delim $$

To appear in AIJ.

Languages With Self-Reference II: Knowledge, Belief, and Modality

Donald Perlis

Computer Science Department
and
Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland 20742
perlis@mimsy.umd.edu
(301) 454-7931

Negative results of Montague and Thomason have diverted research in propositional attitudes away from syntactic ('first-order') approaches, encouraging modal formalisms instead, especially in representing epistemic notions. We show that modal logics are on no firmer ground than first-order ones when equally endowed with substitutive self-reference. Nonetheless, there may still be remedies, hinging in part upon a distinction between 'dynamic' and 'static' notions of provability and belief (an earlier version of this paper emphasized a somewhat different distinction).

descriptors: knowledge representation, commonsense reasoning, self-reference, modal logic, belief, knowledge, provability, auto-epistemic reasoning, introspection.

O. Introduction

The focal point of this investigation is a result of Montague [23], whose customary interpretation as an argument in favor of modal logics for belief and knowledge as opposed to a classical first-order approach, I challenge. (For other responses to Montague, see [1,4,33].) I shall argue that modal logics are on no firmer ground than first-order logics when equally endowed with substitutive self-reference. Both modal and first-order treatments of knowledge and belief for commonsense reasoning can readily lead to inconsistencies. Yet there still may be remedies, depending on the particular forms of commonsense reasoning (and specifically of auto-epistemic reasoning) considered.[1]

Let us write Bel(x) and Know(x) to indicate that x is believed, respectively known, by an implicit agent g. The syntactic status of x is one of the issues to be addressed. If Bel and Know are predicate symbols, then x is an ordinary first-order term which in particular may be the name of a sentence.[2] On the other hand, if Bel and Know are modal operators, then x will be a well-formed formula. In [26,28] it was suggested that for an intelligent reasoner g, a self-referential language is desirable in order to represent (to g itself) such notions as that g has a false belief. We may write, for instance,

$$(\exists x)(\mathrm{Bel}(x) \,\&\, \neg \mathrm{True}(x)).$$

But if this very wff is a belief of g, then it too can serve (either in quoted first-order form, or in formula -- modal -- form) as an argument within another belief formula. I have contended [28] that this is such a basic aspect of language and thought that any reasonable representational mechanism for commonsense reasoning must include facilities for expression of self-reference and syntactic substitutions. We will see that this has significant consequences regarding consistency and modal treatments, in that apparent advantages of the latter over non-modal ('syntactic') ones disappear in the presence of self-reference.

Now, the theorems of a proposed theory S for beliefs of an agent g can be viewed as themselves being the conclusions held by g, so that S is thought of as g's own reasoning context. Alternatively S can be

---

[1] An earlier version of this paper [30] proposed remedies along somewhat different lines, that I now feel to be of less generality and usefulness.

[2] See [21] for an early call for naming, or reification, in AI.

viewed as the theory of someone, h, other than g, who is reasoning *about* g's conclusions. In the former case, g will be able to reason about g's own conclusions, and in the latter h may reason about g's reasoning about g's own conclusions. Of course, even more complex scenarios are possible, and have been considered in the literature, e.g., [4,8,14,16,19,39]. In any case, such conclusions by g amount to beliefs of g. This will play a key role in our analysis.

A variety of theories has been considered for the study of belief and knowledge, many of them modal. S5, to be presented below, is perhaps the most famous of these. For now, I simply observe that S5 and other similar theories (modal *or* classical first-order) are very limited as theories of epistemic behavior of intelligent agents. In effect, they are view knowledge (or belief) fixed once and for all in a timeless world; there are no processes, no mistakes, no guesses, no decisions, no plans, no goals, no new information. This is understandable: the idea here and elsewhere in formal studies has been to get a simple view right first, before turning to the complexities of real reasoning.

Still, the thrust of these theories has been one of deduction, that is, of additional beliefs an agent should come to, given certain original beliefs. There is in this an underlying, if not explicit, notion of process. While at times such complexities can be safely ignored for the sake of simplicity of analysis, at other times the very analysis can be impeded by such limitations. The more complex problem of real-time ongoing reasoning actually suggests key ideas about the nature of knowledge and belief that cannot easily be seen from the more restricted idealized view. This will emerge in sections VI and VII. While we will not here endorse an explicitly process-oriented formalization of knowledge and belief, the underlying idea of process will serve in our analyses. Key to this is the idea that an agent's set of beliefs may change over time.

Montague and Thomason showed that certain apparently plausible formalizations of the concepts of knowledge and belief have turned out to be internally inconsistent. I will recall some of these results here, and supply still more. Does this mean we must surrender formalism altogether, and look only at heuristic algorithms? I do not think so. Some of our results (Theorems 6 and 7) and an Open Problem point in directions that may be fruitful. What perhaps should be concluded is that timeless approaches may not be a

good way to study knowledge and belief, or at least that timeless theories should be formulated *with an eye to* an underlying temporal framework.

My examination of formalizations of the propositional attitudes of belief and knowledge will lead to a distinction between two rather different kinds of theory I shall call 'static' and 'dynamic.' To offer a brief preview, I call attention to the highly introspective feature present in theories of belief and knowledge. These are, largely, theories of *self*-belief and *self*-knowledge. That is, reasoning is specified so as to allow conclusions like Bel α or Know α from already having concluded α, and so on, in a succession of ''layers.'' This layering however has been collapsed in most formalisms into a single flattened sea of conclusions. In the next section, I indicate in intuitive terms why this is suspect in general. Later sections give technical difficulties and potential solutions still within a flattened context. Roughly, dynamic theories, although flattened, are devised to take account of the (implicit) layers in a way congenial to commonsense reasoning, while static theories are not.

To set the stage for the rest of this paper, I briefly comment on the nature of knowledge and how it contrasts with that of belief. Indeed, the differences seem striking. Whereas knowledge is firmly tied to the notion of truth (as evidenced in the first order schema Know('α') $\rightarrow \alpha$), that of belief is another matter entirely. In fact, knowledge has at times been characterized as simply true belief (or alternatively as justified true belief; but see [11]). In any case, there need be no presumption at all in agent h that a sentence believed by agent g is for that reason true. It may be true or false, and h's calling a sentence σ a belief of g distinctly raises the possibility of σ being false. That is not to say, however, that the agent g believing the sentence σ regards σ as anything but true.

Here I am taking the word ''belief'' in the (admittedly imprecise) sense of g's very definitely believing σ to be true, rather than merely suspecting σ to be true (as in ''I believe so''). It may seem advisable to regard the beliefs of an agent g as simply some set or other of wffs. For some purposes this is too lax an approach; see [29] for a suggested narrowing of the definition of belief. However, as with knowledge, much of the literature has tended to treat belief as obeying very stringent ''ideal'' conditions such as logical omniscience (for example, obeying Bel$(\alpha \rightarrow \beta) \rightarrow$ (Bel $\alpha \rightarrow$ Bel β)). It is this approach that I will explore here.

(See [6,7,8,17,19,26,27] for exceptions.) We will find that many such approaches lead to inconsistencies, but eventually find others that are more promising.

The remainder of the paper addresses the following topics: section I presents a sketch of the underlying intuitions motivating my analysis of introspective reasoning, which will crop up at a technical level again toward the end of the paper. Then II briefly reviews the importance of self-reference in commonsense reasoning, in particular with regard to formal substitution. In III general troubles with paradoxes of substitutive self-reference are reviewed, and modified consistent substitution-assertion rules are given based on [10,28]. Then IV reviews modal logics and problematic aspects of first-order analogues of modal logics for knowledge and belief in the presence of substitution *a la* Montague and Thomason, and in V we see that modal logics in the presence of substitution are problematic as well.[3] In VI, these problems are studied in terms of formal provability, providing a consistent static theory STAT of belief and knowledge; and in VII a result of Löb as well as an example of Moore lead us to the aforementioned distinction between dynamic and static notions of provability and belief and suggest a dynamic theory DYNA. Then in VIII, I summarize and suggest where more attention may be needed.

I. Flattened layers of introspection

Smith [34, p.40] tells us ''Perfect self-knowledge is obviously impossible... The self can never be viewed in its entirety, because there is no place to stand -- no vantage point from which to look.'' In terms of our earlier remarks, the ''obviously'' here may mean that at best one can look back at ones ''entirety'' of a time prior to the present. There are layers of time to the phenomena of deduction and introspection. Flattening the layers into one is an enticing formal device, but a problematic one, whose only apparent usefulness is greater ease of study. But *certain* features of reasoning may persist invariant over layers and these may then be ''flattenable,'' without the problems[4] that can beset this formal simplification. This is the issue

[3]I caution the reader that this paper focuses principally on *first-order* modal logics rather than propositional ones, since we are interested in studying how self-reference is handled, and for this, quantification is needed, although the form that quantification takes is not necessarily that of first-order logic; see Theorem 5 and associated discussion.

[4]Compare [29] for a related use of problematic self-knowledge.

addressed in this paper, and motivated in the rest of section I.

If g's beliefs do satisfy some condition C, is it not reasonable for g to (come to) believe C itself? Yes and no. It may be that in coming to believe C, g's beliefs change in ways that may invalidate certain conditions, even C! As a simple example, if g has (initially) only two beliefs, A and B, then it is a fact, C, that g has only two beliefs. But if g then comes to believe this fact C about itself, it thereby ceases to have only two for it has now come by a third, C, which is therefore no longer true. On the other hand, if C were instead the fact of having*at least* two beliefs (rather than *only* two), then g's coming to believe C would not render C false.

Moreover, it seems clear that not all beliefs of an agent g *should* remain believed as new statements become believed. An obvious example is that the belief ¬Bel β ("β is not one of my beliefs") should not remain if β comes to be believed. This is an example of what we can loosely regard as a *local perturbation* on g's belief set S. That is, many beliefs are contingent on small details, and can easily change.

Certain other beliefs, however, seem to be of more permanent character, such as

$$\text{Bel}(\alpha \rightarrow \beta) \rightarrow (\text{Bel } \alpha \rightarrow \text{Bel } \beta),$$

which asserts a global feature of g's reasoning processes. Of course, such a belief need not be true, or remain true over time. But it is perhaps plausible that *some* such general beliefs might remain true (for g) over a long period, and that g might come to believe *those*.

Herein lies the risk: will g's coming to believe a general belief γ about its very set of beliefs, alter that set in such a way that γ no longer holds, or so that it contradicts other beliefs of g? As with the A-B-C example, it depends on γ. We will be concerned here to outline some broad categories of γ's. But one thing we do want, is to be able to make local perturbations without thereby being forced to alter whatever global beliefs are held. Coming to believe β should not force giving up, say, $\text{Bel}(\alpha \rightarrow \beta) \rightarrow (\text{Bel } \alpha \rightarrow \text{Bel } \beta)$. Theories that tend to respect this requirement we call *dynamic*; ones that do not leave such global wffs invariant under local perturbations we call *static*.[5]

[5]We will not be able to give necessary and sufficient conditions defining these terms. Nevertheless, they will serve our needs in directing our analysis and reformulation of various theories.

Another approach would be to openly embrace a hierarchy of theories, each looking back into the previous one. This latter suggestion we will not explore further here, although it is one that deserves attention (see [6,9,17,29]). In this paper we will confine attention to single (flattened, ideal) theories of knowledge and belief, to see whether any can be made to avoid the kinds of difficulties mentioned.

II. Self-reference as a principal of language and thought

In [26] the term *syntaxal* was used to describe languages having terms representing the syntactic features of that same language, the idea being that such self-descriptive power is essential for many purposes within natural language and commonsense reasoning. Rieger [32] uses the term *referenceability* for a similar notion, namely, that it is often important to reason (and communicate) about a particular feature of an utterance, *viz.*, ''That you said *howdy* struck me as unusual.'' Here elements of speech themselves are being referenced; for this some device is needed in a formalization. The quotation mechanisms presented in [10,26,28] are a possibility. An alternative device is found in modal logic, in which formulas (if not arbitrary expressions) are allowed to be operands to other expressions. One of our concerns here will be to contrast these approaches.[6]

We will use the expression ''self-referential'' as a gloss for either of these concepts (referenceability and syntaxality), noting that this leaves open whether all syntax is to be available for reference, as opposed, say, to whole formulas alone, or other aspects of language. This ambiguity will serve our needs, however, when we consider the alternatives of modal and first-order languages.

A further feature of natural language, and one that effective self reference appears to hinge upon, is that of substitutivity. By this I mean the ability to refer to the result of making alterations in a statement, such as ''If you had said *John is here* instead of *Mr. Smith is here*, I would have understood who you meant.'' The ability to form and compare such variations on our utterances is so elementary and

[6]Boolos [2] and Smorynski [35] deal with another point of contact between self-reference and modal logic, namely, the use of a modal operator for the provability relation, and the corresponding treatment of self-reference as it derives from Gödel's Incompleteness Theorem [13]. This has connections with the present work, in that provability is one plausible model for a belief predicate as treated here; however, our focus is on surmounting certain paradoxes related to belief and knowledge rather than studying provability *per se.* See section VI below.

fundamental to our use of language, that it is hard to imagine taking seriously any proposed formal language for a mechanical version of natural language processing that does not have a corresponding facility. Indeed, it seems tantamount to the ability to represent the very fact that language involves using symbols (e.g. *John*) to stand for other entities (e.g., John himself), as in ''You used *Mr. Smith* to refer to John.'' Modal logic (as well as first-order logic) is in general broad enough to allow expression of such notions (if desired).

However, individual substitutions are not enough, as already suggested above in contrasting *John* and John. The very concept of substitution should be expressible, as in ''If the subject is made plural, the verb should be also, so that if *people* is substituted for *person* as subject of a sentence σ, then the verb should be changed to its plural form.'' That is, the expression σ that is undergoing internal changes is not specified in detail, for a general rule is being given. This means that variables are needed to refer to expressions in the language. This amounts to little more than the ability to recognize symbolic strings, and so is not a computationally unreasonable condition to place on a language. Now here a modal logic in its usual form may need to be extended to allow variables for this purpose, whereas first-order logic does not require such modification.

Our starting point then is the contention that giving up such substitutions would be an unrealistic simplification of any formal language for commonsense reasoning. This would be analogous to a reasoning system g that behaves as follows: First, g makes assertion A and then is asked why it chose to make that particular assertion, instead, say, of a similar one with ''John'' used instead of ''Mr. Smith''. But g, having no concept of making a particular assertion made up of elements of some language, let alone of altering those elements, simply fails to comprehend what is asked of it. Of course, the design of full-fledged reasoning techniques to deal with such cases may involve many things; I contend however that an adequate treatment of self-referential substitution is one of them. Thus before turning to specific aspects of belief and knowledge, I explore some aspects of substitutive self-reference in first-order and modal logics.

III. Preliminary results

We shall call a theory (over a language L) with mechanisms for expressing *and* asserting substitutions *unqualifiedly substitutive.* The hallmark of an unqualifiedly substitutive language is that it possesses an operator or predicate Sub(P,Q,a,n) directly asserting[7] the result of substituting in an expression P the expression Q for the nth occurrence of the subexpression a. I.e., if P[Q/a,n] is the expression that results from the indicated substitution, then we are requiring Sub(P,Q,a,n) to be provably equivalent to P[Q/a,n]. Note that Sub here is to be an actual symbol (predicate or otherwise) of L, while P[Q/a,n] is a meta-notation denoting some actual expression of L, namely the one resulting from the actual performance of the substitution. Of course, for the above-mentioned equivalence to be meaningful, the substitution must result in a well-formed formula of L.

It turns out that for the applications to be pursued here, a rather special variation on the Sub operator is required, namely one in which the substitution of Q for a in P be performed for precisely all occurrences of a in P except the last. Therefore I will write simply Sub(P,Q,a). Contexts will vary slightly in that sometimes certain occurrences of terms will be quoted.[8]

As will be seen, the *asserting*[9] of the results of substitutions, i.e., relating the referenced syntactic elements to their intended meanings, runs into paradoxes of self-reference. Firstly, a means of unquoting quoted elements is needed, i.e., of saying formally that "α" carries the meaning α.[10] That is, Sub(P,Q,a) can be thought of as consisting of two conceptually distinct aspects: forming the new expression, and asserting it. These we can conveniently distinguish by writing the formula True(sub(P,Q,a)) where sub is a function producing (a name for) the expression that the indicated substitution leads to, and True asserts this expression. Again of course this can be meaningful only if the substitution leads to a wff of L.

---

[7]That is, Sub(...) is to behave intuitively as if it were True(sub(...)) where sub is a function symbol. This will appear in more detail below. The reason I do not simply import True and sub wholesale here is that I have in mind various applications not all of which fit the mold of predicates and arguments (i.e., modal theories will have operators instead of predicates).

[8]This is, again, since both modal and non-modal contexts will arise. I have blurred details of quotes so as not to constantly write two forms for every predicate expression. The intention is always to have substitutions result in wffs of the appropriate sort for the context. I also have written P[Q/a] for the result of substitution in either case.

[9]Here again I mean simply that we are using a formula with either an operator or predicate whose intuitive interpretation is to be that of the truth of its operand or argument.

[10]This is often represented as defining a truth predicate: True('α') is to tell us that the sentence "α" is true , so that True('α') and α should hold in the same models of a suitable theory.

However, this apparently cannot be done in such a direct way, for as Tarski [37] showed, the schema

$$\text{True}(\text{`}\alpha\text{'}) \leftrightarrow \alpha$$

leads, in any reasonably expressive language, to inconsistency.

The Sub concept described above is the key ingredient here. Unless care is taken, it will be possible to use the variables that range over expressions in such a way that they refer to that very symbol, Sub, and then it is often only a short step to paradox. I formulate this as a theorem in the first-order case; later I will present it as well for modal theories. In the present form it can be considered a variation on Russell's paradox as well as on Tarski's result above. Since Theorems 1 and 2 below amount to variants on ideas already present in the literature (in particular [10,26,28,37]), their presentation here will be abbreviated, especially regarding quotation conventions. The reader can skim quickly through these preliminary results without loss.

For precision's sake I offer the following definition:

> **Definition:** Let S be a first-order theory over a language L containing a 3-place predicate symbol Sub together with the axiom schema Sub('P','Q',a) ↔ P['Q'/a] where P['Q'/a] is as previously described, for all wffs P and Q and terms a of the language L (which is assumed to contain a constant 'α' for each wff α of L). Then S is said to be *unqualifiedly substitutive*.
>
> **Theorem 1:** Let S be an unqualifiedly substitutive first-order theory. Then S is inconsistent.
>
> **Proof:** We use R(x) to abbreviate ¬Sub(x,x,y) and then R('R(y)') abbreviates ¬Sub('¬Sub(y,y,y)','¬Sub(y,y,y)',y), which by the schema is equivalent to ¬¬Sub('¬Sub(y,y,y)','¬Sub(y,y,y)',y). Here we are using the special substitution feature mentioned earlier, so that all except the last occurrence of y in ¬Sub(y,y,y) is replaced by '¬Sub(y,y,y)'. Thus we have R('R(y)') is equivalent to ¬R('R(y)'), a contradiction.

An alternative (although less precise) argument is as follows: Define True(x) to be Sub(x,a,a) and apply the schema for Sub. This yields True('α') ↔ α for each α, which as mentioned above was shown in

[37] to be inconsistent under fairly general conditions.

Part of what is required for the above kind of arguments is the mechanical ability to find and erase a symbol and put another in its place. As we saw above, this is fundamental to the expression of everyday (and significant) features of natural language and reasoning. Of course, all this depends on Sub having axioms that give it the intended meaning of actual symbolic substitution and assertion of the result, and so we can conclude that this is not possible without qualification, in a consistent first-order system.

In [10,28] the difficulty of formalizing a truth predicate in first-order languages was circumvented, based on ideas in [12] and [18]. Specifically, it was found that the above problematic schema can be replaced by

$$\text{True}(\text{`}\alpha\text{'}) \leftrightarrow \alpha^*$$

for all sentences α, where α^* is essentially[11] the result of replacing in α all subformulas of the form ¬True(‘...’) by True(‘¬...’). I will refer to this as GK (the Gilmore-Kripke schema).[12] GK is consistent in a broad setting. I am using the notation in [28] here.[13] It turns out that this approach can be applied fairly directly as well to the Sub predicate, and leads us to the following result:

> **Theorem 2:** A (‘‘qualifiedly substitutive’’) first-order theory S formed from extending a consistent[14] theory S´ not involving the symbol Sub, by the addition of the (qualified) schema Sub(‘P’,‘Q’,‘a’) ↔ (P[‘Q’/a])*, where now we define α^* to be the result of replacing ¬Sub(‘P’,...) by Sub(‘¬P’,...) in α, is consistent.
>
> **Proof:** First extend S´ to S´´ by adjoining consistently a function symbol sub and monadic predicate symbol True with axiom schema GK. Then define Sub as follows:

---

[11] there are some qualifications regarding the form of α; see [10,28] for details.

[12] adapted from Gilmore’s work [12] on formalizing set theory, to capture ideas of Kripke [18] regarding truth predicates.

[13] However, I frequently abuse notation in that quotes and even parentheses may be left off. Thus Bel 0=1 .→ 0=1 abbreviates Bel(‘0=1’) → 0=1.

[14] Actually we need a consistent theory with at least one infinite model, but this is a very minor restriction, for any consistent theory can be relatively interpreted in a consistent theory with infinite models.

$$\mathrm{Sub}(x,y,z) \leftrightarrow \mathrm{True}(\mathrm{sub}(x,y,z))$$

It follows that this extension S´ ´ is consistent, and clearly S is a subtheory of S´ ´ .

One thing I wish to investigate here (section V) is the extent to which the same result holds for modal theories. First I turn to a question addressed by Montague [23] concerning first-order analogues of certain modal theories.

IV. Modal theories and first-order analogues

The advantages of first-order logic over modal logic were pointed out by Montague [23] (I will review these later). However, Montague found that the ''obvious'' approach to using first-order logic instead of modal logic can lead to inconsistencies. Here I look again at Montague's results, to see whether the simplicity of his original suggestion for using first-order logic can be preserved somehow, and just why a modal syntax seems to manage what first-order syntax does not.

It will turn out that both answers are forthcoming, namely, we will see why modal syntax (sometimes) avoids contradiction, and we will be led to a better understanding of how to represent propositional attitudes in self-referential contexts (whether modal or first-order).

A modal language is characterized by modalities, i.e., operators that can be applied to formulas to produce new formulas, which however are not definable in terms of the standard propositional connectives. The most familiar examples are the necessity and possibility operators, sometimes written Nec and Poss, where usually one is defined in terms of the other, e.g., Poss α iff $\neg$Nec $\neg\alpha$. Thus Nec α and Poss α are (modal) formulas, where α is any formula in the language in question. Indeed, α may itself contain one or more instances of Nec or Poss or both. Thus a modal logic has an extended notion of formula, in which for any formula α and any modal operator M, Mα is also a formula. In particular one can adjoin new operators (and axioms and rules) to a first-order language, creating thereby first-order modal logics. This is the case of primary interest for us.

It is a well-known result that the standard connectives (e.g., & and $\neg$) are sufficient to define all truth-functional operators in a propositional language, so operators that are not truth-functional are called modal

to distinguish them from those already definable. Note that indeed the intuitive sense of Nec and Poss depends on more than the mere truth or falsity of their applicands. For instance, unless we are fatalists, a fire may have truly occurred in the house across the street yesterday, without it thereby being *necessary* that such a fire occurred there yesterday. On the other hand, we may feel inclined to say it was necessarily true that when the temperature of the house reached 451 degrees Fahrenheit, the books began to burn. In both cases, we may suppose the statement which is claimed to be necessary or not, is a true one. But in one case it merely *happens* to be true, and in the other it apparently *follows* from a general law about the ignition point of paper. This is not to say that this is a perfectly unambiguous distinction; nonetheless it serves to illustrate operators that cannot be treated as mere shorthands for expressions built of the propositional symbols. This has been taken as evidence that classical propositional logic therefore is inadequate to the task of representing non-truth-functional operators, and that modal logic should be introduced when such operators are needed.

In the hands of Hintikka [15] and Montague [see 25], modal logics for representing concepts such as knowledge and belief have become powerful tools, and consequently a modal extension of first-order logic is regarded as a standard and natural representational medium for dealing with such matters. Thus once again the suggestion appears that extensions to standard logics are needed to represent appropriately the concepts of natural language, especially of belief and knowledge.

This unfortunately is not without its disadvantages. For one thing, first-order logic is much better understood than any modal logic formalisms, and consequently easier to apply coherently. Secondly, if some reins are not placed on the proliferation of new logics except when the latter are shown to be genuinely different (if not also useful!) from first-order logic, then we will end up with a tower of babel, and research will probably suffer.

However, other avenues are open within first-order logic. One that appears promising is to use, instead of a formula as such, rather a quoted formula or term, so that the intended operator applies to such terms instead of formulas. That is, the operator becomes a predicate symbol: Nec('α') or Poss('α'). Then, so the hope goes, the corresponding axioms can be formulated satisfactorily without going beyond first-

order logic.

This allows us to state a third technical benefit that would accrue from a first-order approach to propositional attitudes; in particular, in the words of Montague [23], ''if modal terms [i.e., modal operators] become predicates, they will no longer give rise to non-extensional contexts, and the customary laws of predicate calculus may be employed.'' For instance, if in fact Bill and Kathy have the same phone number, a modal wff such as

Bel(John, phone(Bill)=277-1265)

when coupled with Leibniz' Rule of Substitutivity (that equal terms may be substituted for one another without disturbing logical equivalence), yields

Bel(John, phone(Bill)=phone(Kathy))

even though John may *not* know this. Thus modal treatments combined with normal substitution practice is problematic, and special conventions are required to keep the unwanted consequences at bay. This suggests the attractiveness of remaining within a first-order language.

This is not to say that no problems remain in a first-order setting, of course. However, a first-order approach would instead involve the wff

Bel(John,''phone(Bill)=277-1265'')

which no longer has phone(Bill) as a term; rather the entire second argument to Bel is one constant term. Other similar difficulties that arise in substitution in modal contexts (sometimes referred to as opacity versus transparency of the modal operator in question), when treated instead via arguments that are quoted formulas, do not occur in first-order logic. The abandonment of first-order logic then is not to be taken lightly.

Motivated by these concerns, Montague [23] applied this approach to a modality for necessity. That is, writing Nec('α') instead of Nec α he obtained a quotational first-order construction. Montague proposed axioms for such a formulation, in analogy with standard axioms in the corresponding modal treatments. Unfortunately he found these versions to be inconsistent, whereas each corresponding modal operator version M is consistent. This seemed to be strong evidence in favor of the modal treatment. However, it appears that the inconsistency Montague uncovered hinges on certain fundamental expressive strengths of

quotational first-order languages which are lacking in usual propositional modal languages. That is, first-order logics have richer sets of formulas than have traditional modal logics. Variables allow the formation of (self-referential) wffs that otherwise would not appear in the language, and thus more is being asserted in first-order logic than in the corresponding modal logic. The question then arises: if a modal theory M is made self-referential (i.e., endowed with expression and assertion of substitutions), is it still consistent?[15]

One particular modal theory of interest is S5. Its language is that of propositional logic together with a modal operator. It was first studied as a formalization of the intuitive notion of necessity, with modal operator Nec[16], but also serves as a (tentative) formalization of the notion of knowledge. When I have knowledge in mind, I use Know instead of Nec, and when I have belief in mind I use Bel. In the immediate sequel I will employ Know. Note that α is a formula in a language containing Know, so that arbitrary nestings of Know are permitted.

S5 has the following axiom schemata:[17]

K: $\text{Know}(\alpha \rightarrow \beta) \rightarrow (\text{Know}\ \alpha \rightarrow \text{Know}\ \beta)$

T: $\text{Know}\ \alpha .\rightarrow \alpha$

I: $\neg\text{Know}\ \alpha \rightarrow \text{Know}\neg\text{Know}\ \alpha$

as well as all (substitution instances of) tautologies, and the following rules of inference:

---

[15]It is of separate interest whether a first-order version of a modal logic can be kept suitably "weak" so as not to intrude, via its variables, new kinds of wffs that destroy a faithful match with the modal logic. This has been explored by [des Rivieres & Levesque 33]. Our purposes here are somewhat different, namely, how to represent propositional attitudes in an explicitly self-referential context. Our contention is that apart from a desire to avoid inconsistency, there should be an underlying intuitive model justifying ones axioms. Then presumably whatever underlying intuitive model justifies the use of any particular modal formulation should apply as well to the full first-order formulation, unless that model itself indicates a principled argument to the contrary.

[16]often written as L or a box $\Box$.

[17]In the literature, schemata K, T, and I are sometimes called Distribution, Knowledge, and Negative Introspection, respectively, and rule N is called Necessitation and sometimes written as RN. I is also sometimes called schema 5, since it is the distinguishing schema of S5. If I is dropped, the resulting system is called T (not to be confused with schema T, although schema T is the characteristic schema of the system or theory T). If schema T is dropped as well, the resulting theory is called K (with then characteristic schema K). Theory S4 has schemata K and T and the "Positive Introspection" or "PosInt" schema $\text{Know}\alpha \rightarrow \text{KnowKnow}\alpha$ as well as rule N; so S4 is stronger than T; it turns out that PosInt is provable in S5 so that S5 is stronger than S4. Thus K, T, S4, and S5 are in increasing order of strength, and all have rule N and schema K; any such system (at least as strong as system K) is called (essentially) a *normal* system of modal logic. Another system, G, is studied in [2]; note that there 'normal' is used as here except that rule N is not required. We refer the reader to [5] for more information on the (enormous) variety of modal systems studied. For a recent modal logic specifically designed for AI, see [20].

MP: from α and $\alpha \rightarrow \beta$ infer β

N: from α infer Know α

S5 as presented above is a propositional theory, although it also has been studied in a predicate context, with an underlying first-order language supplemented with the operator Know, and with the usual logical axioms and rules of inference as well as the axiom schemata K, T, and I, and rule N. It is in the first-order context that we are interested, so that in the remainder of this paper S5 will be taken to mean first-order S5.[18] Note that not only is the language extended beyond classical propositional or first-order languages, but also there is a non-classical rule of inference. This requires comment. Rule N is to be applied strictly to actual theorems of S5, not to any hypothesis α we may wish to employ in proving, say, $\alpha \rightarrow \beta$. That is, although *if* α were a theorem of S5 then so would be Know α, this does *not* entail that the wff $\alpha \rightarrow$ Know α is a theorem of S5. In particular, extensions to S5 formed by adjoining new axioms are not assumed to obey rule N for any wffs other than the original ones provable in S5 itself, unless otherwise stated.

As a theory of knowledge, S5 has the following intuitive interpretation: Know α means α is known (by some thinking agent). Then K, T, and I may be plausible for an ''ideal thinker'' g. Schema K says that g's knowledge is closed under modus ponens; schema T that whatever g knows is true; and schema I that g knows whenever it doesn't know something (i.e., it can introspect negatively). Also, rule N is plausible if the agent is smart enough to know everything that can be established in S5, which may seem easy to grant for an ideal thinker. This presupposes that we are viewing S5 as *our* ''external'' theory *about* g's ''internal'' knowledge. However, rule N then has the further consequence that g will end up taking as its own knowledge all the axioms and theorems of S5 so that S5 ends up also being g's own theory after all. In this light, rule N serves as a kind of positive introspection mechanism.[19]

---

[18] See [2,5] for more detail on S5 and its uses.

[19] The arguments to Know in S5 are interpreted as propositions rather than (quoted) sentences as in a syntactic first-order approach. This has some benefits, such as corresponding to Pierre's and Peter's knowing the ''same'' proposition that *Londres est jolie* and *London is pretty*, respectively. Also there are technically elegant semantics (due to Kripke; see [5]) that can be provided for S5 and other modal systems. However, this does not alter the fact that Pierre and Peter *express* their propositions sententially, nor that actual sentences are very important features of language. Thus our comments about the central role of self-referentiality and syntax seem to hold up. There are other qualifications regarding propositions as objects of knowledge or belief as well; for an overview of much of the philosophic literature, see [36].

Montague studied several systems related to S5, with the particular aim of changing Know into a predicate symbol applied to names of formulas. I need not present details of these modal variants in order to state the following result of his:

> **Theorem** [Montague 23 (Thm 3)]: Any first-order ''arithmetical'' theory having the schema T´ , namely, Know('α') $\rightarrow \alpha$ for each closed wff α, and also satisfying condition N´ , that |-- Know('α') whenever |-- α, is inconsistent.

Note that schemata I and K have been left out. I will henceforth however drop the '´ ' on T and N, letting context determine whether a modal or first-order schema is meant. Also, I will retain the names T and N even when the predicate symbol is other than Know, e.g., Bel or the usual provability predicate Thm described later. The term ''arithmetical'' need not concern us; it is a gloss for a technical requirement that has the effect of allowing asserted substitutions into wffs.[20] We can establish an alternative theorem with a quick proof, if we stipulate a *sub* function directly, to get a variation on Montague's result that is more tailored to our needs. I begin with a definition.

> **Definition:** If S is a first-order theory with function symbol *sub* of three arguments, and supplied with a distinct function term 'α' for each expression α (such that the free variables of 'α' are those of α) as well as axioms sub('P','Q','a') = 'P['Q'/'a']', i.e., the name of the result of the indicated substitution, then S is *first-order self-referential.*

> **Theorem 3:** Let S be a first-order self-referential theory having a monadic predicate symbol Know and axioms Know('α') $\rightarrow \alpha$ for each closed wff α, and satisfying the condition |-- Know('α') whenever |-- α. Then S is inconsistent.

> **Proof:** Much as in the proof of Theorem 1, let R(x) abbreviate the formula ¬Know(sub(x,x,'y')), so that R('R('y')') abbreviates

[20]Moreover, the use of substitution is virtually tantamount to the introduction of a certain amount of arithmetic in any case (see Quine [31]), and I have argued that substitution is an essential feature of commonsense reasoning.

¬Know(sub(‘¬Know(sub(‘y’,‘y’,‘y’))’,‘¬Know(sub(‘y’,‘y’,‘y’))’,‘y’)),

which is equivalent to

¬Know(‘¬Know(sub(‘¬Know(sub(‘y’,‘y’,‘y’))’,‘¬Know(sub(‘y’,‘y’,‘y’))’,‘y’))’).

So R(‘R(‘y’)’) is equivalent to ¬Know(‘R(‘R(‘y’)’)’). We further abbreviate R(‘R(‘y’)’) as RR, so that we have RR iff ¬Know(‘RR’). If we then use the axiom Know(‘RR’) → RR, we get Know(‘RR’) → ¬Know(‘RR’), so that Know(‘RR’) is impossible and therefore ¬Know(‘RR’) is proved. But this is (equivalent to) RR, so RR is proved. But then by the postulated inference condition, we deduce Know(‘RR’) after all, a contradiction.

What does this result tell us? It appears that even a very weak subtheory of S5,[21] when ‘‘translated’’ into a first-order context, goes awry, at least in the presence of substitutivity. But is this reason to think that the modal version is better off? It is true that S5 (and therefore its subtheories) are consistent. But S5 by itself is not in a substitutive context. So the question arises as to whether modal theories such as S5 remain consistent when augmented with substitution capabilities.

In a similar vein, Thomason [38] has provided another apparent failure of our intuition, as follows: If an agent g believes (a suitable theory of) arithmetic and also g’s beliefs (given as arguments to the predicate Bel) satisfy the following conditions:

$$\text{Bel}(\text{`}\alpha\text{'}) \rightarrow \text{Bel}(\text{`Bel}(\text{`}\alpha\text{'})\text{'})$$

$$\text{Bel}(\text{`Bel}(\text{`}\alpha\text{'}) \rightarrow \alpha\text{'})$$

$$\text{Bel}(\text{'}\alpha\text{'}) \text{ for all valid } \alpha$$

$$\text{Bel}(\text{`}\alpha \rightarrow \beta\text{'}) \rightarrow (\text{Bel}(\text{`}\alpha\text{'}) \rightarrow \text{Bel}(\text{`}\beta\text{'}))$$

then g is inconsistent in the sense that g will believe all wffs.

To relate this to the intuitions of section I, I offer the following critique for Thomason’s and Montague’s theories as theories of omniscient ideal reasoning: In each case, an axiom schema (either T or the second schema above) attributes to g a global belief about g’s own beliefs. This self-viewing is best taken as

[21] In fact all we need are schema T and rule N along with first-order logic and self-reference, so that not even the full system T is essential here, since schema K is not used.

layered or step-like, and when instead it is flattened in one fell swoop then internal contradictions can arise. In effect, when g takes a position regarding the contours of its set S of beliefs, it may be embracing a ''new'' belief β which, if we force $\beta \in S$, can run into a self-referential dilemma. If g is aware of this, it might prudently choose to be more circumspect about its use of self-reference and in the process better merit our designation of it as ''omniscient'' or ''ideal''.

V. Substitutive modal logic

If we endow a modal logic M with the property of substitutivity in the form of an operator Sub(P,Q,a), with the intention that this thereby create suitable conditions for referenceability within such an extended version of M, we have at least two available approaches. We can let P and Q be quoted expressions and Sub a predicate symbol, or we can let P and Q be formulas and Sub another modality.

Let us begin by exploring the first alternative. Since we already know that a first-order unqualified substitutive theory is inconsistent (Theorem 1), then so will be any modal theory M that extends such a first-order theory. Therefore, if we endow S5 with a predicate symbol Sub, we can't allow it the unqualified substitution axioms as well. What then if we use only qualified substitution axioms of the sort known to be consistent in the first-order case? That is, can we extend S5 to include Sub(x,y,z) $\leftrightarrow$ True(sub(x,y,z)) together with the consistent treatment of True and sub mentioned earlier, and thereby retain consistency in the modal theory that results? Unfortunately, the following result shows that we cannot.

> **Theorem 4:** If M consists of S5[22] extended by the Sub predicate with axiom Sub(x,y,z) $\leftrightarrow$ True(sub(x,y,z)) and associated qualified axioms for True and sub, then M is inconsistent.
>
> **Proof:** We proceed by defining R(x) to be the formula
>
> ¬Know Sub(x,x,'y')
>
> Then R('R(y)') -- which we will abbreviate as RR -- is
>
> ¬Know Sub('R(y)','R(y)','y')

[22] An anonymous referee has pointed out that the proof does not use schema I, hence the theorem actually holds if S5 is replaced by any modal system with schemata T and K and rule N, i.e., for any logic as strong as modal system T.

i.e.,

$$\neg \text{Know True}(\text{sub}(\text{`R(y)'},\text{`R(y)'},\text{`y'}))$$

Now sub(‘R(y)’,R(‘y’),‘y’)) = ‘RR’ and so

$$\text{True}(\text{sub}(\text{`R(y)'},\text{`R(y)'},\text{`y'})) \leftrightarrow \text{True}(\text{`RR'})$$

whose right-hand-side is equivalent to RR since True simply strips off quotes from its argument except when the symbol True itself is directly negated in that argument. What we have then is

$$\text{True}(\text{sub}(\text{`R(y)'},\text{`R(y)'},\text{`y'})) \leftrightarrow \text{RR}$$

and from rule N and schema K we then get

$$\text{Know}\,[\text{True}(\text{sub}(\text{`R(y)'},\text{`R(y)'},\text{`y'})) \leftrightarrow \text{RR}]$$

and then

$$[\text{Know True}(\text{sub}(\text{`R(y)'},\text{`R(y)'},\text{`y'}))] \leftrightarrow [\text{Know RR}]$$

It follows that

$$[\neg\text{Know True}(\text{sub}(\text{`R(y)'},\text{`R(y)'},\text{`y'}))] \leftrightarrow [\neg\text{Know RR}]$$

But RR is equivalent to ¬Know True(sub(‘R(y)’,‘R(y)’,‘y’)), and so we get that

$$\text{RR} \leftrightarrow \neg\text{Know RR}$$

Now, Know RR → RR by schema T, and so Know RR → ¬Know RR, which means that ¬Know RR is a theorem of M. But we have just seen that RR is equivalent to ¬Know RR, so then RR also is a theorem of M, and by rule N so is Know RR, a contradiction.

We then consider the second alternative mentioned above, namely, that Sub(P,Q,a) be a modality in which P and Q are formulas. Here we are faced with a difficulty of syntax, if we wish to keep to our underlying premise in this paper, namely, that not only should language be substitutive and assertional, but that the very feature of substitutions should be expressible, in the form Sub(x,y,z) where x, y, and z are variables. This becomes problematic when x and y are intuitively to range over formulas (rather than names of formulas). What is called for is a quasi-second-order modal logic, in which the arguments to which modalities are applied (namely predicates, or more generally, relations) can be the values of variables. Thus we wish to write Sub(X,Y,z) where X and Y are predicate variables, and z is an individual variable that ranges

over names of expressions. However, it turns out that it is not necessary to adopt such a syntax, for even without variable arguments to modalities, contradiction arises.

Definition: S is an *unqualifiedly substitutive* modal logic if S has a modality Sub(P,Q,A) and the (by now familiar) substitution axioms using P[Q/A], where P, Q and A are wffs. That is, Sub(P,Q,A) is equivalent to the result of substituting Q for all but the last occurrence of A in P. (We need not even use names at all, for instead of arbitrary expressions, it suffices to refer to whole formulas.)

Theorem 5: Any unqualifiedly substitutive modal theory is inconsistent.

Proof: We simply pick an arbitrary wff, say P, and consider the formula (*) below:

(*) Sub(¬Sub(P,P,P) , ¬Sub(P,P,P) , P)

The indicated substitution will then replace the first two P's in the first argument ¬Sub(P,P,P) of (*), with the second argument, which also is ¬Sub(P,P,P). This results in

(**) ¬Sub(¬Sub(P,P,P) , ¬Sub(P,P,P) , P)

which is in fact simply ¬(*)!! That is, (*) is equivalent to ¬(*), which is a contradiction.[23]

So S5 and even weaker systems such as T are inconsistent with either form of self-reference that naturally arises. Thus any advantage in a modal language seems to be lost, and we might as well remain with a classical first-order language. Still, this does not settle problems of formal representation of belief and knowledge. Whether formulated in terms of classical first-order substitutive or modal substitutive languages, special axioms and rules of inference for propositional attitudes are problematic. We now turn to remedies of this situation, hinging on separating the two mutually troublesome features, namely the schema

[23]It is of some interest that the requirement for ''variable substitution'' has been obviated by the very means of substitution. That is, in some sense the significance of variables is precisely that they allow for the possibility of substitution. This is not so apparent in first-order logic, where variables are central to the entire structure; but in modal logic where seeming arguments (to modalities) are not usually treated in argument fashion, the presence of substitutions is evidently just what brings things back to the first-order fold (at least in regard to self-referential paradox).

T -- Know(‘α’)$\rightarrow\alpha$ -- and the rule N for inferring Know(‘α’) from α.

VI. Belief as provability

Our results indicate that modal logics, when endowed with sufficient power to represent substitutions, face the same inconsistencies found in first-order treatments. Much of the appeal of these logics is then lost, since one might as well then simply stay within first-order logic and employ a stratagem there for retaining consistency, instead of hunting for an analogous stratagem in modal logic. Indeed, we saw from Montague's and Thomason's theorems and from our Theorems 4 and 5, that there are severe difficulties whether the formal syntax is dressed in first-order or modal clothing. In effect, if we are going to have substitutivity of even a very mild sort we have to choose something less than full use of both the notion Know $\alpha .\rightarrow \alpha$ (schema T) on the one hand, and on the other hand |-- Know α whenever |-- α (the familiar rule N).[24] What principled justification can be given for this, and what principled decision can be made toward a resolution?[25] One point of view emerges from the idea of provability.

Consider an agent g, whose conclusions are to be represented. Here I use the term ‘‘conclusions’’ as a deliberately neutral ground between knowledge and belief: whatever is in the agent's reasoning to be used as if trustworthy. While this is still rather vague, it is sufficient for our purposes.[26] What we can say is that an agent's conclusions would seem to be tightly related to what the agent can prove (establish, decide, conclude), so that a natural idea is to explore ideas of provability in an effort to characterize possibilities and limitations on formal treatments of belief and knowledge. This idea was the basis for much of Konolige's efforts in [16], in which however provability was represented as a concept external to a given reasoner, i.e., one agent might reason about the provability relation of another agent but not about its own provability relation. In [17] Konolige comes closer to self-provability, but retains a kind of hierarchical approach in

---

[24]I henceforth routinely drop quotes. All theories will be first-order (non-modal).

[25]Asher and Kamp [1] pursue much the same question, and with a similar course by applying methods for truth predicates to a knowledge predicate. Their work involves a hierarchy of decisions about knowledge, and remains fairly close to Kripke's original idea [18] in that it is model-theoretically oriented. However, the needs of artificial intelligence (and possibly even of logic) seem better served by a more syntactic and proof-theoretic (i.e., computational) approach, as argued in [10,28]. I will comment again on their approach below.

[26]I refer the reader to [29] for discussion of the issue of what constitutes a belief.

which what is provable at one level is then recorded as provable in the next. Here I am more concerned with a ''flattened'' theory that involves a predicate for its very own provability notion, yet in equal standing with the other predicates in the language; that predicate itself then forms part of the grist for those very proofs to which it is intended to refer. [2] and [35] examine formal notions of provability, but in relation to arithmetic rather than epistemic concerns. In this and the next section, I study the extent to which various notions of provability are applicable to belief and knowledge. Initially I focus on the classical static theories; in section VII, I turn to commonsense constraints.

Let us pursue the idea that an agent g's beliefs[27] are its theorems. But then if g is to have a Bel predicate of its own, it may be a kind of provability predicate. What do we mean, formally, that Bel be a provability predicate? Several things may suggest themselves. One, given in [3], follows.

> **Definition:** A *provability predicate* for a theory S is a wff P(x) that satisfies rule N -- if α is a theorem of S, then so is P('α') -- and also schemata K and S4's PosInt:
>
> $$\text{K: } P(\alpha \rightarrow \beta) \rightarrow (P\ \alpha \rightarrow P\ \beta)$$
>
> $$\text{PosInt: } P\ \alpha \rightarrow PP\ \alpha.$$

So, what is available for a provability predicate? There are many possibilities, most of which have little to do with provability. However, one provability predicate has played a special role in logic; it is due to Gödel, and we write it as Thm.[28] In this section I will focus on Thm. In the next section we will find that for certain kinds of commonsense beliefs, this will not do, and we must examine alternative ''introspection'' predicates that are not provability predicates at all (as we have defined them).

Thm stands for the usual Gödel predicate symbol for provability in a ''suitable'' theory of arithmetic S, i.e., Thm is defined as

$$\text{Thm}(\text{`}\alpha\text{'}) \leftrightarrow (\exists x)(\text{Proof}(x, \text{`}\alpha\text{'})$$

where Proof(x,α) in turn formalizes in terms of arithmetic the proof-theory of S: it says x is (the gödel

---

[27] We start here with belief, letting knowledge come in later. It turns out that knowledge is quite a tricky notion; see [11].

[28] Also often written as Prov or Bew (for Beweis).

number of) a proof of (the wff with gödel number) 'α'. Thm then pins down the mechanical details of what goes into a proof. This makes it static, for *no* new beliefs (axioms) can be adjoined now, without undoing the intended meaning of Thm. Thm lays out explicitly all the steps allowed in a proof, and even says that these are the only ones allowed. This explicit sense of Thm roughly corresponds to the specification of a mechanical listing of all and only the wffs that can be established by g (a recursively enumerable, but not recursive, set of wffs).

But how much of Thm is needed, anyway, in actual use of a belief predicate? A reasoner need not (and cannot) know *all* about itself, but might well benefit from knowing *certain* things about itself.

In fact, Thm gives so much detail about proofs within a theory S that it inflexibly binds S away from any new knowledge. Thus if S is extended to S´ , the syntactic definition of Thm, if it is to now express proofs in S´ , must change. Moreover, certain facts will be new simply because they are not provable within S, and thus Thm cannot ever express them with reference to the theory for which it is formulated. In fact, the same happens with any provability predicate, as we shall see, regarding auto-epistemic knowledge of certain sorts.

Now, Montague's result shows that we must give up schema T, if we are to retain rule N and substitutivity (and consistency). That is, not *all* wffs of the form Bel $\alpha .\rightarrow \alpha$ will be theorems of our agent g's theory S. What does this mean in terms of Thm? Intuitively, Thm $\alpha \rightarrow \alpha$ means that each theorem α of S is true (in a fixed standard model). Now, if S really has such a model, then each of these statements Thm $\alpha .\rightarrow \alpha$ is correct; that these cannot all be *provable* in S is an interesting limitation of a computational mechanism to fully express its own computational behavior. This is closely allied with Gödel's second incompleteness theorem and also Löb's theorem, discussed later. However, the underlying idea of the limitation has an intuitive sense to it, and will be the basis for the general picture that will emerge. The idea is that of section I, that the actual processes of a reasoner g can never be known in full detail by g itself, except by g's gaining this as new information which in turn changes g's structure so that what g has gained is a faithful picture of what it *was*, not what it *is*.

As I have stated, Know α is often taken to mean α is among those beliefs of g that are true. Then Know α means *to g* that α is one of its true beliefs, even though in general g cannot identify which these are! Indeed, each of g's beliefs is individually believed (by definition) by g; as soon as any one is seen to be false, it is no longer believed.[29] So g cannot isolate its true beliefs from the rest; it simply can refer to them in the abstract, just as it can refer to its entire belief set. In effect, g may believe that (the extension of) Know is a proper subset of (the extension of) Bel, but can give no examples of the relative complement (i.e., a false belief)!

Nevertheless, using Know as true belief, we can now employ schema T so that it applies to Know rather than to Bel. This manages to get around some of the difficulties we have seen. Specifically, the following result provides one way to endow g with its own knowledge and belief predicates and yet avoid inconsistency. This is a static approach, so that g will not be able to accommodate new beliefs and yet retain the intended meanings of Bel and Know.

> **Theorem 6:** Let S be any consistent qualifiedly substitutive first-order theory, not containing the symbol Bel. Then there is a consistent first-order theory STAT(S), which is an extension of S having predicate symbols True, Bel, and Know, and obeying rule N for Bel, with axiom Know $\alpha \leftrightarrow$ Belα & Trueα, where True satisfies schema GK.[30]
>
> **Proof:** Let True be as in GK, and then let Bel be Thm and let Know be as stated, both as extensions by definitions. Rule N is automatic for Bel (inherited from Thm).

How much of S5 does this result give us? We get rule N for Bel; schema T for True and Know; schema K for Bel, True, and Know. Theorem 6 does not give schema I for any of True, Know, or Bel.

---

[29]Try to imagine a reasoner g having simultaneous beliefs Bel α *and* ¬True α, as in "I believe 1337 is prime, but it is not!"

[30]In [30] a GK version of rule N was used for Know: α** | Know α, where ** is the Gilmore operator applied to Know rather than to True. This has the unfortunate consequence that even straightforward (non-paradoxical) instances of theorems of g were not provably Known by g. For instance, even many *harmless* theorems such as β v ¬β did not have corresponding theorems Know[β v ¬β]. Thus we have dropped this approach and retained the Gilmore technique for the predicate True alone. Similarly, GK is not very satisfactory as a rule for Bel , for it severely undermines the introspection properties. For instance, if ¬Bel β is a theorem, so should be Bel¬Bel β; but GK will not provide this. Also, as for Know, theorems of the form β v ¬β should have counterpart theorems Bel[β v ¬β], but again GK will not provide this in general.

However, we do get S4's positive introspection schema PosInt for Bel, since Thm happens to obey Thm α $\rightarrow$ ThmThm α.[31]

Example: Belief biconditionals and cats.

A key technique implicit in Montague's and Thomason's results, as well as Tarski's and Gödel's, is the Fixed Point or Diagonalization Lemma (e.g., see [22]). This allows us to find, given a predicate P(x), a wff α such that $\alpha \leftrightarrow \neg$P('$\alpha$') is provable in a suitable theory S. Substitutivity (qualified or not) is one criterion that makes S suitable (see [31]). The principal application for us is the following: Let Bel be a monadic predicate symbol of a theory S. We may then let BB be such that S has the theorem BB $\leftrightarrow \neg$Bel BB (which I refer to as the belief biconditional).

Let C be (a formalization of) the sentence ''The cat is on the mat,'' e.g., On(cat,mat). That is, C is a plain ordinary wff without self-reference, and without use of the predicate symbols Bel or Know or True or Thm. Its truth then should be determinate, even if unknown. Thus g should be safe in concluding not only C v ¬C and Know C v ¬Know C, but also Know(C v ¬C) and Know(Know C v ¬Know C). This follows from Theorem 6, and forms an intersting contrast with the wff BB. Let S satisfy Theorem 6. Then

STAT(S) |-- Know(C v ¬C)

STAT(S) |-- Know(Know C v ¬Know C)

STAT(S) |-/- Know(BB v ¬BB)

STAT(S) |-/- Know(Know BB v ¬Know BB)

STAT(S) |-- Bel(α v ¬α) for all wffs α.

---

[31]It is of interest that Asher and Kamp [1] similarly arrive at a (semantical) framework for Bel, in which schema K and PosInt are preserved but schemata I and T are not. However, as mentioned earlier, their treatment has no corresponding proof theory for direct comparison to our work.

Thus tautologies seem to behave well with respect to Bel, and *ordinary* wffs such as C also do so with respect to Know. But some explicitly self-referential wffs like BB fail to do so. In short, STAT(S) seems to represent a somewhat reasonable static theory of belief and knowledge. But it does not satisfactorily answer our doubts about schema T for Bel. For this, we need to look more closely at what use beliefs are put to, in order to assess what role schema T might or might not be reasonably expected to play.

VII. Belief as introspection

Theorem 6 above provides a version of Bel (and Know, etc) that will suffice for certain purposes, such as nesting of belief sentences. As such it may be fine. But more might be desired. Let us see what this might be, by returning to the issue of schema T. Now, we know we may not have the full schema T (for Bel) in g's theory S, if S is to obey rule N (for Bel). And perhaps not all instances of schema T (for Bel) are even *plausible* for g to conclude. But perhaps *certain* instances are both plausible and possible. So first we should ask whether commonsense reasoning has need for these.

Another desideratum arises from rule N itself, for this rule does not necessary apply to new axioms that may be adjoined to S, even though the idea of rule N as an introspection facility should require this. We would like to have a formulation of Bel that is dynamic, in that as g gains new beliefs, the broad characteristics of Bel do not change, and thus that the rules for Bel should also not change simply because g has some additional beliefs. That is, there should be a core theory of belief that is invariant under mere accretion of (at least some[32]) information. If rule N is to be in this core, then it must be applicable to new (local) beliefs that arise.

We lead into this by observing that a different approach to introspection than the 'static' one of Thm can be conceived. In particular, as Löb's theorem will show, certain wffs when adjoined to a theory S result in the meaning of Thm becoming out of date, in the sense that Thm will express provability in the original but not the extended theory. This is because Thm is so pinned down by its arithmetical definition to exact procedures of proof, that it cannot refer abstractly or generically to the general concept of proof. It is

[32]that is, what we loosely called "local perturbations" in section I.

statically tied to one and only one set of conclusions. On the other hand, a reasoner g may not even be aware of its precise algorithm for reasoning, and yet refer to it generically in ways that do not depend on details (and therefore may remain consistent with a greater variety of extensions).[33] That is, perhaps a kind of referring to ones theorems or beliefs can be made, without explicitly stating in detail just how they arise. Possibly then Thm is then too fi ne-grained and unrealistic for commonsense reasoning, both conceptually (who could ever know all their mental processes?) and formally (Löb's theorem below).

Where does introspection arise in commonsense reasoning? One prominent place is in non-monotonic reasoning, in which account is taken of *not* having (believing, proving) a certain wff. For instance, one can believe *a la* Moore [24] that one always knows (or believes) that one has an elder brother if this is in fact the case, without necessarily knowing a way in which that conclusion or belief (that one had an elder brother) would be arrived at.[34] This is then *generic* or *abstract* information about ones set of conclusions. Just how such a notion might be approached and how it should differ from Gödel's explicit Thm will be taken up below. The point however is that an intelligent agent g may not need to know in any great detail just how its mental feats are accomplished; certainly human beings are in this situation. In fact, we will see formal requirements virtually forcing us into this position when we try to give auto-epistemic weight to Bel.

Now, can we think of a situation in which g ought to believe an instance of schema T, i.e., of Bel α $.\rightarrow \alpha$? Well, there are trivial cases, the ones in which α is already a theorem. These of course are instances which STAT(S) will also produce. So how about non-trivial ones? Yes, following (or rather reversing) the example of Moore. We can suppose g to believe ''if I believe I have an elder brother, then I have an elder brother'' even if g does not believe ''I have an elder brother.'' This would be a kind of infallibility belief for g, but a special one regarding brothers, and it seems perfectly plausible that an agent might have good grounds for such a belief as this.

---

[33]This raises a number of issues, only some of which will be dealt with in the remainder of this paper. One I will not touch, is that of hierarchical 'cycles' of reasoning over time, as the agent realizes more and more about its (changing) inference algorithm(s). However, this would appear to be a key one for future work. See [9] for very interesting results on hierarchies of theories of arithmetic, and [6] for an approach to commonsense reasoning viewed as a step-like process.

[34]Hence, from the *not knowing* (of an elder brother), one concludes the *not having*.

Moore's (unreversed) example itself is also of interest; it has the form: $\alpha \rightarrow$ Bel α. Although this is not an instance of schema T, it is also not obvious that it is consistent with STAT(S) as long as Bel remains a provability predicate. Of course, here too there are trivial (or vacuous) cases, when $\neg\alpha$ is a theorem of S.

> **Definition:** A wff α is *simple auto-epistemic over theory S* if α is of the form Bel $\beta \rightarrow \beta$, or $\beta \rightarrow$ Bel β, or $\neg$Bel β, where β is in the language of S. (S may or may not contain the symbol Bel.) I name the three types: MAE is the set of wffs of the Moorean form $\beta \rightarrow$ Bel β, RAE of the reverse form Bel $\beta \rightarrow \beta$, and NAE of the negative form $\neg$Bel β. Note that MAE wffs include instances of PosInt, and all RAE wffs are instances of schema T. I will sometimes abbreviate ''simple auto-epistemic'' as ''simple AE.''

This is not to say that all wffs of interest in auto-epistemic reasoning must be of one of the three given forms. Far from it. However, these three types seem to be the simplest ones and arguably the commonest, and also plenty of questions arise even for them.

Note that Thm does not satisfy schema T -- this is essentially Gödel's theorem on consistency proofs, and also can be seen in Montague's theorem 3. That is, not all wffs Thm $\alpha .\rightarrow \alpha$ will be theorems of g. However, a theorem of Löb carries this much further, so that not a single *instance* of T will be provable except trivial ones for which α itself is provable. Call a wff Thm $\alpha .\rightarrow \alpha$ *trivial* if α is a theorem of S. We suppose here (as throughout the paper) that S has sufficient substitution properties, in this case the Fixed Point Lemma referred to earlier.

> **Löb's Theorem:** [see 2,3,22,35] If P is a provability predicate for a consistent theory S, then no non-trivial instance of schema T (for P) is provable in S.[35]

> **Corollary:** No NAE or non-trivial RAE wffs are provable in S (if P is Bel).

---

[35] This is a very striking result that seems counterintuitive at first. One consequence is that *no* wff of the form $\neg$P α can be a theorem of S if S is consistent. This also arises out of Gödel's theorem on consistency proofs, which is easily proved from Löb's theorem and conversely. Gödel's result -- his Second Incompleteness Theorem -- is that
$\neg(\exists x)$[Thm x & Thm $\neg$x] is not a theorem of S.

Thus if Bel is formalized (defined) as Thm, then *no* non-trivial RAE wff is a theorem of S. This is stronger than Montague's result that *some* RAE wff will fail to be provable (i.e., schema T in its full form clashes with rule N). Now we are faced with the fact that *each* instance of schema T (except trivial ones) clashes with the Thm interpretation of Bel.

What is going on here? After all, if a system (or reasoner) g does happen to have beliefs or theorems that are true (in some standard interpretation for which Thm has the standard meaning), then all wffs Thm α $.\rightarrow \alpha$ will also be true in that interpretation. But why then cannot g be made to prove this, since it is true? Well, on our suggested intuitions from section I, g can *come* to do so (by being given this extra information), but in the process g will change (as a formal system) and Thm will then characterize what g was, not what g is. Put differently, the result of Löb tells us g cannot know Thm to capture precisely *its* means of drawing conclusions; in fact it will not as soon as g thinks that it does! Or in AI terms, an ideal[36] g can never fully catch up declaratively with its own procedures for drawing conclusions.[37] For instance, the wff Bel 0=1 $.\rightarrow$ 0=1 will not, by Löb, be provable in S, even though it will be consistent with any reasonable such S. Now if we extend S to S´ = S + Bel 0=1 $.\rightarrow$ 0=1, S´ will be consistent but then Bel cannot be a provability predicate for S´ (unless S´ is inconsistent). Thus provability predicates are static: their properties do not remain invariant over additions of even very modest new introspective information.

Note that $\neg P\alpha$ v $\neg P\neg\alpha$ is a kind of consistency statement for a provability predicate P. So ¬Bel α v ¬Bel $\neg\alpha$ can also be regarded as a kind of self-consistency belief. Now, while this may in general be too strong for a realistic agent g, still certain cases of it seem unassailable. For instance, ¬Bel 0=1 should be concludable by g, on the basis that ¬ 0=1 and that Bel ¬ 0=1. Yet Löb precludes this for provability predicates. Thus we separate two studies: static provability systems, and dynamic introspection systems that allow for the incorporation of new beliefs while retaining invariant general or generic information about beliefs as a whole. The latter requires giving up provability predicates as models for Bel; in their place we

---

[36] i.e., consistent, logically omniscient, and knowing sufficient arithmetic.

[37] A similar notion is exploited in [29] to characterize certain forms of default reasoning. Also Konolige [17] treats a similar theme from a different formal perspective.

substitute what we shall call ‘‘introspection predicates.’’

It is true that Löb’s theorem applies to more than simply Thm; *any* provability predicate P(x) gives us |-/- ¬Pα for any α. However, the two postulates for a provability predicate in addition to rule N, namely PosInt and schema K, are not ones we should necessarily assume for Bel. That is, even though they may be true about g’s reasoning, they are not facts g will necessarily know about itself. We then have the following result, toward a dynamic theory of belief.

> **Theorem 7:** Let S be any consistent qualifi edly substitutive fi rst-order theory not including the predicate letter Bel in its language, and let AE be any set of Moorean or reverse auto-epistemic wffs over S. Then there is a consistent extension DYNA(S) in which all AE wffs are provable.
>
> **Proof:** Let M be a model for S, and then form a model M´ of S+AE by interpreting Bel as truth in M; this will satisfy all auto-epistemic wffs of AE. For given a wff Bel $\alpha .\rightarrow \alpha$, if α is true in M then α (and hence Bel $\alpha .\rightarrow \alpha$) already holds in M´ ; and if α is not true in M, then Bel α is false in M´ and so again Bel $\alpha .\rightarrow \alpha$ holds. For a wff $\alpha \rightarrow$ Bel α, if α holds in M´ then it also holds in M and so Bel α is true in M´ , making $\alpha \rightarrow$ Bel α true in M´ ; and if α does not hold in M´ then $\alpha \rightarrow$ Bel α holds trivially. Then the theory of M´ extends S and has all wffs in AE as theorems.
>
> **Corollary:** DYNA(S) above can be taken to obey DYNA(S) |-- Bel α whenever S |-- α.
>
> **Proof:** The same model M´ in the proof of Theorem 7 is a model of Bel α for all theorems α of S, and so again the theory of M´ serves.

It is necessary to restrict S and AE as above, namely S must not contain the symbol Bel, so that AE will not contain wffs in the language of S.[38] If α were allowed to be any wff in the *extended* language

[38]And thus the Corollary does *not* provide the full rule N for Bel.

including Bel, then we could easily create Liar-type wffs and run into Tarski's theorem.[39] But we have already done what Thm (or any provability predicate) cannot do, in having even one non-trivial auto-epistemic belief present. Note that in modelling Bel as provability for S, we have not made Bel the same as Thm, for Bel is in the language of the extension S+AE, not of S. That is, Bel is not Thm for S+AE, and so Löb does not apply to the extension. But by the same token, Bel then is not an introspection predicate in the extension either.

Do we get similar results to STAT(S) for cats and BB here? Yes, from the Corollary, letting S have the GK schema for True, and defining Know α be Bel α & True α as before, we have:

DYNA(S) |-- Know(C v ¬C)

DYNA(S) |-/- Know(BB v ¬BB)

DYNA(S) |-- Bel(α v ¬α) for all wffs α in the language of S.

Some natural cases have been missed, since we are working with a restricted language in which there is no nesting of belief or knowledge. Thus we still have not achieved our goal of making g highly introspective as to its own beliefs. In particular, rule N is present only in very restricted form, applied to theorems of S but not to the extension in which the predicate Bel enters the language. Is anything else lacking? What might we want, for Bel to be an introspection predicate? Several things may suggest themselves. However, among the simplest is what I shall call the *double-N* rule, NN, which amounts to our familiar rule N from S5 and its converse $N sup -1$, that is, α is a theorem of g iff Bel α is a theorem of g. Thus g can recognize (prove) it has α as a theorem (i.e., g can prove Bel α), precisely when it really does have α as theorem. This makes Bel in some sense ''correct''. Whether such a g is still realizable as a purely first-order theory is another matter. We are using rules of inference (N and its converse) outside of standard first-order logics, unless the logic in question obeys NN as a consequence of its axioms (note, for instance, that schema T in a theory makes $N sup -1$ is redundant). Thm does happen to obey NN, although due to Löb

[39]Or Montague's theorem; e.g., from MAE wffs we would get rule N, and from RAE wffs we would get schema T, making the by now familiar deadly mix.

this will not help us here.

Definition: P(x) is an *introspection predicate* for a theory S if it obeys rule NN:

S |-- Bel α iff S |-- α for all wffs α.

Lemma: If P and Q are introspection predicates for S, then for every term t,

S |-- P(t) iff S |-- Q(t).

Proof: trivial.

It might appear from the Lemma that at most one predicate (up to equivalence) could satisfy NN, thus forcing it to coincide with Thm. For NN seems to characterize fully just what atoms of its associated predicate (e.g., Bel) can hold. After all, Bel α will be forced on g as a conclusion whenever (and only when) α itself is a conclusion. This might then seem to limit our formal choices for Bel very severely, indeed perhaps force us back to Thm and the loss of simple AE wffs. However, this is not necessarily the case, and leads to an open problem below. Roughly, Löb might not ruin our chances at an auto-epistemic formalization for Bel, because there might co-exist more than one introspection predicate for the same theory.[40]

Now we might ask whether introspection should go even further than our new definition. In particular, the following may seem reasonable to consider:

Definition: A theory S with introspection predicate Bel is *fully-introspective* if whenever a wff α in the language of S is *not* a theorem of S, then ¬Bel α *is* a theorem of S.

A fully introspective reasoner then would always be able to tell correctly for every wff whether it believed it or not: S |-- Bel α or S |-- ¬Bel α for each α.

[40]The beliefs of g will still be (the same as) the theorems of S; but it is not obvious that this necessitates, say, S |-- Bel $\alpha .\rightarrow \alpha$ iff S |-- Thm $\alpha .\rightarrow \alpha$, despite the Lemma.

Now, such a reasoner, if consistent and knowing sufficient arithmetic, would have a non-recursively-enumerable set of theorems. Still, it may be of epistemological interest to know that in principle such reasoning could be envisioned. However, it is not to be.

> **Theorem 8:**[41] Every fully-introspective qualifiedly-substitutive first-order theory S is inconsistent.
>
> **Proof:** We form BB as usual, so that BB ↔ ¬Bel BB is a theorem of S. Now either BB is a theorem of S or it is not. If BB is provable then (rule N, from Bel's being an introspection predicate) so is Bel BB. But also from the biconditional we get ¬BB, a contradiction. Suppose then that BB is not provable in S. Then by negative introspection we get ¬Bel BB, hence (from the biconditional again) BB, and finally (rule N) Bel BB, contradiction. (Note that we used only rule N, not full NN, so actually we have a stronger result than the stated one.)

We seem to be stuck then with (at best) the more modest introspection notion of rule NN for Bel. To recapitulate: static notions of introspection as in Thm are subject to Löb while auto-epistemic versions of Bel ought not to be. We can do pretty well in a static (Thm- and NN-based) theory of belief, *if* we avoid consideration of AE wffs (Theorem 6). And conversely we can do pretty well in an AE theory of belief *if* we avoid NN (Theorem 7).

It is getting both together that remains problematic. We offer as a "*Belief Doctrine*" that Bel should satisfy both NN *and* a wide variety of cases of MAE, RAE, and NAE. It is then worth seeing whether some introspection predicate (other than Thm) might achieve this. In short, what kind of theories DYNA(S) obey rule NN? We know that if Bel is such a predicate, and if it obeys (within a theory S) schema K, then by Löb it cannot obey PosInt and so will *not* coincide with Thm. In our terms Bel would be dynamic and generic, failing to correspond in detail to the actual reasoning mechanisms of g. But we have argued that this is appropriate for introspective reasoning. I leave the existence of such an AE- and NN-based (but not Thm-

[41] This was inspired by a conversation with Richard Weyhrauch, Sardinia, October 1986. This result is not necessarily negative. Commonsense reasoners may well be inconsistent (see [29]), and yet have interesting formal properties ([6]).

based) Bel as an open problem:

Open Problem: What subsets of MAE, RAE, and NAE are consistent with NN?

Regarding RAE in particular, we know of course $\varnothing$ is fine (take Bel to be Thm), and that using *all* reverse wffs is never so (Montague). In fact, we cannot include Bel BB $\rightarrow$ BB for the same reason. But is there any non-empty subset of RAE that is consistent with NN? If so, then there can be more than one introspection predicate for one and the same theory: a provability predicate will not be the only choice, despite the Lemma.

In terms of our notion of flattened layers, certain global statements such as PosInt and K will not allow mundane local statements such as RAE to be present. If we ''force'' them into an extension, we simply end up changing the theory so that we are, after all, looking back at the earlier ''entirety'' rather than the present (new) one. The Open Problem then is asking how much local perturbation (individual instances of simple AE wffs) we can get away with and yet preserve a useful amount of globality in the form of rule NN. Note that we did get a *weak* answer, in the Corollary to Theorem 7, since rule N is obeyed there for wffs α that do not themselves contain the symbol Bel, and this already gives a number of cases useful in commonsense.

VIII. Conclusions

When a formal language is endowed with self-referential capabilities, especially in the presence of unqualifiedly substitutive mechanisms, difficulties of contradiction can easily arise. This holds for modal as well as (pure) first-order logics. However, the features of self-reference and substitutivity appear fundamental to any broad knowledge representation medium. Moreover, when remedies are taken, the modal treatments seems to offer no advantage over the first-order ones, and indeed the latter carry advantages of their own.

One can argue that although an agent g can't *know* its beliefs to be true, still they *might* be true by good luck (or by the clever design of the agent's reasoning devices by a godlike artificial intelligencer), and all g's inference rules might be sound as well. But then, if g is an ideal reasoner, wouldn't it be appropriate

for g to believe this too? Wouldn't such an ideal g be able to believe Bel $\alpha .\rightarrow \alpha$ for all α? The odd answer (which we have seen in Montague's Theorem 3) is: not if g's beliefs are to be consistent, which of course they must be if they are to be true. But this can be seen as an overly bold flattening of an essentially layered concept of Know. Of course, Thm is also a kind of flattening, but one in which no new information is to be brought in, by the very concept of Thm which pins down precisely what is allowed.

In fact, Löb shows us even more: that schema T can be allowed for *no* α except trivial cases, unless we give up provability as the measure of belief; then PosInt and other vestiges of pinned-down provability must be left aside. But since auto-epistemic reasoning depends on (certain instances of) schema T, agents will have to rely on dynamic (perturbation-tolerant) reasoning about their own beliefs. They cannot fully introspect; in particular, they will have to rely on pragmatic means to tell, for instance, that they do *not* have a certain belief.

I have been occupied here in showing that, after all, a flattened picture of commonsense may be available, a once-and-for-all set of wffs closed under certain procedures. This is what DYNA(S) really does, and it does so by leaving out wffs that might force the meaning of Bel to no longer be an introspection predicate. That is, the tradition of research that I have been exploring throughout this paper is in the mold of finding a fixed set of conclusions that g can believe, *but* allowing a reasonable amount of introspective self-reference at the same time.

Perhaps ironically, the static and flattened provability predicates, such as Thm, which obey PosInt and schema T, and which were found to thwart some attempts at a commonsense view, are the ones that would force a major *change* in any agent dealing with them, in a cycle of ever-expanding interpretations of its growing mechanisms of proof. Thus our dynamic version of Bel is not really one for a real-time agent at all. The conclusion, then, is that people have sought fixed formulations for belief using what are intrinsically non-fixed notions of self-description. For a single (and hence flattened) theory of belief to be viable, it must deal with predicates that are tolerant of self-description in a context of simple AE wffs; such theories are what I have called dynamic. The real trouble is that if Bel is made to look at itself closely enough, then it ends up describing a theory different from the one being investigated. This is fine if a cycle of ever-

stronger theories is the focus of interest. However, single theories are vastly simpler to study, and so it is worth seeing how far this flattened approach can be carried.

The formalization of (fixed theories of) knowledge and belief still faces conceptual difficulties, especially in the case of agents whose beliefs are closed under logical consequence. In particular, it is unclear whether rule NN can be made consistent with reasonable commonsense instances of schema T. But it also appears that the study of agents with limited reasoning power, as has been initiated in [6,7,8,19,26,27] is great need of further study. Key to those approaches is the *absence* of the rule of inference "from α infer Know α" (with respect to a deductive engine which is sound and complete). Although some of these latter efforts utilize modal formulations, our work here strongly suggests that this is more a matter of taste than any real technical distinction, and thus that it may be preferable to stick with a common formal language to facilitate comparison in future work.

Acknowledgement

This research was supported in part by grants from the following institutions:

(1) U. S. Army Research Office (DAAG29-85-K-0177)

(2) The Martin Marietta Corporation

I would like to thank the following individuals for discussions that prompted me to carry out the elaboration of the ideas presented herein: Ray Reiter, Nils Nilsson, Jack Minker, Jennifer Drapkin, Michael Miller, Rosalie Hall, Brian Haugh, Kurt Konolige, Dana Nau, Jim Reggia, Maria Simi, Jim des Rivieres, Hector Levesque, Richard Weyhrauch, Bill Gasarch, Barry Richards, and Ian Pratt. I would also like to thank Pat Hayes for encouragement, and an anonymous referee for many helpful comments.

Bibliography

(1) Asher, N. and Kamp, H. The knower's paradox and representational theories of attitudes, *Proceedings, Theoretical Aspects of Reasoning About Knowledge*, 1986, 131-147.

(2) Boolos, G. *The Unprovability of Consistency*. Cambridge University Press, 1979.

(3) Boolos, G. and Jeffrey, R. *Computability and Logic*, 2nd edition. Cambridge University Press, 1980.

(4) Burge, T. Epistemic paradox, *J. Phil., 81* (1984), pp.5-29.

(5) Chellas, B. *Modal Logic*. Cambridge University Press, 1980.

(6) Drapkin, J. and Perlis, D. Step-logics: an alternative approach to limited reasoning. *Proc. European Conf. on Artif. Intell.*, 1986, 160-163.

(7) Eberle, R. A logic of believing, knowing and inferring. *Synthese 26* (1974) pp.356-382.

(8) Fagin, R. and Halpern, J. Belief, awareness, and limited reasoning: preliminary report. *IJCAI 85*, pp.491-501.

(9) Feferman, S. Transfinite recursive progressions of axiomatic theories, *J. Symbolic Logic, 27* (1962), 259-316.

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

(11) Gettier, E. Is justified true belief knowledge? *Analysis 23* (1963), 121-123.

(12) Gilmore, P. The consistency of partial set theory..., in: T. Jech (ed.) *Axiomatic Set Theory*. Amer. Math. Soc., 1974.

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

(14) Halpern, J. and Moses, Y. A guide to the modal logics of knowledge and belief: preliminary draft. *IJCAI 85*, pp.480-490.

(15) Hintikka, J. *Knowledge and Belief*. Cornell University Press, 1962.

(16) Konolige, K. A first-order formalisation of knowledge and action for a multi-agent planning system, in: J. E. Hayes et al (eds.) *Machine Intelligence 10*. Wiley, 1982. pp.503-508.

(17) Konolige, K. A computational theory of belief introspection. *IJCAI 85*, pp.502-508, Los Angeles (Ninth International Joint Conference on Artificial Intelligence).

(18) Kripke, S. Outline of a theory of truth, *J. Phil., 72* (1975), pp.690-716.

(19) Levesque, H. A logic of implicit and explicit belief. *Proc 3rd National Conf. on Artificial Intelligence*, 1984, pp.198-202.

(20) Levesque, H. Foundations of a functional approach to knowledge representation. *Artificial Intelligence 23*, 1984, pp.155-212.

(21) McCarthy, J. First order theories of individual concepts and propositions. In J. E. Hayes et al (eds.) *Machine Intelligence 9*. Wiley, 1979. pp.129-147. Also in R. Brachman and H. Levesque (eds.) *Readings in Knowledge Representation*. Morgan Kaufmann, 1982.

(22) Mendelson, E. *Introduction to Mathematical Logic*, 3rd edition. Wadsworth, 1987, Belmont, CA.

(23) Montague, R. Syntactical treatments of modality.... *Acta Philos. Fenn. 16*, (1963) pp.153-167.

(24) Moore, R. Semantical considerations on non-monotonic logic, *Artificial Intelligence 25*, (1984) 75-94.

(25) Partee, B. (ed.) *Montague Grammars*. Academic Press, 1976.

(26) Perlis, D. Language, computation, and reality. Ph.D. Thesis, U of Rochester, 1981.

(27) Perlis, D. Nonmonotonicity and real-time reasoning, *AAAI Workshop on Nonmonotonic Reasoning*, 1984.

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

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

(30) Perlis, D. Self-reference, knowledge, belief, and modality. *Proc 5th National Conference on AI*, 1986, pp.416-420.

(31) Quine, W. Concatenation as a basis for arithmetic. *J. Symb. Logic, 11* (1946).

(32) Rieger, C. Conceptual memory... Ph.D. Thesis, Stanford University, 1974.

(33) des Rivieres, J. and Levesque, H. The consistency of syntactical treatments of knowledge. *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge*, pp. 115-130. 1986. Morgan Kaufmann.

(34) Smith, B. Varieties of self-reference. *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge*, pp. 19-43. 1986. Morgan Kaufmann.

(35) Smorynski, C. *Self-Reference and Modal Logic*. Springer-Verlag, 1985, New York.

(36) Stich, S. *From Folk Psychology to Cognitive Science: the Case Against Belief*. MIT Press, 1983.

(37) Tarski, A. Der Wahrheitsbegriff in den formalisierten Sprachen, *Studia Philos., 1* (1936), 261-405.

(38) Thomason, R. A note on syntactical treatments of modality, *Synthese* 44 (1980), pp 391-395.

(39) Vardi, M. A model-theoretic analysis of monotonic knowledge. *IJCAI 85*, pp.509-512.