Self-Reference, Knowledge, Belief, and Modality

Donald Perlis

University of Maryland
Department of Computer Science
College Park, Maryland 20742

ABSTRACT

An apparently negative result of Montague has diverted research in formal modalities away from syntactic ("first-order") approaches, encouraging rather weak and semantically complex modal formalisms, especially in representing epistemic notions. We show that, Montague notwithstanding, consistent and straightforward first-order syntactic treatments of modality are possible, especially for belief and knowledge; that the usual modal treatments are on no firmer ground than first-order ones when endowed with self-reference; and that in the latter case there still are remedies.

I. INTRODUCTION

We are in this paper particularly concerned with the concepts of belief and knowledge, in their relation to (and in the avoidance of) self-referential paradox. Let us write Bel(x) and K(x) to indicate that x is believed, resp. known, by an implicit agent g. The syntactic status of x is one of the issues to be addressed. If Bel and K are predicate symbols, then x is an ordinary first-order term which in particular may be the name of a sentence, as in Bel("Snow is white"). On the other hand, if Bel and K are modal operators, then x will be a well-formed formula, as in Bel(Snow is white). In [21] it was suggested that for an intelligent reasoner g, a self-referential language is desirable in order to represent such notions as that g has a false belief, this itself being a likely belief of g. We may write, for instance, (Ex)(Bel(x) & ¬True(x)). But if this very wff is to be a belief of g, then it too can serve (either in quoted first-order form, or in formula -- modal -- form) as argument within another belief formula. We contend that this is such a basic aspect of language and thought that any reasonable representational mechanism for common-sense reasoning must include facilities for expression of self-reference and syntactic substitutions. (Note that Rieger [25] calls essentially this same notion *referenceability.*) We will show 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. A longer version of the present paper [22] contains proofs of theorems.

The example of an agent having a false belief is a key one. For it distinguishes between what we might call weak and strong languages. In particular, traditional modal languages with an operator Bel for belief do not ordinarily allow variable operands; this would amount to something like a second-order modal language. This means that having a false belief is not straightforwardly representable in such languages. Of course, the same can be said for traditional first-order languages. Thus traditional logics tend to be weak. However, the obvious remedy of introducing names for formulas as in "Snow is white" above, leads to familiar problems of inconsistency. While this has been taken by some to mean that epistemic notions such as Bel should be left as modal operators rather than risk inconsistency in a first-order setting with names, on the other hand it is too weak to accommodate the needs of artificial intelligence (as in the false belief case). Here we will investigate the introduction of names into formal treatments of belief and knowledge, and ways to retain consistency while retaining as well the strong feature of referenceability.

II. PRELIMINARY RESULTS

We shall call a theory T (over a language L) with mechanisms for expressing *and asserting* all such substitutions *unqualifiedly substitutive.* The hallmark of an unqualifiedly substitutive language is that it possesses an operator Sub(P,Q,a,n) directly asserting 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, only a rather special use of an operator such as Sub 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 we will write simply Sub(P,Q,a). Contexts will vary slightly in that sometimes all occurrences of a will be identical, sometimes one occurrence will be quoted. We beg the reader's indulgence in sloppily using the same notation for both cases. We also write P[Q/a] for the result of substitution in either case.

This research was supported in part by grants from the following institutions: U. S. Army Research Office (DAAG29-85-K-0177) and the Martin Marietta Corporation.

As will be seen, the *asserting* 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 "A" carries the meaning A. This is often represented as defining a truth predicate: True("A") is to tell us that the sentence "A" carries a true meaning, i.e., A. 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, as a gloss for Sub(P,Q,a), the (perhaps pseudo-) 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.

For precision's sake we offer the following definition: Let T 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) $\leftrightarrow$ 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 "A" for each wff A of L). Then T is said to be an *unqualifiedly substitutive* theory.

Theorem 1: Let T be an unqualifiedly substitutive first-order theory. Then T is inconsistent.

For the proof of this and subsequent results, see longer version.

In [19] and [21] the difficulty of formalizing a truth predicate in first-order languages was circumvented, based on ideas in [6] and [14]. 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 T formed from extending a consistent theory T' not involving the symbol Sub, by the addition of the (qualified) schema Sub("P","Q",a) $\leftrightarrow$ P[Q/a]*, where α^* is the result of replacing $\neg$Sub("P",...) by Sub("$\neg$P",...) in α, is consistent.

What we wish to investigate eventually (section IV) is the extent to which the same result holds for modal theories. First we turn to a question addressed by Montague [16] concerning first-order analogues of certain modal theories.

III. FIRST-ORDER ANALOGUES

There are some solid technical benefits that would accrue from a first-order approach to propositional attitudes; in particular, in the words of Montague [16], "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." Motivated by these concerns, Montague [16] applied this approach to a modality for necessity. That is, writing Nec ("A") instead of Nec A 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 modal languages. That is, first-order logics have richer sets of formulas than have traditional modal logics. For 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?

It is of separate interest whether a first-order logic 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 26]. 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, and 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.

Montague studied several systems related to S5, with the particular aim of changing Nec into a predicate symbol applied to names of formulas. We need not present details of these modal variants in order to state the following variation on a result of his, where we freely adopt the symbol I (for information) in place of Nec. (Note that under an assumption of omniscience, S5 plausibly formalizes the "information" a reasoning agent may have. We for the moment avoid the terms "knowledge" and "belief" in favor of this more neutral expression.)

If T is a first-order theory with function symbols *sub* and *quote* of three and one arguments, respectively, and supplied with a term "α" for each wff α as well as axioms defining sub and quote appropriately, i.e., quote(e) = "e" for each constant symbol e, and sub(P,Q,a) = "<P[Q/a]>", i.e., the name of the result of the indicated substitution, then T is *first-order self-referential*.

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

What does this result tell us? It appears that even a very weak subtheory of S5, 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 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.

IV. 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 unqualifiedly 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) ↔ 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 extended by the qualified Sub predicate with axiom Sub(x,y,z) ↔ True(sub(x,y,z)) and associated axioms for True and sub, then M is inconsistent.

We then consider the second alternative mentioned above, namely, that Sub(P,Q,a) be a modality in which P and Q are formulas. It turns out that even without variable arguments to modalities, contradiction arises. Specifically, we define T to be an *unqualifiedly substitutive* modal logic if T 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 expression, it suffices to refer to whole formulas.)

Theorem 5: Any unqualifiedly substitutive modal theory is inconsistent.

So S5 itself is inconsistent with either form of self-reference that naturally arises. We now turn to remedies of this situation, hinging on separating the two troublesome features, namely the schema I("α")→α, and the rule for inferring I("α") from α. This will at last split I into the two cases of Bel and K.

V. CONSISTENT FORMALIZATIONS

We suggest (as is fairly common) that Kx means x is among those beliefs of g that are true. It is important to emphasize that K is to be a symbol in g's own language, so that Kx means *to g* that x is one of its true beliefs, even though in general g cannot identify which these are! That is, in general g can only refer in the abstract to its knowledge (true beliefs). Indeed, all g's beliefs are (by definition) believed by g; as soon as any one is suspected of being false, it is no longer believed. 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 K) is a proper subset of (the extension of) Bel, but can give no examples of the relative complement! Thus general assertions about K (such as that Kx → x) are part of g's external view of itself, so to speak, comparing its belief set to an unauthenticated outer world of truth, while assertions about particular elements are part of an internal view of Bel relevant to working directly with individual beliefs as things to use in planning and acting.

That is, we are suggesting two postulates, one for individual beliefs (from x infer Bel x), and one for the totality of beliefs and knowledge ((x)(Kx → x)). It is mixing the two that is problematic. A judicious approximation to a mix is however possible, as the following results indicate.

Theorem 6: Let T be any consistent qualifiedly substitutive first-order theory. Then there is a consistent first-order theory Int, which is an extension of T having predicate symbol Bel, and obeying the subsumed rules |-- Bel("α") iff |-- α.

A still stronger result would seem to arise if we simply formally identify Bel with the predicate Thm via the use of Godel numbers. This of course requires incorporating a certain amount of number theory into the agent's reasoning, but given the rather powerful assumptions that go into most logics of knowledge (e.g., that all logical consequences of an agent's knowledge are also known to the agent), this seems easy to grant. Moreover, the use of substitution is virtually tantamount to the introduction of a certain amount of arithmetic in any case (see Quine [24]), and we have argued that substitution is an essential feature of commonsense reasoning.

We then are left with the suggestion that a theory along the lines of Int is appropriate for a formalization of belief. It allows for introspection, to the extent that if α is believed (affirmed) then that very fact is also believed, and conversely. But it makes no claim that totality of its beliefs need be true, even though each separate belief is of course asserted, and hence taken to be the true. The strong contrast with a logic of knowledge is shown in the following result, which is based on [6,14,19,21].

Theorem 7: Let T be any consistent qualifiedly substitutive first-order theory. Then there is a consistent first-order theory Ext, which is an extension of T having predicate symbol K, and axioms K("α") → α for each wff α, and obeying the (subsumed) rule that |-- K("α") whenever |-- α^*, where α^* is the result of replacing ¬K("...") by K("¬...") in α.

From α can be inferred Bel("α"), but to infer K("α") more is needed, namely α^*, the positive form of α. This can be interpreted in various ways depending on the underlying conceptualization of the formalism either as the agent's-eye-view of the world, or as our own god's-eye view. The longer paper describes more fully the significance of this and of the *; the latter is the critical distinction between K (Ext) and Bel (Int). Note that for most wffs α, α^* *is* α.

As with Int and belief, we suggest Ext as a possibly appropriate formalization of the notion of an agent's knowledge. But contradiction will arise if the agent tries to combine Int and Ext into one theory (i.e., with Bel and K conflated). We hope to have suggested why such a combination is not appropriate. More motivational discussion is provided in the longer paper. Elsewhere [23] we have investigated further the ramifications of the idea that an agent's beliefs are not all true (known), and that a rational agent will believe *that*.

One more observation is in order. An agent g may reason about *both* its beliefs *and* its knowledge, simply by combining the theories Int and Ext, but keeping K and Bel as separate predicates. One can even relate them judiciously, such as by the axiom Kx → Bel x. This we state in the following theorem. The extent to which theories such as S5 can be viewed as "incorporated" within theories such as Omni below is discussed in the longer paper.

Theorem 8: Let T be any consistent qualifiedly substitutive first-order theory. Then there is a consistent first-order theory Omni, which is an extension of T having predicate symbols Bel and K, the axioms of Int and Ext as in Theorems 6 and 7, and axiom Kx → Bel x.

Note that if we introduce Kx by definition to be Bel(x) & True(x), then we can simply use axioms and rules for True as in [21]. This then provides a slightly sharper version of Theorems 7 and 8, in which for instance K("¬K("α")") may be inferred from [¬Bel("α") v True("¬α")].

VI. 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* his 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 Bel x → x? The odd answer (which we have seen in Theorem 3) is: not if g's beliefs are to be consistent, which of course they must be if they are to be true. This can be seen also as an illicit identification of Bel with K.

The proposed theories Int, Ext, and Omni appear relevant to the study of omniscient reasoning. For limited reasoning, alterations will be needed. Further related work, especially to the latter, includes [1,2,3,4,5,7,8,9,12,15,17,27].

ACKNOWLEDGEMENTS

I would like to thank the following individuals for highly motivating discussions that prompted me to carry out the elaboration of the ideas presented herein: Ray Reiter, Nils Nilsson, Jack Minker, Jennifer Drapkin, Michael Miller, Brian Haugh, Kurt Konolige, Dana Nau, and Jim Reggia.

REFERENCES

(1) Drapkin, J. and Perlis, D. Step-logics: an alternative approach to limited reasoning. Proc. Eur. Conf. on Art. Intell. 1986.

(2) Drapkin, J. and Perlis, D. A preliminary excursion into step-logics. Proc. Intl. Symp. on Methodologies for Intell. Systems 1986.

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

(4) Fagin, R., Halpern, J., and Vardi, M. A model-theoretic analysis of knowledge. Proc. 25th IEEE Symp. on Foundations of Computer Science, 1984, pp.268-278.

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

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

(7) Halpern, J. and Moses, Y. Towards a theory of knowledge and ignorance. AAAI workshop on Nonmonotonic Reasoning, 1984.

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

(9) Hintikka, J. Knowledge and belief. Cornell University Press, 1962.

(10) Hughes G., and Cresswell, M. An introduction to modal logic. Methuen, 1968.

(11) Israel, D. What's wrong with nonmonotonic logic? Proc. First Annual National Conference on Artificial Intelligence, 1980.

(12) Konolige, K. A computational theory of belief introspection. IJCAI 85, pp.503-508.

(13) Kripke, S. Semantical analysis of modal logic. Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik, 9, (1963), pp.67-96.

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

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

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

(17) Moore, R. Reasoning about knowledge and action. IJCAI 77, pp.223-227.

(18) Partee, B. (ed.) Montague grammars. Academic Press, 1976.

(19) Perlis, D. Language, computation, and reality. Ph.D. thesis. U of Rochester, 1981.

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

(21) Perlis, D. Languages with self-reference I: foundations. AIJ 25, 1985.

(22) Perlis, D. Languages with self-reference II. U of Md Tech Report.

(23) Perlis, D. On the consistency of commonsense reasoning. U of Md Tech Report.

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

(25) Rieger, C. Conceptual memory... Ph.D. thesis. Stanford University, 1974.

(26) des Rivieres, J. and Levesque, H. The consistency of syntactical treatments of knowledge. To appear, Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge, 1986.

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