AUTOCIRCUMSCRIPTION

Donald Perlis

Department of Computer Science
and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

Abstract

Reasoning can be used to select among various possible interpretations of events. But how are these possibilities determined? We isolate two key technical features of circumscription (consistency and minimization), and use the first as the basis for a reformulation of the circumscription principle in a way related to possibility, self knowledge, and negative introspection. The second (minimization) then can be separately expressed on its own. Conceptual clarity and a kind of validity are results of this separation, as well as a computational means to determine (sometimes) when a wff is *not* among a reasoner's conclusions.

I. Introduction

Assessing possible states of affairs with respect to given knowledge is an essential feature of making sense of the world. This is because our knowledge is generally very incomplete, so we need to reason about the gaps in our knowledge. Once we have isolated various possibilities, we may choose among them, perhaps selecting one as a plausible conjecture until evidence convinces us to leave it. Thus such conclusions are defeasible, and this provides a basis for non-monotonic reasoning.

But a question that in some sense comes first, is: How do we determine that a wff is indeed possible, with respect to what we do know? One idea is that to recognize a gap (ignorance) in our knowledge *is* to recognize alternative possibilities. The absence of knowledge that, say, α, amounts to the possibility that $\neg\alpha$. In effect, we are theory-builders; we build tentative theories about the world on the basis of what we determine that it is consistent to postulate. That is, if we can find a consistent interpretation of our concepts, we may view it as a possibility: x is possible if we do not know (cannot prove) x is false.[1] Once we have found such an interpretation, we may try to assess whether it is worth taking seriously. This bears on another fact, namely that consistency-testing is in general undecidable. This is in fact why most formalisms for non-monotonic reasoning do not have effective proof procedures. Circumscription has the virtue of being the one

[1] In [17] I discuss some further concepts of possibility in commonsense reasoning.

fairly general (semi)-decidable formalism available. This will lead us to some insights into the prior problem of determining possibilities.

In [12] McCarthy introduced the idea of circumscription, and it was soon recognized as a powerful and important new technique in artificial intelligence and logic. However, the foundational and conceptual status of circumscription, in its various versions[2], has remained unsettled (see Davis [1], Etherington et al [2], Lifschitz [8,9]; McCarthy [12,13], Perlis & Minker [22]) and also bound up with the equally unsettled status of various theories of default reasoning. Here we focus on one of the key advances provided by McCarthy's original insight, namely a way to finesse consistency proofs. McCarthy exploited the intuition[3] that if a true interpretation of certain axioms can be established in such a way that a particular predicate letter P is re-interpreted via a stronger formula Z (i.e., one for which $Z \rightarrow P$ is provable), then Z is indeed a possible reading for P, i.,e., $P \rightarrow Z$ (and hence $P \leftrightarrow Z$) is consistent with those axioms. He further observed that this establishing might be possible within the very theory itself having those axioms. Thus McCarthy has discovered a technique for determining, at least in some cases, when a particular wff is *not* a logical consequence of others.[4] This is a very striking result, for in general the logical consequences of a set of wffs are at best only semi-decidable; thus there is hope now that in many cases of interest to commonsense reasoning, a computationally viable mechanism may be available to determine when a wff is *not* one of those among what a particular agent knows (can conclude). This is of importance because of the central role that ''what I don't know'' plays in non-monotonic reasoning in general.

Now, McCarthy and others have studied this from the point of view of asserting $P \rightarrow Z$ once Z is established as an alternative possible interpretation of P. The typical conclusion then has the form: from $\neg Zx$ conclude $\neg Px$ (since $P \rightarrow Z$). This has seemed appropriate to the main goal of modeling minimizing assumptions as in default reasoning. But formulations to date have made the passage from the possibility of Z to $P \rightarrow Z$ (and hence to $\neg Px$ given $\neg Zx$) in one fell swoop, rather than first recording the possibility as a result in its own right and then with a further axiom (when desired) basing the minimizing of P via $P \rightarrow Z$ on the former. However, there are at least three advantages to separating the possibility (or

---

[2]At my last count there were at least eight versions of circumscription on the AI market. My apologies for introducing yet another here. I offer the mitigating plea that *this* one is not really circumscription at all, in the sense that it does not aim at minimizing extensions, though it does borrow outright the truly novel portion of McCarthy's original schema.

[3]That is *almost* true (see [1,2]). Shepherdson [25] uses a similar idea to give ''inner models'' of set theories.

[4]In effect, circumscription involves a relative consistency proof, that identity of P and Z is consistent with (and relative to) A[P].

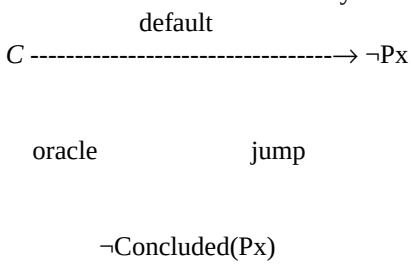
relative consistency) of ¬Px given ¬Zx from the *conclusion* that in fact ¬Px holds (given ¬Zx):

1. It is not always the case that one wishes to minimize all possible interpretations, yet one might still want to know of their viability. Indeed, McCarthy early on [12] characterized circumscription as ‘‘a rule of conjecture,’’ which suggests something tentative like possibility or viability rather than a definitive conclusion.

2. This is directly related to negative introspection: how is it determined that a given wff is *not* among the reasoner’s conclusions? While this is in general undecidable, many cases of it are of critical importance in commonsense reasoning. An approach to possibility then might afford a computational means to decide (many such) cases. This is the main issue addressed in this paper.

3. The semantics of possibility are simpler to study, and in particular the conclusion of possibility is *sound*: if there is a true interpretation Z of P for which, say, Zb is false, then Pb *cannot* be provable (unless the original theory is inconsistent).

These (and especially 2.) will be taken up at greater length below. In the present section we aim mainly at discussing some of the choices available for formal mechanisms involved in such an undertaking. What we wish to do, then, is formalize the idea that if there is a true interpretation of a set of axioms, in which P is re-interpreted as Z, then from ¬Zx we conclude only ¬Concluded(Px). Clearly this is different from concluding that ¬Px. An overview of the theme we are pursuing can be indicated in the following diagram, where *C* stands for some mechanism that sets up conditions under which it may be desirable to conclude non-monotonically that ¬Px.



Here *C* could be a portion of McCarthy’s circumscription schema (as used below), or appropriate aspects of Reiter’s [23] or McDermott & Doyle’s [14] forms of non-monotonic reasoning. The point is that the conclusion of the default has been broken into two steps, first recognizing explicitly the fact that the default conclusion is not prevented by anything known, and then the actual passage to the default. In Perlis [19] these steps are called the ‘oracle’ and the ‘jump’ respectively. However, here we will be exploring the extent to which a variation on circumscription can actually render the oracle computationally feasible.[5]

[5]The name ‘oracle’ was chosen before to indicate an intractable (even undecidable) problem -- namely, that of determining non-provability -- so that now it may be inappropriate.

Note that a fully ''internal'' solution is impossible within a fixed theory T. That is, Gödel [4] and Löb [11] showed that if T is consistent, then under very general conditions $|\!\!/\!\!-_T \neg Thm_T(\alpha)$ for *any* wff α. Thus we must augment T, in order to get conclusions such as $\neg Thm_T(\alpha)$. Even here we are stymied by the fact that the wffs α for which $|\!\!/\!\!-_T \alpha$ form a semi-undecidable set in general. However, we may succeed in getting many particular cases, and this is what circumscription will allow us to achieve.[6]

II. An Introspective Treatment

We begin by recalling one standard form of circumscription, the so-called predicate[7] circumscription schema (McCarthy [12]) where Z is an arbitrary wff:

$$(1) \qquad A[Z] \,\&\, (\forall x)(Zx \rightarrow Px) \;.\!\rightarrow.\; (\forall x)(Px \rightarrow Zx)$$

Here A[P] is given as (a conjoined set of) axioms, and A[Z] is this conjunction rewritten by substituting Z for P. It is natural to think of Z as a candidate interpretation for (the underdetermined) P. For instance, suppose A[P] is the conjunction of Pb and $(\forall x)(Qx \rightarrow Px)$; this might formalize the idea that b is winged, and that flyers are winged. However, A[P] really says very little as to what winged things (P-things) are: all we know is that b is one, and flyers (Q-things) are winged. This leaves open what the extent of winged things is. Possibly b is the only such, possibly not. The predicate circumscription schema above has as consequence that, after all, the class of winged things is the smallest class Z satisfying A[Z], i.e., the union of the class of flyers and the object b. For we need only choose Z to be the formula Qx v x=b. Then A[Z] and $Zx \rightarrow Px$ follow from A[P], so the schema yields $(\forall x)(Px \rightarrow Zx)$, i.e., $(\forall x)(Px \rightarrow .\, Qx \text{ v } x=b)$.

Now suppose we wish to know that Qx v x=b is a *possible* interpretation of P, i.e., that winged things *may be* precisely b and the flyers, but we are not prepared to conclude, yet, that this is *true.* Then we must avoid the final part of the schema, namely $(\forall x)(Px \rightarrow Zx)$. What can we put in its place? We presumably want something like Possible$(\forall x)(Px \rightarrow Zx)$. However, there is a simpler and more useful formulation, namely we first rewrite the consequent in contrapositive form and rearrange to emphasize the negative conclusion:

$$A[Z] \,\&\, (\forall x)(Zx \rightarrow Px) \,\&\, \neg Zy \;.\!\rightarrow.\; \neg Py$$

[6]Actually, because we will not use Thm but rather another predicate, the Gödel-Löb result might not apply; see Perlis [21] for a discussion of these issues in a general setting.

[7]Below we will switch to *formula* circumscription; our comments here apply equally to both versions.

and then instead of the boldness of ¬Px we use ¬Concluded(Px), i.e., rather than assert that Px is *false*, we simply record that it is not concluded that Px, where for brevity we employ the symbol K rather than Concluded:

$$(2) \qquad A[Z] \,\&\, (\forall x)(Zx \rightarrow Px) \,\&\, \neg Zy \;.\rightarrow.\; \neg K(Py)$$

Thus this amounts to a kind of meta-conclusion about the reasoner's own conclusions: the reasoner is (to be) endowed with the ability to introspect that it has not been able to draw certain conclusions. Determining the truth of A[Z] can be seen then as a discovery that we have built a viable theory or possible world. That is, A[Z] can be regarded as asserting that Z is a possible interpretation of P, or represents a possible world with respect to the knowledge A[P], since it says that Z satisfies all that we know about P anyway (namely, A[P]); then if y does not satisfy Z (P's possible interpretation), Py cannot have been known. The advantage to using ¬K(Py) -- or Possible(¬Py) -- instead of Possible($\forall$x)(Px $\rightarrow$ Zx) will emerge below.

There is another question of interpretation that arises. Namely, what is the role of the subformula ($\forall$x)(Zx $\rightarrow$ Px) here? In circumscription proper, it is essential in dealing with disjunctive axioms (see McCarthy [12]).[8] But if we seek simply to determine ignorance rather than to minimize, its appropriateness is unclear. There are two viewpoints that naturally arise: we can dispense with this subformula altogether, seeking the most general range of possible interpretations of P, or we can retain it with an eye to possible future use in minimizations or defaults.

This is worth exploring a bit more. Suppose that our axioms A[P] consist of Pa v Pb and a≠b. Then a reasonable ignorance-tester should show that neither Pa nor Pb is concluded from A[P]: ¬KPa & ¬KPb. On the other hand, if we accept as a default (or P-minimizing) rule that ¬K(Px) $\rightarrow$ ¬Px, we immediately run into inconsistency (which is what prompted McCarthy's use of the additional subformula in question). Now, this does not mean that it is a mistake to state that ''neither Pa nor Pb is concluded from A[P].'' It simply means that such observations do not lend themselves directly to default reasoning. Hence if our interest is primarily in determining what is (not) concluded, it may be appropriate to dispense with the subformula; and for later use in defaults it may be appropriate to retain it and thereby find fewer candidate wffs determined to be unconcluded. Of course, we could just as well undertake *both* approaches at once, employing different notations: GenK and MinK -- general concludability and minimizing concludability. It is the latter (MinK, which

---

[8] If P is viewed as *unusual* this says P's substitute Z introduces no new unusual entities.

is (2) above) that bears direct relation to standard circumscriptive formalisms and default reasoning, namely we have the obvious result that

(3) $\qquad (2) + \neg K(Py) \rightarrow \neg Py \quad \text{entails} \quad (1)$

In what follows, except for one explicit application to default reasoning, we will avoid the added complexity of juggling two versions, and employ a single version (GenK, (4) below) without the subformula, for conciseness of exposition. However, much of what we say will apply equally to the alternate version. Thus the version we will study has the form

(4) $\qquad A[Z] \,\&\, \neg Zy \;.\!\rightarrow.\; \neg K(Py)$

whose purpose is not to elucidate circumscription so much as to borrow a portion of the underlying idea of circumscription in order to address the negative introspection problem.[9]

Before proceeding into details, we mention that Konolige [6] and Levesque [7] have undertaken similar tasks. Konolige studies the problem of drawing conclusions on the basis of knowing what an agent does not know. He uses a modal *propositional* logic for this purpose, and thus can retain decidability; this represents a rather limited language, however. Levesque pursues the same goal via a special modal semantics that does not have (in its quantificational version) a corresponding semidecidable proof-theoretic component; this serves his purpose since his logic is not intended as an effective calculus for a reasoning agent. Autocircumscription, on the other hand, is so intended, and it is important then that it be semidecidable (and it is) even though a quantificational language is used.

III. Technical Details

There is one obvious aspect in which a complication arises: the syntax of the underlying first-order language must include names for wffs, so that they can appear as terms in formulas. That is, we must reify wffs as first-order objects. This is not new, however; ways to do this are given in (Feferman [3] and Perlis [18]). Feferman meets the technical requirement as follows: for each wff w in the language, there is a designated term t_w whose free variables are those of w. Thus t_w is a function symbol with variables (or constant symbol if w has no free variables).[10] Call a language 'reified' if it is so endowed with terms. Note that since t_w is itself in the language, it gives rise to other terms naming wffs in which t_w

[9]Vladimir Lifschitz has suggested an elegant generalization of (4), along the lines of $K(Px) \leftrightarrow (\forall p)(A(p) \rightarrow px)$, where p is a second-order predicate variable. This brings out the idea of possible worlds more forcefully and provides positive introspection as well.

[10]This device is sometimes called 'quasi-quoting.'

appears. Since context makes clear the usage, we will simply use ‘w’ for t_w, or even just *w* itself; however, t_w itself, being a term, does not contain the wff w, nor any predicate letters at all.

Thus we may employ a predicate expression K(‘w’) where ‘w’ is the name of a wff, to mean that the wff (named) ‘w’ is concluded by our reasoning agent g. The predicate symbol K can just as well be read as ‘‘g knows (or believes)’’ its argument, or better yet, ‘‘I know’’ the argument. If w has free variables, then the wff K(t_w) (also written as K(‘w’) above and as Kw) does too; this allows for quantifying into what otherwise might appear to be opaque contexts.

Suppose then that L is a reified first-order language, with predicate symbol K. Then we can present a revised schema, in a form we call *autocircumscription* (for it is designed to isolate the feature of determining what is (not) known to the agent *itself*, rather than what is (not) true in the outer world). For greater generality we turn to *formula* circumscription (McCarthy [13]), written in a first-order version. Let A[P] be a finite conjunction of wffs of L, where P is a tuple of predicate letters Pv’0.4m’$_0$,...,P_n appearing in A[P]. Let Z be the tuple of wffs Zv’0.4m’$_0$,...,Z_n, and W[P,x] any wff to be minimized (and in which the predicate letters P and variable(s) x may figure). Then one version of formula circumscription is:

$$A[Z] \;\&\; (\forall x)(W[Z,x] \rightarrow W[P,x]) \;\&\; \neg W[Z,y] \;\; .\rightarrow. \;\; \neg W[P,y]$$

The idea is that the predicate letters Pv’0.4m’$_0$,...,P_n in the original axiom A[Pv’0.4m’$_0$,...,P_n], are open to interpretations Zv’0.4m’$_0$,...,Z_n, respectively (which are to be chosen suitably, and where care is taken to avoid clash of variables -- see Mott [16]). Then if W[P] is ‘‘minimal’’ with respect to A[P], and fails at y under the Z-interpretation, W must also fail at y for the original P.[11]

The corresponding *autocircumscription* schema for an axiom set A = A[Pv’0.4m’$_0$,...,P_n] is:

AUTO: $$A[Z] \;\&\; \neg W[Z,y] \;\; .\rightarrow. \;\; \neg K\, W[P,y]$$

The formula of which ignorance is being tested is represented by W[P,y]. The idea, as before, is that if there is an interpretation Zv’0.4m’$_0$,...,Z_n of the predicates Pv’0.4m’$_0$,...,P_n such that A[Zv’0.4m’$_0$,...,Z_n] holds, then this is a possible interpretation, so that we must have been ignorant of any fact about W[P,y] that happens to fail for W[Z,y].

---

[11]Thus W[P,x] is a special wff *designated* for minimization.

One feature of autocircumscription is that, unlike minimizing versions of circumscription, no special license is needed in choosing the wff W. In fact, *all* wffs can be ignorance-tested at once with impunity. Thus we regard autocircumscription as a schema not only for Z but also for W.[12] For the same reason, in the schema we can assume that Pv'0.4m'$_0$,...,P_n are *all* the predicate letters in A[P]. Thus we do not bother to write A[P] anymore but just A. We introduce the notation AUTO[A] to stand for the above schema (over *all W's*) together with A itself, making AUTO[A] an extension of A: AUTO[A] = A + AUTO.

Now, what is preferable about this version over the one with Possible(∀x)(W[P,x] → W[Z,x])? Well, the latter would require us to write axioms for Possible in a way that would to break apart (∀x)(W[P,x] → W[Z,x]) into subformulas so that knowing, say ¬W[Z,c], would allow a conclusion about the possibility of ¬W[P,c]. Schema AUTO avoids this by placing K directly where desired.

Another idea is to minimize the predicate K itself with ordinary circumscription, rather than go to the trouble of inventing a new version. However, to make it useful, axioms relating K to the actual provability conditions of the underlying theory would be necessary, and this is not at all easy.[13] The present approach, on the other hand, requires no particular axiomatization of K, for it serves via AUTO simply to record when a wff is *not* a theorem of A. If A does have theorems of the form Kα, this is not necessarily a problem, as long as it respects the intended meaning of K, as we now define.

Definition: A theory T is *autoconsistent*[14] (for the predicate letter K) if the language of T is reified and has the predicate letter K and for every wff α

1. T |– Kα implies T |– α and
2. T |– ¬Kα implies T |⊬ α

Definition: A theory B *autoextends* theory A -- or is an autoextension of A, or is autoextensional over A -- for the

---

[12]Certain versions of circumscription, such as in [13] and [20], probably could facilitate corresponding versions of autocircumscription in a single (higher-order or set-theoretic) formula.

[13]See Perlis [21]. Nevertheless, provability provides a key to a sound *semantics* for autocircumscription; this will be taken up in the theorems below. Also, Lin [10] has carried out such a K-minimization in the case of a propositional modal formulation of circumscription.

[14]Note the similarity to the notion of a stable autoepistemic theory (see Moore [15] and Stalnaker [26]); however, our notion is weaker, as it must be by Theorem 7.7 in Perlis [21] showing in effect that suitably reified (substitutive) stable theories are inconsistent. That is, fully (positive *and* negative) introspective consistent theories with self-reference do not exist.

predicate letter K, if B is an extension of A, the language of B is reifi ed and has the predicate letter K, and for every wff α

1. B |– Kα implies A |– α and
2. B |– ¬Kα implies A |⊬ α

Thus a theory is autoconsistent iff it is an autoextension of itself; and theorems of an autoextension B of a theory A do not violate the intended meaning of Kx, namely that x is a theorem (of the original theory A). If A is autoconsistent, then A itself has explicit information as to its own proof theory; sometimes this is too much to expect (see [4,11,21] and the discussion after Theorem 3 below).

Lemma: If T is autoconsistent, then T is consistent. If T is consistent and S autoextends T then S is consistent.

Proof: Suppose T is autoconsistent. Let α be any wff. Then if ¬Kα is not a theorem, T is consistent (since *all* wffs are theorems of inconsistent theories). Yet if ¬Kα *is* a theorem, then autoconsistency requires that α not be, so again T is consistent.

Now suppose T is consistent and S is an autoextension of T. If S were inconsistent then, for any wff α of T, both Kα and K¬α would be theorems of S, hence both α and ¬α would be theorems of T, contradicting T's consistency.

Theorem 1: If T is consistent and does not involve the symbol K in its proper axioms, then AUTO[T] is consistent.

Proof (suggested by a referee): Interpret K as "true" in any model of T; this will automatically satisfy all instances of the autocircumscription schema.

However, we can require more of AUTO[T], namely that it autoextend T, for this is after all the intension behind autocircumscription. We also may want T to have axioms involving K. We then have the following result.

Theorem 2: If the only proper axioms of T = T[Pv'0.4m'$_0$,...,P_n] that involve K are literals, and if T is autoconsistent, then AUTO[T] is consistent and even autoextensional over T.

Proof: Call a wff *K-free* if it does not involve the predicate symbol K. Since T is autoconsistent then by the

Lemma, T is consistent. Let M be a model of the K-free axioms of T. Then interpret 'K(α)' in M as T |– α, for each wff α in the language of T. That is, Kx is interpreted over D=Domain(M) as meaning that x is the name of a wff that is provable in T. This makes M a model of (all of) T, since the literal axioms involving K will be satisfied from the autoconsistency: if Kα is an axiom of T then so is α and thus (by the interpretation of K in M) Kα is true in M. And if ¬Kα is a theorem of T, then (any instantiation over D of) Kis false in M, for if it were true then (that instantiation of) α would be a theorem of T which in turn (by autoconsistency) would bar (the instantiation of) ¬Kα from theoremhood.

But M is also a model of AUTO[T]. For AUTO[T] consists of T plus conclusions of the form ¬K W[P,y] given antecedents T[Z] & ¬W[Z,y] for any vector of wffs Z. We have to show that if the antecedents hold in M then so does the conclusion ¬K W[P,y], i.e., W[P,y] is not a theorem of T for any y in D. But if (some instantiation over D of) W[P,y] were a theorem of T=T[P], then (the corresponding instantiation of) W[Z,y] would be a consequence of T[Z], and so W[P,y] would hold in M. This contradicts the antecedent ¬W[Z,y], which we are assuming to hold in M. This shows that (an instantiation of) W[P,y] cannot be a theorem of T after all. Thus (each instantiation of) K W[P,y] is false in M, so ¬K W[P,y] is true in M. This shows that M is a model for AUTO[T], and therefore AUTO[T] is consistent.

Now AUTO[T] also autoextends T, since if AUTO[T] |– Kα for some wff α of T, then Kα is true in M, so T |– α. And if AUTO[T] |– ¬Kα, then again ¬Kα is true in M so T |⊬ α.

Note that the proof does not make any use of information as to *which* wff W is the one being ignorance-tested. This corroborates our earlier claim that we may as well consider AUTO[T] to *simultaneously* be applied to *all* wffs W.

Corollary: If T is consistent and does not involve K in its proper axioms, then AUTO[T] is also consistent and autoextends T.

These results provide a sharp formal distinction between autocircumscription and the standard (minimizing) varieties. For as has been shown by Davis [1], Etherington et al [2], and Mott [16], peculiarities and even inconsistencies can occur when a circumscription schema (or second-order axiom) is adjoined to certain theories. Indeed, these authors have been at pains to isolate special cases in which consistency is preserved. But for autocircumscription, consistency is

preserved, barring the use of K for defaults. If we try to recapture actual minimizations by adjoining default rules using 'K,' then we again have the full force of minimizing circumscription and so once more the specter of inconsistency can arise.

But even here we have a positive result. We recall a definition from [21]: a *Moorean autoepistemic* (or MAE) wff is one of the form $\alpha \rightarrow K\alpha$. The prototypical example [15] is the sentence ''If I had a brother I would know it.'' These can be regarded as defaults (true instances of having a brother *typically* are known), or, as Moore prefers, as autoepistemic beliefs (true instances of my having a brother *are* known to me). The distinction between these, important for some purposes, is not critical here, so we refer to MAE wffs as defaults.

Theorem 3: If T is autoconsistent and the only proper axioms of T involving K are literals and at most one MAE wff $\alpha_0 \rightarrow K\alpha_0$ where α_0 is a K-free sentence, then AUTO[T] is consistent.

Proof: Construct M as in the proof of Theorem 2, such that if T $\not\vdash \alpha_0$ then M $\not\models \alpha_0$; this is easy since α_0 is K-free. Now M will satisfy the MAE wff above since either T $\vdash \alpha_0$ and so $K\alpha_0$ is true in M, or T $\not\vdash \alpha_0$ and so by construction M $\not\models \alpha_0$; in either case the MAE wff is true in M. Thus M is a model for T and then also for AUTO[T] as in the proof of Theorem 2. So AUTO[T] is consistent.

Unfortunately, we cannot guarantee AUTO[T] to be an *auto*extension T in the above theorem. For instance, if neither α_0 nor $\neg\alpha_0$ is a theorem of T, then we may find both $\neg K\alpha_0$ and $\neg K\neg\alpha_0$ as theorems of AUTO[T]. But the MAE wff and $\neg K\alpha_0$ produce $\neg\alpha_0$, which together with $\neg K\neg\alpha_0$ violates autoextensionality (and of course also autoconsistency). We see this, however, not as a defect of autocircumscription, but rather as a general feature of default (or non-monotonic) reasoning lying outside of ignorance-testing per se.[15]

It is also easy to see that the restriction to *one* MAE default is necessary (in general). For if T has the axioms PvQ, $P \rightarrow KP$, $Q \rightarrow KQ$, then AUTO[T] yields $\neg KP$ and $\neg KQ$, which in turn yield $\neg P$ and $\neg Q$, contradicting PvQ. While this in no way is a comment against $\neg KP$ and $\neg KQ$ (they are literally true about T) it is a comment against the free use of defaults.[16]

---

[15] This issue is the point of another paper, in progress, on the essentially process-oriented nature of default reasoning.

IV. Applications

A. Suppose I have the belief that I am more knowledgeable than Bill about LISP, and in particular that if I don't know some proposition about LISP then neither does he. Now, this will allow me to infer ¬Know(Bill,y) *if* I can first infer ¬Know(me,y). This is where a technical ignorance-prover will be of use. Taking *Know(me,x)* to be the first-order predicate *K(x)*, the autocircumscription schema will facilitate establishing that I do not know (certain cases of) y, so that useful conclusions (Bill's ignorance of y) can follow.

Note that this example is *not* contingent on, or even significantly related to, the separate issue of whether the proposition y happens to be *true*. Nonetheless, interesting conclusions are derivable, namely that it is indeed unknown to me, and also unknown to Bill. Also note that this allows us at the same time to recognize two competing theories about the world: that y is true, and that y is false. While clearly we know y v ¬y already (a tautology), we do not know, without the negative introspection that autocircumscription affords us, that either of these is a possibility, let alone both. Further reasoning then might lead us to accept or reject one or the other of these theories.

B. As our second example, we offer one raised by McCarthy [personal communication, 1984]: How can an agent decide that, on the basis of all it knows, the question as to whether Ronald Reagan is (currently) standing or seated is indeterminate? Here the K predicate together with the autocircumscription schema, solves this problem. For instance, let an agent have the following axioms A[Seated,Standing]:

{Seated(Bill), ¬Seated(Sue), Seated(x) ↔ ¬Standing(x), Ronald Reagan ≠ Bill, Ronald Reagan ≠ Sue}

Then letting Zv'0.4m'$_0$(x) be x=Bill, and Z_1(x) be x≠Bill, we find A[Zv'0.4m'$_0$,Z_1]. But then taking P_0 to be W[P_0, P_1] AUTO[A] will give us $\neg Z_0(y) \rightarrow \neg K(\text{`Seated(y)'})$, from which we get ¬K('Seated(Ronald Reagan)'). Similarly we can show ¬K('Standing(Ronald Reagan)'). Thus both Seated(Ronald Reagan) *and* Standing(Ronald Reagan) have been shown not to be provable from the axiom set A.

C. Our final example is Moore's Brother Problem, alluded to earlier. Let us suppose that we know Fc (Carl, c, is a friend). We postulate the MAE wff Bc → KBc (if Carl is my brother, I will know it). The aim is to be able to derive ¬KBc (I do not know Carl to be my brother) and then ¬Bc (Carl is not my brother); and indeed, more generally, ¬KBx.

[16] See Reiter & Criscuolo [24] and Hanks & McDermott [5] for more on interacting defaults. It would be nice not to have to make this restriction; however, in [19] and [21] evidence is given suggesting that *any* formalism for default reasoning that formally represents too much of its own behavior will face problems of inconsistency.

Thus our theory T[B,F,K] has the axioms Bc → KBc and Fc.

Note that by Theorem 3, AUTO[T] will be consistent: T is consistent and involves K only once, in a single MAE wff applied to the sentence Bc. We show that AUTO[T] |– ¬KBx. We simply interpret Bx as always false; that is, let Z_0x be x≠x. Then T[Z_0,F,K] is readily provable, and so is ¬Z_0x; we get immediately ¬KBx. Consequently from the MAE wff we find ¬Bc, and we have established that Carl is not the agent's brother!

It would be more natural, not to mention convenient, to postulate the more general MAE wff Bx → KBx; however Theorem 3 then does not apply since Bx has a free variable and so is not a sentence. In fact, in this case the above example will fail with disjunctive information such as Bc v Bd. However, the failure will occur not at the stage of negative introspection or oracle (¬KBc & ¬KBd) but at the stage of default conclusion or jump (¬Bc & ¬Bd). Of course, a case like this flies in the face of the original default or autoepistemic belief.

V. Conclusions

Certain more specialized forms of reasoning such as defaults and auto-epistemic conclusions may be viewed as embellishments of theory-building (proving possibility) which in turn lends itself to formalization via autocircumscription. This will not work in every case, due to the undecidability of consistency. But perhaps most cases of interest to commonsense reasoning can be so handled.

Acknowledgement

I wish to thank Halina Przymusinska, Vladimir Lifschitz, Brian Haugh, Kave Eshghi, Martin Davis and anonymous referees for very eye-opening and constructive criticism. This research has been supported in part by the U.S. Army Research Office (DAAL03-88-K0087) and the Martin Marietta Corporation.

References

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

(2) Etherington, D, Mercer, R., and Reiter, R. [1985] On the adequacy of predicate circumscription for closed-world reasoning. *Computational Intelligence 1*, 11-15.

(3) Feferman, S. [1984] Toward useful type-free theories. *J. Symbolic Logic, 49*, 75-111.

(4) Gödel, K. [1931] Uber formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme I, *Monatsh. Math. Phys.*, 38, pp. 173-198.

(5) Hanks, S. and McDermott, D. [1987] Nonmonotonic logic and temporal projection. *Artificial Intelligence, 33*, 379-412.

(6) Konolige, K. [1982] Circumscriptive ignorance. Proc. AAAI-82, pp. 202-204.

(7) Levesque, H. [1987] All I know: an abridged report. Proceedings, AAAI-87, pp. 426-431.

(8) Lifschitz, V. [1986] On the satisfiability of circumscription. *Artificial Intelligence 28*, 17-28.

(9) Lifschitz, V. [1987] Circumscriptive theories: a logic-based framework for knowledge representation (preliminary report). Proceedings, AAAI-87, pp. 364-368.

(10) Lin, Fangzhen [1988] Circumscription in a modal logic, Proc. of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge (Moshe Vardi, ed.) pp. 113-127. Morgan-Kaufmann.

(11) Löb, M. [1955] Solution to a problem of Leon Henkin. *Journal of Symbolic Logic*, 20, pp. 115-118.

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

(13) McCarthy, J. [1986] Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence, 28*, 89-118.

(14) McDermott, D. and Doyle, J. [1980] Non-Monotonic Logic I. *Artifi cial Intelligence 13*, 41-72.

(15) Moore, R. [1985] Semantical considerations on non-monotonic logic. *Artifi cial Intelligence 25*, pp. 75-94.

(16) Mott, P. [1987] A theorem on the consistency of circumscription. *Artifi cial Intelligence 31*, 87-98.

(17) Perlis, D. [1981] Language, computation, and reality, PhD Thesis, Univ. of Rochester, Rochester NY.

(18) Perlis, D. [1985] Languages with self reference I: foundations. *Artifi cial Intelligence 25*, 301-322.

(19) Perlis, D. [1986] On the consistency of commonsense reasoning. *Computational Intelligence 2*, 180-190.

(20) Perlis, D. [1987] Circumscribing with sets. *Artifi cial Intelligence 31*, 201-211.

(21) Perlis, D. [1988] Languages with self reference II: knowledge, belief, and modality. *Artifi cial Intelligence, 34*, 179-212.

(22) Perlis, D., and Minker, J. [1986] Completeness results for circumscription. *Artifi cial Intelligence 28* 29-42

(23) Reiter, R. [1980] A logic for default reasoning. *Artifi cial Intelligence 13*, 81-132.

(24) Reiter, R. and Criscuolo, G. [1981] On interacting defaults. Proceedings of the Seventh International Joint Conference on Artifi cial Intelligence, pp 270-276.

(25) Shepherdson, J. [1951] Inner models for set theory I. *J Symbolic Logic, 16*, 161-190.

(26) Stalnaker, R. [unpublished] A note on non-monotonic logic. Manuscript, Philosophy Department, Cornell University.