,...,Pn] P-reduces N. (By ´model´ here is meant ´normal model´, i.e., a model in which equality is interpreted as identity. This, incidentally, illustrates the pointlessness of choosing P0 to be the equality predicate, for then two distinct elements necessarily cannot be identical and so all (normal) models are minimal for equality.)

As an example, consider again McCarthy´s axiom A[P]: a=/=b &. Pa v Pb. Here P0 is just P. It is easily seen that the P-minimal models are precisely ones of the form {Pa a=a b=b c1=c1 c2=c2 ...} or {Pb a=a b=b c1=c1 c2=c2 ...} where the number of ci´s may be none or any other cardinality. In particular, M1 = {Pa} and M2 = {Pb} are two such models. But {Pa Pb}, although it is a model of A[P], is not minimal.

The clearly desirable situation would be to have a definition of model appropriate to the proof theory of the circumscriptive schema, i.e., affording a completeness result of the form: B is a consequence of A[P] by circumscription, i.e., a theorem of A[P]/P, iff B holds in all P-minimal models of A[P]. That this does not hold in general, as will be discussed below,

indicates that at present there are unclear areas in the foundational status of circumscription.

First however is stated a positive result, variants of which have been given in Davis [1980] (for what is often called ´domain´ circumscription), in McCarthy [1980] (for predicate circumscription), in Minker and Perlis [1984] (for ´protected´ circumscription), and extended by Etherington [1983] to formula circumscription.

Soundness Theorem: For any formula B

$$A[P] \mid\!\!-\!\!-_P B \text{ implies } A[P] \mid\!\!=\!\!=_P B$$

where P is a vector of predicate symbols P0,P1,...,Pn and the P-single-turnstyle and P-double-turnstyle mean the antecedent together with the circumscriptive schema A[P]/P has the consequent as a theorem, and that the consequent holds in any P-minimal model of the antecedent, respectively. (Note that $A[P] \mid\!\!-\!\!-_P B$ is the same as $A[P]^* \mid\!\!-\!\!- B$.)

Again, the example above will illustrate this. Since $A[P] \mid\!\!-\!\!-_P \neg Pa \vee \neg Pb$ as we saw earlier, then it follows that $\neg Pa \vee \neg Pb$ holds in the models M1 and M2. Of course, we also see directly that this is the case.

Unfortunately in general the converse, which would provide a full completeness theorem, does not hold, as shown by Davis [1980]. Let A[N] be Peano arithmetic (with the postulates N(0), $(x)(N(x) \rightarrow N(x+1))$, etc.) Then the N-minimal models contain N-extensions isomorphic to the natural numbers, so that the formulas B relativized to N that are true in these models are precisely those which are true in arithmetic. But no recursive first-order theory, including one of the form $A[N]^* = A[N] + A[N]/N$, has as its theorems precisely those sentences true of the natural numbers, nor even its N-relativized theorems. [Etherington, Mercer, and Reiter (6) noticed that for certain other arithmetical theories A[P] considered

in (4), A[P]* may be inconsistent even though A[P] is consistent. Specifically, A[P] may fail to have minimal models.]

Kueker [1984] has found the following simpler illustration: Let I[P] be the theory Pa, Px<-->Psx, a =/=sx, sx=sy-->x=y. Then models of I[P] are of two types: those that satisfy the sentence (x).Px-->[-(Ey)x=sy --> x=a] and those that do not. But any minimal model is isomorphic to the natural numbers N, and is of the former type. Kueker has shown that this sentence is not a theorem of I[P]* = I[P]+I[P]/P, which demonstrates that I[P] is not P-complete. The obvious candidate for Zx in the circumscriptive schema, namely

$$Px \;\&.\; -(Ey)x=sy \dashrightarrow x=a,$$

achieves nothing.

Nevertheless, certain partial converses do hold, which have rather broad application. First some terminology. A theory A[P] is P-complete if A[P] |P= B implies A[P] |P-- B for all B, i.e., if the converse to the Soundness Theorem holds for A[P] and P0. (Note that the full converse to Soundness, which is false, is simply the assertion that every theory is P-complete for every wff P0.) A[P] is P-characterizing if it has theorems of the form

$$(x)(Pix<-->Wi1x) \vee \ldots \vee (x)(Pix<-->Wikix)$$

for each i=0,...,n where the W´s do not involve P0,...,Pn.

Minker and Perlis [1985] exploit these concepts in the following partial completeness result: If A[P]* is P-characterizing then A[P] is P-complete. As a special case, the following can be obtained as a corollary: If A[P] has only finite models, then for all sentences B, A[P] |P= B iff A[P] |P-- B.

As with much of commonsense reasoning techniques, circumscription naturally

presents itself as a candidate for a reasoning mechanism that could in principle be used in an intelligent robot, for instance in conjunction with a theorem prover. However, the fact that a schema or infinite set of axioms is involved, presents practical difficulties, especially in the necessary choice of which instance(s) of the schema to use. That is, efficiency or effectiveness questions arise.

In this regard, Lifschitz [1984] has shown the significance of a subclass of theories vis a vis circumscription: the separable theories. Separable theories A[P] are those which are formed, using conjunctions and disjunctions, from formulas containing no positive occurrences of P0 and formulas of the form

$$(x)(E(x) \dashrightarrow P0(x))$$

where E is a predicate that does not contain P0. (These appear related to the P-characterizing theories, and may afford fruitful terrain for further investigation.) Such theories turn out to afford expression by means of a single wff replacing the (infinite set of wffs of the) circumscriptive schema, thereby avoiding the problem of selecting an instance of the schema. In effect, Lifschitz finds an instance (for separable theories) that is optimal.

E. Real-time introspection

Now we take another point of view. If smart systems do in fact perform default reasoning, and if they do do so by effective means (as they must), then what is decided is not derivability in the logical sense, but something different. Israel [1980] argues that a sequence of logics is a better way to view the situation. Just such an approach is undertaken in an experimental reasoning system studied by Perlis [1981, 1984]. Similarly, Konolige [1982, 1984] has explored certain modes for representing deductions agents may perform, without

investing them with full logical consequence as their means of inference. Here we wish to suggest a specific and yet quite general kind of self-referential inference that can be applied to the above ideas about defaults.

Consider a reasoning system S, in which the deductions S performs are seen as occurring over time. Here we do not want to assume that at some time S stops inferring new things. We picture S as a 'computer individual'--to borrow a phrase from Nilsson--that goes on thinking as it interacts with the world. Thus there is no time at which it has 'got' all its conclusions. However, at any time, there are conclusions it has, and ones it has not. We endow S with a mechanism Introspect(X), that allows it to infer whether it already knows X. The result of performing Introspect(X) is that either Know(X) or -Know(X) will appear in S's database, depending on whether in fact X was or was not already in S's database. It matters not that S may know A and A-->X; as long as X itself is not present as a separate item in the database, -Know(x) is returned.

This seemingly rather obtuse procedure in fact is anything but. For recall that S is continuously thinking, deducing. If at any moment before the Introspection is done, X is inferred, then Know(X) will appear instead of -Know(X). And if it is of sufficient interest to S to apply Introspect at all, then it is most likely that S already will have been trying to decide X, and would have proven X from such simple axioms as A and A-->X before a complex default procedure could have finished.

Specifically, imagine S wanting to know whether X. S both considers proving X and denying X (proving -X). If S knows A and A-->X, and also a default rule such as -Know(X) --> -X, then S can more quickly prove X then invoke the default. Even if Introspect is invoked before X is proven, the default still will not yet have been accomplished, for another step is needed. And if by some chance the default is invoked to prove -X only later to be con-

tradicted by the direct proof of X, we still have a conflict resolution to fall back on, namely, that not only are X and -X in conflict, but also are Know(X) and -Know(X); now the latter is easy. If Know(X) has been deduced from X, then it takes precedence over -Know(X). Once this is done and -Know(X) is removed, the basis for -X is gone and it too can be removed. So in any case, we end up with X and Know(X) as desired.

The lesson we derive from this is as follows. The database changes in many ways. Yet S needs to keep reasoning *as* it changes, and *with* its axioms that are changing: about it, with it, as it changes.

We have implemented a system that operates along the lines specified above, with an introspection device for self-knowledge. In order to keep the introspections rapid, only a small ´working subset´ of the entire database is searched when an introspection is performed: the set of ´currently used´ beliefs. These are kept in a queue that is updated in a least recently used basis, in analogy with human short-term memory. The entire architecture is highly reminiscent of the production systems of Newell and Simon [1972], and indeed early experiments with our system show that a "short-term memory" size that works best for trial problems to date is approximately that of human "short-term memory" as well.

IV. Conclusion

We have discussed the implications of self-reference with regard to both language and beliefs. As stated earlier, the problem of self-reference must be dealt with in order to provide a computer with any reasonable sense of intelligence. As evidenced from the preceeding it is clear that this is not an easy problem. On the other hand, however, there is a good deal of evidence indicating that the current research is making significant headway.

Bibliography

Bobrow, D. and Winograd, T. [1977] An overview of KRL, Cog. Sci. 1.

Brachman, R. [1978] A structural paradigm for representing knowledge, Rep. No. 3605, Bolt Beranek & Newman.

Davis, M. [1980] The mathematics of non-monotonic reasoning. Artificial Intelligence, 13 (1,2), pp. 73-80.

Etherington, D. [1983] Formalizing non-monotonic reasoning systems. University of British Columbia, Dept. of Computer Science, Tech. Report 83-1.

Gilmore, P. [1974] The consistency of partial set theory without extensionality, in T. Jech (ed.), Axiomatic Set Theory, Amer. Math. Soc.

Israel, D. [1980] What's wrong with non-monotonic logic? Proc. AAAI-80, pp. 99-101.

Kripke, S. [1975] Outline of a theory of truth, J. Phil 72.

Kueker, D. [1984] Another failure of completeness for circumscription, (photocopied notes), Week on Logic and Artificial Intelligence, Univ. of Maryland, Oct. 22-26.

McCarthy, J. [1980] Circumscription--a form of non-monotonic reasoning. Artificial Intelligence, 13 (1,2), pp. 27-39.

McCarthy, J. [1984] Applications of circumscription to formalizing common sense knowledge, Workshop on Non-monotonic Reasoning, Mohonk, Oct. 17-19.

McDermott, D. [1982] Non-monotonic logic II: non-monotonic modal theories. Journal of the ACM, 29 (1), pp. 33-57.

McDermott, D. and Doyle, J. [1980] Non-monotonic logic I. Artificial Intelligence, 13 (1,2), pp. 41-72.

Minker, J. [1982] On indefinite databases and the closed-world assumption. Lecture Notes in Computer Science, v. 138, pp. 292-308, Springer. (6th Conference on Automated Deduction).

Minker, J. and Perlis, D. [1984] Applications of protected circumscription. Lecture Notes in Computer Science, v. 170, pp. 414-425, Springer. (7th Conference on Automated Deduction).

Minsky, M. [1974] A framework for representing knowledge. MIT AI Lab Memo 306.

Moore, R. [1983] Semantical considerations on non-monotonic logic. Proc. 8th IJCAI, pp. 272-279.

Newell, A. and Simon, H. [1972] Human Problem Solving. Prentice Hall.

Perlis, D. [1981] Language, computation, and reality. Ph.D. Thesis Univ. of Rochester.

Perlis, D. [1984] Non-monotonicity and real-time reasoning. Workshop on Non-monotonic Reasoning, Mohonk, Oct. 17-19.

Perlis, D. and Minker, J. [1985] Completeness Results for Circumscription. To appear, Artificial Intelligence J.

Reggia, J. and Nau, D. [1984] An abductive non-monotonic logic. Workshop on Non-monotonic Reasoning, Mohonk, Oct. 17-19.

Reiter, R. [1978] On closed world databases. In: Logic and Databases, Gallaire, H. and Minker, J. (eds.), Plenum, pp. 55-76.

Reiter, R. [1980] Equality and domain closure in first-order databases. Journal of the ACM 27 (2), pp. 235-249.

Reiter, R. [1980a] A logic for default reasoning, Artificial Intelligence 13 (1,2), pp. 81-132.

Reiter, R. and Criscuolo, G. [1981] On interacting defaults. Proc. 7th IJCAI, pp. 270-276.

Russell, B. [1908] Mathematical logic as based on the theory of types, Amer. J. Math. 30.

Tarski, A. [1944] The semantic conception of truth and the foundations of semantics, Philos. and Phenom. Res. 4.