

Towards a Paradigm Shift in Belief Representation Methodology

John A. Barnden

Computing Research Laboratory
New Mexico State University
Box 30001, Las Cruces, NM 88003

KEY PHRASES: belief representation, propositional attitudes, iterated attitudes, commonsense reasoning, metaphors of mind, AI methodology.

ABSTRACT

Research programs must often divide issues into manageable sub-issues. The assumption is that an approach developed to cope with a sub-issue can later be integrated into an approach to the whole issue — possibly after some tinkering with the sub-approach, but without affecting its fundamental features. However, the present paper examines a case where an AI issue has been divided in a way that is, apparently, harmless and natural, but is, actually, fundamentally out of tune with the realities of the issue. As a result, some approaches developed for a certain sub-issue cannot be extended to a total approach without fundamental modification. The issue in question is that of representing and reasoning about people’s beliefs, hopes, intentions and other “propositional attitudes”, and/or interpreting natural language sentences that report propositional attitudes. Researchers have, quite understandably, de-emphasized the problem of dealing in detail with nested attitudes (e.g. hopes about beliefs, beliefs about intentions about beliefs), in favour of concentrating on the sub-issue of non-nested attitudes. Unfortunately, a wide variety of approaches to attitudes are prone to a deep but somewhat subtle problem when they are applied to nested attitudes. This problem can be very roughly described as an AI system’s unwitting imputation of its own arcane “theory” of propositional attitudes to other agents. The details of this phenomenon have been published elsewhere by the author: the present paper therefore merely sketches it, and concentrates instead on the methodological lessons to be drawn, both for propositional attitude research and, more tentatively, for AI in general. The paper also summarizes an argument (presented more completely elsewhere) for an approach to attitude representation based in part on metaphors of mind that are commonly used by people. This proposed new research direction should ultimately coax propositional attitude research out of the logical armchair and into the psychological laboratory.

1: INTRODUCTION

Research programs must often divide issues into manageable sub-issues. This problem-reduction method has often been used in AI, both at the “macro” level of the major division of the field into natural language processing, knowledge representation, perception, and so on, and at the “micro” level of divisions of individual subfields — for instance, the division of natural language understanding into syntax, semantics and pragmatics.

No doubt, such problem reductions have led to useful advances. On the other hand, the fruitfulness or safety of traditional problem reductions has certainly been questioned. Thus, the general nature of my claim in this paper — that problem reduction can lead to major difficulties — is hardly novel. My real point, however, is more specific. It is both to give further exposure to a specific case of the dangers of problem reduction, a case that is of considerable technical interest on its own merits, and to draw some specific methodological morals not only for research in the particular AI subfield in question but also for AI in general. Note, however, that the idea that problem reduction is necessarily bad is *not* one of these morals.

Our particular case of the dangers of problem reduction arises in what we can call the “propositional-attitude subfield” of AI. This subfield is concerned with two inseparable questions: (a) how to represent and reason about the beliefs, hopes, intentions and other “propositional attitudes” of cognitive agents (people mainly); and (b) how to interpret “attitude reports” — natural language sentences that report propositional attitudes. It turns out to be presentationally convenient, as well as sufficient for most of our purposes, to approach the propositional attitude subfield through (b). Of course, (b) intimately involves (a), for several reasons that we will not go into. What our approach does skirt is the question of attitude representations that do not emanate directly or indirectly from natural language understanding processes. Our considerations can, however, be extended to address such representations.

Some prototypical examples of attitude reports are “*Mike believes that Voluptia is clever*”, “*Mike intends to kick the ball*”, and “*Mike wants Voluptia to win*”. Here the clauses “*Voluptia is clever*”, “*[Mike] to kick the ball*”, and “*Voluptia to win*” are called the “complements” of these reports. Since the complements are not themselves attitude reports, the sentences are *non-nested* attitude reports. An example of a *nested* (or *iterated*) attitude report is

George believes that Mike believes that Voluptia is clever.

Although this example involves the nesting of one belief inside another belief, it seems more common in practice for a nesting to be that of an attitude of one type inside an attitude of a different type: e.g. the nesting of a belief within a hope, or an intention within a belief, and so on. However, it is common in the field to stick just to belief, which is generally regarded as the prototypical sort

of propositional attitude. For the most part we will adopt this simplifying practice, for expository purposes.

Another *expository* simplification is that we will ignore the fact that in practice the interpretation of a sentence is always within the context of the speaker's assumed attitude towards the content of the sentence. This has the effect of embedding the sentence within at least one level of attitude nesting, so that an overtly non-nested attitude report actually leads to a *nested* attitude representation internally. If we took account of this we would make our discussion much more complex without altering the fundamental issues (or so I submit — there is no space to argue the point here).

Few authors have explicitly allowed a concern with nested attitudes to have a fundamental effect on the basic representational features of schemes for attitude representation and attitude-report interpretation. In most cases, the detailed technical features of a scheme are designed with a focus on non-nested attitudes, and then extended in a simple way to nested attitudes. There are some major exceptions, including Creary (1979) [which is strongly related to ideas in Church 1951, 1973, 1974], Cresswell (1985), and Wilks & Ballim (1987). See also Creary & Pollard (1985) and Wilks & Bien (1983). Many other authors have placed a lot of emphasis in other ways on nested attitudes — consider work such as Konolige (1983), Maida (1986), Rapaport (1986), and the large body of research based on modal logic. Unfortunately, a wide variety of schemes for attitude representation are prone to a deep but somewhat subtle problem when they are applied to nested attitudes. This problem can be *very roughly* described as an AI system's unwitting imputation of its own "theory" of propositional attitudes to people, where the arcaneness of that theory is such that any ordinary person is unlikely to think in terms of it. The problem has been overlooked, despite the work that has been done on nested attitudes, and despite connections to some theoretical issues involving non-nested attitudes. Certain methodological tendencies, which are the focus of this paper, have contributed to this neglect. They include: the practice of uninhibitedly using atomic symbols to stand for complex properties and relationships (other than attitudes themselves); and the neglect of natural language sentences *about* the entities, properties and relationships proposed in AI theories themselves.

The above casting of our problem in terms of theory imputation is a little inaccurate. More accurately, the problem is one of "*implausible entailment*".¹ Let us assume that a natural language understanding system called SYS converts incoming sentences into formal representational expressions that are intended to capture the meaning of the sentences. We say these expressions

¹ Elsewhere I have used the term "unwarranted entailment". However, by "unwarranted" I mean "unsound" in the sense mentioned below, *not* "implausible". Rather, implausible entailments are a special case of unsound ones. I now feel that it is clearer to avoid the term "unwarranted".

are SYS's "formal renderings" of the sentences. We will mostly be considering pairs of reports of the form

((1)) *George believes that Mike believes that C*

((2)) *George believes that D*

where D is an *an explication or elaboration of the idea that Mike believes that C*. It then often turns out that SYS's scheme for assigning formal renderings forces one of the sentences to follow from the other, in effect. If D involves *arcane, theoretical notions used in the formal approach itself*, and is therefore at a meta-theoretical level, the entailment is likely to be highly *implausible*.

This implausible-entailment phenomenon, which is detailed in Barnden (1986a, 1987a,b; see also 1983, 1986b) does not arise for all formal approaches, but is sufficiently pervasive and subtle to be noteworthy. Although the phenomenon involves specialized, meta-theoretic sentences, it nevertheless presents a practical problem. Also, it is by no means divorced from long-standing issues in AI and philosophy: in particular, it is strongly connected to "referential opacity" and the "(hyper-)intensionality" of attitude reports, to the "paradox of analysis" in philosophy [Langford 1942, Moore 1942], to the problem of "logical omniscience" [e.g. Levesque 1984, Fagin & Halpern 1987, Barnden 1988], to the puzzles about identity statements raised in the context of Fregean accounts of sense and denotation [e.g. Linksy 1983], and to problems with theory-laden terms [e.g. Partee 1982] and synonymy [e.g. Mates 1950]. These issues can themselves largely be seen as unsound-entailment issues, but work on them does not appear to have had much to say about our special implausible-entailment phenomenon. In fact, I have seen no evidence to suggest that AI researchers or philosophers already knew of the implausible-entailment problem at all. To my knowledge it has not been discussed in the literature.

The plan of the paper is as follows. Section 2 describes a particular version of our implausible-entailment problem. Section 3, the core of the paper, diagnoses the problem, discussing the role of the various methodological tendencies giving rise to it. Section 4 motivates and summarizes an approach to remedying the problem by appeal to commonsense views that people have of attitudes. In particular, the approach brings in the metaphors for mind that appear to be prevalent in realistic discourse about attitudes. This suggests that propositional attitude research should turn more towards psychological research than it has done. Section 5 sets out some morals for AI as a whole as well as for the propositional-attitude subfield. Section 6 is the conclusion.

2: THE SYMPTOMS

In the interests of brevity, we shall only consider in detail the way in which our implausible entailment problem arises when SYS renders attitude reports into expressions of *quotational logic* in a prototypical way. (See e.g. Haas 1986, Maida 1986, Morgenstern 1986, Perlis 1985, and Quine 1981 for a sample of quotational-logic approaches.) The implausible entailment problems raised by other approaches are analogous, and will be described briefly. I do not claim to have explicitly demonstrated that the problem exists in all major recent approaches to propositional attitudes. (In particular, it does not arise in existing modal-logic approaches, but there are other problems with these). However, I submit that my demonstration of its existence in a variety of relatively recent important approaches is enough to elevate the problem to the status of a trap lying in wait for propositional attitude researchers in general.

In fact, it even seems allowable to place the “burden of proof” on the designers of future attitude representation schemes — that is, to oblige the designers to show that their schemes do *not* suffer from the problem. At present I know of no precise method for demonstrating that a scheme avoids the problem. It seems fair, however, to recommend that a designer try to construct examples of the problem that are analogous to those reported briefly below and discussed at greater length in Barnden (1987b).

2.1: The Quotational-Logic Case

We will see that a SYS using quotational logic can easily effect an implausible conflation of a pair of sentences of the form (1), (2), in the sense of implausibly giving them the same formal rendering. Suppose that SYS would render the sentence “*Voluptia is clever*” as the formula `clever(voluptia)`. (We adopt throughout the convention that a person with a given proper name, e.g. Voluptia, is denoted in the logic by an individual constant formed by de-capitalizing the proper name, e.g. to get `voluptia`. We take any given proper name to refer to a unique person.) Then SYS would use the *quotation* of that formula in rendering an attitude report whose complement is that sentence. For instance, the attitude report

((3)) *Mike believes that Voluptia is clever*

would be rendered as

((4)) `BF(mike, 'clever(voluptia)')`.

Here BF is the symbol for a belief predicate that relates agents to formulae expressing their beliefs. It is important to realize that the use of BF *does not necessarily convey that the agent is entertaining*

the formula in his/her head. BF is used simply to give theoretical characterizations of what it is that believers believe.

The quotational expression ‘**clever(voluptia)**’ (including the quotation marks) is just a special sort of individual constant symbol, denoting the formula **clever(voluptia)**. Notice that we have not left first-order logic, although the logic is unusual in that its formulae themselves form part of the world which the logic is being used to represent.

The general idea, then, is that the rendering of an attitude report is achieved by quoting the rendering of the report’s complement, and then using this quotational expression as an argument in an application of the attitude predicate BF. More precisely, we assume that SYS proceeds according to the following rule:

Interpretation Rule: SYS renders an attitude report of form *X believes that C* as the formula $\text{BF}(x, Q)$, where x is the internal symbol corresponding to proper name X , and Q is the *quotation* of the formal rendering of the complement C of the sentence.

The rule captures the basic idea of the methods used in, for instance, the AI proposals of Haas (1986), Maida (1986), Morgenstern (1986) and Perlis (1985). Notice the following crucial fact:

Compositionality Effect: If the complements $C1$, $C2$ of two attitude reports “*X believes that C1*” and “*X believes that C2*” have the same formal rendering, then so do those attitude reports themselves.

The Interpretation Rule is over-simplified in many respects, such as in ignoring the important issues of lexical ambiguities and of pronouns within the complements. More pertinently, the rule covers only one way in which attitude reports can be read (interpreted) — namely, a fully “de-dicto” way of reading them. The question of the variety of ways in which attitude reports can be read is a complex and controversial one, with most investigators artificially concentrating on just one or two types of reading (typically, de-dicto and “de-re”). However, the above Rule is enough for our simplified account of the implausible-entailment problem in this paper. (A fuller discussion is included in Barnden 1987b.) The important thing to observe is that the type of reading captured in the Rule is one in which, intuitively, all the concepts mentioned in the attitude report’s complement play an explicit role in the agent X ’s belief. For instance, the Rule takes sentence (3) to convey that Mike is thinking both of Voluptia, under some mental representation or other, and of the property of cleverness, again under some mental representation or other. Further, it would be in line with philosophical and AI research on attitudes to assume that these mental representations are restricted in some special way, such as being required to be Mike’s *standard* mental representations for Voluptia and cleverness respectively.

Having laid the groundwork, we are ready to build our argument. I make two simplifying assumptions. First, I assume that we can introduce into English the new verb “to believe-formula” which corresponds exactly to the predicate symbol BF. For instance, “*Mike believes-formula the third formula written on the blackboard*” is intended to mean that Mike is in the BF relationship to the third formula written on the blackboard. Observe that “believes-formula” is an ordinary transitive verb, on a par with other verbs like “writes” or “constructs” that could be used to relate agents to formulae. Crucially, the verb “believes-formula” is *not* a complement-taking attitude verb like “believes” itself.

Second, I also assume that English sentences are allowed to refer to formulae by quoting them. We can therefore have sentences like

((5)) *Mike believes-formula ‘clever(voluptia)’.*

Barnden (1987b) explains in detail how the two assumptions can be avoided, but the simplified discussion they allow is sufficient for the purposes of the present paper.

The crucial observation now is that SYS’s rendering of sentence (5) is formula (4). This is because “believes-formula” corresponds exactly to BF, and the quotation in (4) denotes the same expression `clever(voluptia)` as is denoted by the quotation in sentence (5). So, SYS gives sentences (3) and (5) the same rendering. This conflation is *not* in itself what I object to. *But it causes another, implausible and objectionable, conflation — namely, of attitude reports that have those sentences as complements.* For instance, consider the attitude reports

((6)) *George believes that Mike believes that Voluptia is clever*

((7)) *George believes that Mike believes-formula ‘clever(voluptia)’.*

that respectively have (3) and (5) as complements. The Interpretation Rule is bound to give sentences (6) and (7) the same rendering, simply because their complements are given the same rendering. (Recall the Compositionality Effect above.) Since this latter rendering is (4), the common rendering of (6) and (7) is formed from the quotation of (4):

((8)) `BF(george, ‘BF(mike, ‘clever(voluptia)’))’).`

The trouble is that, in conformity with the intuitive view expressed above about the nature of the readings that our Interpretation Rule assumes, sentence (7) should surely be taken to convey that *George is thinking about Mike, the believes-formula relation and the formula clever(voluptia)*. But sentence (6) certainly does *not* convey that George is thinking about the believes-formula relation and the formula `clever(voluptia)`. After all, George’s could hardly be criticized for

never thinking about that relation and formula, or indeed about formulae at all! In sum, sentence (7) is likely to be false even when sentence (6) is true, although SYS conflates them and is thus committed to them having the same truth value.

Statement (7) does not follow from (6) even if George has views about propositional attitude theory, because his theoretical view of belief states might not be based on relationships to formulae. For instance, George might believe in the existence of propositions as abstract entities, and accordingly view Mike as being in some relation to a proposition rather than to a formula. He might even *deny* that Mike believes-formula `clever(voluptia)`. In that case, from the system's point of view George would be mistaken about the nature of belief.

We actually have to be quite careful in arguing that sentences (6) and (7) are likely to have different truth values, lest we seem to be relying on dubious assumptions. I do *not*, for instance, assume that (7) conveys that George has a visual image of the formula `clever(voluptia)`, or that he is thinking about the symbols it contains. I merely assume that he has in his mind *some* description of or name for the formula. Equally, the argument requires no assumption that the idea of George's that is being appealed to by the word *believes-formula* in sentence (7) explicitly involves a detailed characterization of the "believes-formula" relationship. All I do assume is that (7) conveys that George is entertaining a notion in which it is explicit that some person, known to us as Mike, is in some special relationship, known to us by the name "believes-formula", to some formula, known to us through the symbol string `clever(voluptia)`. I hold that this assumption, though weak, is sufficient warrant for us to say that sentences (6) and (7) can, and often will, have different truth values.

The undesirable conflation of these two sentences is a prime example of the problem this paper is about.

2.2: Other Cases

Although I hold that there is no pressing reason for objecting to the assumptions made above of the existence of the verb "believes-formula" and of the availability of formula quotation in English sentences, it is worth while knowing that one can avoid them. The avoidance is demonstrated in Barnden (1987b), by showing that there is an implausible entailment from the sentence (6) to the sentence "*George believes that Mike bears the relationship signified by the special predicate symbol to the formula at the top of the blackboard*", interpreted in a context in which "the special predicate symbol" is BF and the formula at the top of the blackboard is `clever(voluptia)`. The interpretation requires a partially de-re-like form of reading not covered by the Interpretation Rule above, but that is immaterial to the issue. Also, the implausible entailment is now only one-way, rather than two-way as in the above implausible conflation, but that is also immaterial.

Let us turn now to non-quotational approaches. Barnden (1987b) shows in detail how an implausible-entailment problem arises for the situation-semantics approach in Chapter 10 of Barwise & Perry (1983). This approach casts believers as being in a special relationship to “situations” rather than formulae, to speak very roughly. The problem is that the approach makes the sentence (6) bear implausible-entailment relationships to the sentences

((9)) *George believes that Mike believes-schema schema-of-Voluptia-being-clever*

((10)) *George believes that Mike believes-schema something.*

The word “schema” appeals to a generalization by Barwise & Perry of the notion of a situation. The verb “believes-schema” is analogous to the verb “believes-formula” used above. The notion of entailment that is relevant now is not truth-value-based as it was above, but is rather the notion of “strong consequence” described in Barwise & Perry (1983). It turns out that sentence (6) is a strong consequence (relative to a given “discourse situation” and “speaker connections”) of sentence (9), and that sentence (10) is a strong consequence of (6). These strong consequence relationships can be seen to violate the intended meanings of the three sentences (see Barnden 1987b). The basic observation is that George’s beliefs about belief need have nothing to do with his potential beliefs about Barwise & Perry’s schemata and other theoretical constructs (if indeed he were to be in a position to have any such beliefs!)

Hobbs (1985b) has proposed a first-order logic approach to belief representation that is based on terms denoting events and situations. It appears that this scheme suffers from the implausible conflation of sentence (6) with the sentence “*George believes that Mike believes-situation something which is a clever-situation for Voluptia*”, as long as the phrase “*which is a clever-situation for Voluptia*” is read as a de-re characterization of the situation Mike is believed to believe in. The demonstration of this effect will be published elsewhere.

Barnden (1986a, 1987b) also shows that an implausible-entailment problem arises for the neo-Fregean scheme expounded in Creary (1979) [see also Creary & Pollard 1985], which is an extension of the scheme of McCarthy (1979). In fact we get a problem of implausible conflation, much as we did in the quotational-logic case. Specifically, the sentences (6) and

((11)) *George believes that Mike believes-concept the clever-concept of the Voluptia-concept*

are given the same formal rendering. Here “believes-concept” is a verb analogous to “believes-formula” and “believes-schema”, and relates an agent to a proposition-like concept that the agent believes. The phrase *Voluptia-concept* refers to the standard concept of Voluptia, and the phrase *the clever-concept of* refers to the function taking any person concept to the concept of that person (as characterized by that concept) being clever. Since sentence (11) conveys that George is thinking

in terms of this somewhat arcane function, it will usually be false even when (6) is true. Hence the conflation of the two sentences is implausible.

The arcane functions like the clever-concept function are similar to the “characterizing functions” of Church’s Frege-inspired logic of sense and denotation [Church 1951, 1973, 1974]. Therefore, it is to be expected that this logic is in danger from Creary-like problems. However, I have not demonstrated that the problems cannot readily be circumvented in the logic by other means.

Finally, we should note that implausible entailment problems strongly analogous to those above do *not* arise for modal-logic approaches to propositional attitudes. It is instructive to see why this is so. Modal-logic schemes for interpreting attitude reports (in a de-dicto way) are, essentially, based on a rule just like the Interpretation Rule displayed earlier for quotational logics, except that no quotation is used. Thus, the sentence (3) would typically be rendered as

$$((12)) \quad B_{mike}(\text{clever}(\text{voluptia})).$$

where B_{mike} is a modal belief operator. The crucial difference from the quotational-logic and other schemes considered lies in the semantics of this operator as compared to the semantics of the predicate symbol BF in the quotational case (and analogous symbols in the other cases). Such predicate symbols are interpreted as denoting what we can call “belief relationships” between agents and belief objects (formulae, situations, concepts, propositions, or whatever). In contrast, the meaning of the modal operator B_{mike} is entirely implicit in the truth conditions for formulae that are applications of it. These truth conditions are classically given in terms of of “possible worlds” [see e.g. Chellas 1980]. In the example at hand, the formula $B_{mike}(\text{clever}(\text{voluptia}))$ is true, in a given possible world, if and only if the subformula $\text{clever}(\text{voluptia})$ is true in all possible worlds that are so-called doxastic alternatives (for Mike) to that world. Nowhere in this account is the situation described as Mike being in relation to some object.

The result of the lack of a relationship interpretation for B_{mike} is that we cannot follow a path like the one we took in the quotational case from sentence (3) to formula (4), and thence back to the alternative sentence (5) couched in language referring to the BF belief relationship. The closest analogy that I can see to this path is to go from sentence (3) to the formula (12), and thence back to something like the sentence: *The formula “clever(voluptia)” is true in all of Mike’s doxastic alternatives to the current world.* Although, relative to a given world (the “current” one), this sentence might be taken to be equivalent to formula (12), there is no reason to think that the sentence would actually be rendered as this formula, rather than as a formula that explicitly dealt with possible worlds. (This discussion is difficult to pursue, since proponents of modal-logic interpretations for attitude reports do not deal with sentences that talk explicitly about the possible worlds and other notions within the modal theory itself.)

Although, as we have seen, modal-logic approaches do not suffer from our implausible-entailment problem, they do suffer from other difficulties. These will be briefly discussed in Section 4.1.

2.3: Casting the Problem as Theory Imputation

In the quotational, situation-based, and concept-based systems above, we can intuitively view our implausible-entailment problem as a matter of the system imputing its own theoretical notions to ordinary agents like George. Effectively, the quotational-logic scheme takes sentence (6) to mean the same as (7), and the Creary scheme takes sentence (6) to mean the same as (11); the scheme thereby casts the outer believer in a nested-belief situation as thinking in terms of theoretical notions that form the basis of the scheme itself. In the quotational case, these notions are quite clearly arcane and unlikely to be entertained by an ordinary agent. The situation is more unclear in the Creary case, since one could claim that ordinary people do view each other as entertaining concepts. However, what the analysis (in Barnden 1986a, 1987b) shows is the more subtle point that the scheme casts an ordinary outer believer (George) as thinking in terms of a concept-to-concept function like the “clever-concept” function mentioned above. This function gives an arcane twist to commonsense ideas about concepts.

When the implausible entailment is not a conflation, it still appears that the problem can often be viewed intuitively as theory imputation. For instance, in the Barwise & Perry situation-semantics case, the fact that sentence (10) is a strong consequence of (6) means that the scheme requires an ordinary outer believer in a nested-attitude situation to have thoughts about the believes-schema relation (which is certainly an arcane and elaborate construct, even though it is founded on notions of situation that could with much justice be regarded as commonsensical). However, the theory-imputation view would have been inappropriate if we had had to rely only on the strong consequence from sentence (9) to (6). This entailment relationship is the wrong way round for a theory-imputation view to be natural; and schemes studied in the future may conceivably turn out to have only the wrong-way-round entailments. Thus, the implausible-entailment view of the problem remains primary, even though the theory-imputation view is perhaps more intuitively appealing.

2.4: Importance of the Implausible Entailment Problem

The reader may well be wondering whether our implausible entailment problem really has any importance in the long run, in view of the following questions:

- *Isn't the problem insignificant because of its reliance on very strange meta-level sentences?*
(Note that the strangeness of the sentences lies more fundamentally in their subject matter

rather than in the use of invented language, because such use can be avoided in all the cases of the problem portrayed above.)

- *Isn't the problem merely an artefact of the fine details of the particular rendering schemes considered above?*
- *Isn't there some easy way of avoiding the problem?*
- *Isn't it reasonable to impute one's own view of attitudes to other agents — so why shouldn't an AI system act like this?*

The answer to the first part of the last question is “yes”, and to the second part the answer is “no reason at all” — but there is a misleading ambiguity in the question that I take up at the end of the next section.

The answer to each of the first three questions appears to be “no”. Section 4 will justify a negative answer to the third question. On the second, part of the intent in discussing a variety of schemes was to underscore the fact that the sheer presence of the problem in a specific scheme is not particularly sensitive to the fine details of the scheme (although the details of the way the problem arises are of course very sensitive to the details of the scheme). The fundamental reason for the presence of the problem in the above schemes is the presence of symbols (like BF) standing for the various theoretical “belief relationships” (between agents and belief objects like formula, situations, concepts, and so on). Any scheme having such symbols is likely, therefore, to suffer from a version of the problem, unless great care is taken (as in the scheme of Barnden (1987a,b), to be discussed in Section 4.3).

We now proceed to justify a negative answer to the first question in the list.

(I) The problem has theoretical importance in being linked to certain other well-known, long-standing problems with propositional attitudes. This was noted in the Introduction, and we now briefly review the nature of some of these problems and their relationship to the implausible-entailment problem.

Referential Opacity, Hyper-Intensionality, and the Paradox of Analysis: In the sentence “Mike believes that Jim’s wife is clever” the phrase “Jim’s wife” cannot be replaced by a name for that person or another description of that person without possibly changing the truth value of the sentence, if the sentence is interpreted to mean that Mike is mentally entertaining an internal description analogous to the phrase “Jim’s wife”. (This is the ‘inner-scope’ or ‘de-dicto’ reading of the phrase.) Equally, in “Mike believes that Mary is clever” the name “Mary” cannot (*salva veritate*) be replaced by some description of Mary such as “Jim’s wife”. This non-replaceability of referring expressions is known as referential opacity. Referential opacity is just a special case of the hyper-intensionality of attitude contexts, whereby what is believed (or hoped for, desired, or

whatever) cannot be replaced by something equivalent to it or that follows from it. Thus, George's believing that a figure is an equilateral triangle does not imply that he believes that the figure is a triangle all of whose angles are 60 degrees. This is analogous to my observation that the second occurrence of the phrase "believes that" in the sentence "George believes that Mike believes that whales are fish" cannot *salva veritate* be given an explication in terms of an arcane theoretical account of belief. The paradox of analysis, as stated in Langford (1942) and Moore (1942), is strongly related to hyper-intensionality, is to the effect that, for example, George's believing that a figure is a cube does not imply that he believes that the figure is a cube with twelve edges, even though a cube necessarily has twelve edges as a result of the definition (analysis) of the concept of a cube.

Theory-Laden Terms: Partee (1982) has discussed the interpretations of the word "semantics" in the pair of sentences "Thomason believes that semantics is a branch of mathematics" and "Loar believes that semantics is a branch of psychology". She claims that these sentences are about the very nature of the concept of semantics, and that fixing on a particular meaning for the word "semantics" will cause misunderstanding of one of the sentences. It appears that phrases like "believes that" could be theory-laden in much the same sense, though this issue remains to be properly investigated.

(II) We should note that one's perception of the importance of the implausible entailments is affected by whether the interpretation scheme in question is actually being *used by an AI system or a person* to do natural language understanding, or whether it merely forms part of a *formal semantics* of natural language. In the latter case, if the scheme imposes an entailment that is unsound (i.e. there is a context in which the entailing sentence is true but the entailed one is false) then the scheme is, quite simply, wrong — *no matter how unusual or specialized the sentences in question are*. Thus, unsound entailments between sentences of forms (1) and (2) make entire classes of formal models demonstrably false.

Admittedly, our main concern is not with formal semantics but with the actual process of natural language understanding by AI systems (and possibly people). (Of course, a given natural language understanding system might be founded on some formal semantic theory, and it would be disturbing if that theory were wrong, even if only in a quite arcane way). Let us return therefore to AI systems.

(III) Certainly, in *most* situations an AI system will not need to consider arcane sentences like (7), (9), (10), or (11). Further, the use of formula (8) (or the corresponding formulae given in Barnden (1987b) in the case of the other schemes) as the rendering of sentence (6) may be heuristically adequate in the sense of *usually* leading to reasonable conclusions about George's likely behavior, plans or attitudes. However, AI systems will, presumably, ultimately be applied to *discourse about*

propositional-attitude research itself. Such discourse is of *practical*, though certainly *specialized*, importance, and we are engaging in it right now. In such discourse it may well be important for the system to be able to view George as believing that Mike believes that Voluptia is clever, without the system thereby being forced to view George as thinking in terms of formulae (etc.) — precisely because what might be at issue in the discourse itself is the question of what it is for George to have a belief about a belief. Imaginary discourses of this sort are presented in Barnden (1986a).

While I agree that it may be a long time before AI systems will be discoursing about propositional attitude research, I do not agree that this reduces the importance of the problem. Given the recognized centrality of propositional attitudes in the natural language processing enterprise, it is as well to get as good as possible a treatment of them as soon as possible, rather than waiting until other problems are solved and then perhaps having to partially redo fundamental aspects of those solutions in order to make them fit with a proper treatment of attitudes.

(IV) As we shall see in the next Section, the problem is actually a special case of a somewhat more general problem that is of clear practical importance and that does not involve meta-theoretical sentences. The special problem acts as a useful test case for any approach to the general problem.

(V) Finally, the problem has practical importance at a methodological level, in the sense that it has focussed the work of at least one researcher (me) in a way that has led to what I claim to a fruitful new approach to propositional attitudes (see Section 4.4 and Barnden 1988a,b). Although this new approach might conceivably have been arrived at independently, since it has independent justification, it was as a matter of fact arrived at through our implausible-entailment problem.²

In a sense, (V) may be the most important point of all. The implausible entailment problem is fundamentally to do with the commonsensicality or otherwise of outer agents' views of inner attitude states, in nested-attitude situations. The mentioned new approach to propositional attitudes is also fundamentally to do with this commonsensicality issue, which has hardly been addressed before in the propositional attitude literature.

² Point (V) may appear to be circular, because (i) a theme of the paper is that the importance of the implausible entailment problem makes the failure to notice it methodologically important, whereas (ii) point (V) is using methodological importance to bolster the importance of the problem. The apparent circularity is broken by the fact that the specific type of methodological importance appealed to in (ii) lies in the arrival at the new approach, which has the independent justification alluded to.

3: THE DIAGNOSIS

In this section I discuss the methodological tendencies that have led to the implausible-entailment problem of the last section being overlooked.³ The main tendencies that appear to be at fault are:

- (NO-NEST) Concentration on non-nested attitudes
- (ATOM) Using atomic symbols to stand for complex predicates
- (REF) Concentration on the problems of reference rather than of predication
- (NO-META) Neglecting thought and language *about* proposed theories
- (IMP) Theorists imputing their theories to other agents.

Clearly, tendencies (NO-NEST), (REF) and (NO-META) are types of problem reduction. (ATOM) is too, to some extent, as we will see.

Since our implausible-entailment problem is explicitly to do with nested attitudes, the importance of (NO-NEST) is obvious. However, (NO-NEST) also has a non-obvious role that we will discuss below. We should first realize that our implausible-entailment problem with nested attitudes is actually a special case of a more general implausible-entailment problem arising even with *non*-nested attitudes, but which has hardly been addressed in the propositional attitude subfield either. A non-nested example of the general problem is as follows.

Consider an AI system, SYS, that does not have a predicate symbol, such as **boiling**, that by itself stands for boiling. Instead, suppose SYS formally renders the sentence “*The water is boiling*” as the formula **very-hot(w) \wedge bubbling(w)**, where we assume for the sake of the example that something is boiling if and only if it is very hot and is bubbling, and **w** is a constant symbol denoting the body of water in question. (We use **w** for simplicity, to avoid having to deal properly with the issue of definite descriptions.) If SYS uses the quotational-logic Interpretation Rule adopted in the previous section, then it will render

((13)) *George believes that the water is boiling*

as the formula

((14)) **BF(george, ‘very-hot(w) \wedge bubbling(w)’)**.

But this is also the rendering of

((15)) *George believes that the water is very hot and is bubbling.*

³ As I stated earlier, the problem does not appear to have been addressed in the literature.

Now, this conflation of sentences (13) and (15) may not be very objectionable — certainly, it seems highly likely that if one is true then the other is — although I do argue in Barnden (1986a, 1987b) that there are unusual situations in which the sentences may differ in truth value. But we could strengthen the example to make SYS explicate boiling in some detailed scientific terms rather than in commonsense terms. In that case, the truth of sentence (13) would *usually* fail to imply the truth of the new, scientific sentence corresponding to (15).⁴

The issue here is very much the same as in the previous section — namely, *the explication of some predicate, within an attitude context, in non-commonsensical terms*. The predicate is “is boiling” in the present example, and was “believes that Voluptia is clever” in nested-attitude cases in the previous section. But the reason that this correspondence has not led to the *nested*-attitude version of the implausible-entailment problem being attacked is that the *non-nested* version has not been seen as a problem. And, the non-nested version has not been seen as a problem because it is common practice, in *propositional attitude* research, to render *ordinary* predicates, like “is boiling”, by means of atomic symbols like `boiling`. This is the tendency (ATOM) mentioned in the list above. Therefore, the issue of what to do about explications (commonsensical or not) of *ordinary* predicates within attitude contexts has simply not been given enough attention.⁵ Hence, there has been little opportunity for such explications to suggest, by extension, a consideration of explications of *attitudes* within attitude contexts.

Of course, in AI subfields other than the propositional attitude subfield, ordinary predicates have frequently been given explications, often of a supposedly commonsensical nature. The prime example of this is probably the work on Conceptual Dependency [Schank 1973]. Such work does often bring in the representation of attitudes, but the details of such representation are not the primary concern of the research. On the other hand, in the propositional attitude subfield, ordinary predicates have usually not been given explications, *precisely because* what is focussed on is the representation of *propositional attitudes*.

⁴ Naturally, SYS would also conflate the non-attitude sentences *The water is boiling* and *The water is very hot and bubbling*. My arguments are not directed at *this* conflation, given our presumption that water is boiling if and only if it is very hot and bubbling. Actually, in reality the two sentences would be interpreted in the context of the propositional attitude conjectured to be held by the speaker towards its content, but, by one of the expository simplifications mentioned in the Introduction, we are ignoring the embedding of all sentences in such implicit attitudes. The point is that we are not arguing against SYS representing states of boiling *it* believes in by means of the very-hot-and-bubbling explication.

⁵ This is not to say that existing propositional attitude representation schemes could not handle the explication of ordinary predicates within attitude contexts. They can, but the issue has not been perceived as having important ramifications.

We thus have the paradoxical situation that the very fact that what is focussed on is the representation of attitudes contributes to the neglect of our nested-attitude implausible-entailment problem. This effect occurs because that focus is one contributory factor to (ATOM), which is tantamount to a neglect of the explication of ordinary predicates. This in turn contributes, by extension, to a neglect of the effects of explicating attitudes, and hence to the neglect of our problem.

Thus, to summarize the argument so far, (ATOM) contributes to the neglect of our nested-attitude implausible-entailment problem. It does so by contributing to a neglect of the more general problem discussed above. (ATOM) is supported in part by a concentration on attitudes at the expense of ordinary predicates.

It is interesting that (NO-NEST) also contributes to the neglect of the more general problem (as well as making its obvious contribution to the special, nested-attitude problem.) This is because (NO-NEST) contributes to (ATOM) somewhat. For, if nested attitudes *had* been given more intense attention, the question of explication of attitudes within attitude contexts would have come more readily to light, and this would have suggested an examination of the explication, within attitude contexts, of predicates in general.

Thus, we have a two-way reinforcing effect, whereby the neglect of the more general problem contributes to the neglect of the special problem, and the prime reason (namely NO-NEST) for the neglect of the special problem contributes to the neglect of the general problem.

There are two further influences leading to (ATOM) (and thence to the neglect of our implausible-entailment problem), other than those already noted. One of these other influences is (REF), and the other is the nature of the most commonly used types of logic. Let us take the latter first. Propositional attitude research still predominantly uses, at most, some variant of ordinary first-order logic or of first-order modal logic, although more complex logics (e.g. λ -calculus) have also played an important role, notably in Church (1951, 1973, 1974) and Cresswell (1985). Now, one feature of first-order logic (modal or not) is that it makes it easier to have complex *referential* structures than to deploy complex, explicated *predicates*. For instance, objects can be referred to by complex terms such as

`boss-of(eldest-son-of(father-of(peter), president-of(PTA)))`.

Such an expression is a valid syntactic unit, and can act as an argument in any application of any function symbol or predicate symbol. On the other hand, suppose one wants to represent the complex predicate "... is short-sighted and is son of ...". If there is no predicate symbol standing for this predicate, then one is forced to *separately* explicate each different application of the predicate. For instance, to state that John is short-sighted and the son of Peter we need something like `short-sighted(john) \wedge son-of(john, peter)`, and to state that Bill is short-sighted and the

son of Mary we need $\text{short-sighted}(\text{bill}) \wedge \text{son-of}(\text{bill}, \text{mary})$. The point is that there is no common, valid syntactic unit shared by these formulae. They do share a common structure, but that structure is not in itself a valid syntactic unit. This makes it relatively awkward and unnatural to take an explicative approach to predicates.

Having said this, it is common practice in writings on logic to use notation like “ $P(\text{john}, \text{peter})$ ” and “ $P(\text{peter}, \text{mary})$ ” to stand for formulae such as the two above. The symbol P is not a symbol in the logic itself, but is rather a meta-theoretic symbol standing for a complex parametrized formula. This practice does somewhat lessen the effect I am pointing out, but I still conjecture that the effect is a significant contributor to (ATOM).

I stated also that (REF) contributes to (ATOM). The thrust of (REF) is that the consideration of how to treat noun phrases within attitude-report complements has been commonplace in propositional attitude research (usually under the heading of the de-dicto/de-re opposition or some similar concern), whereas the treatment of other aspects of complements has been minor by comparison. Why is this? One factor is possibly the observation just made about logic (which carries over to a large extent to related formalisms such as semantic networks or frames). There are also no doubt purely historical factors, which I will not attempt to trace. Another factor is probably the primacy that has been given, within treatises on natural language semantics, to the topic of reference to physical objects, and the similar primacy given in many epistemological theories to concepts of physical objects as opposed to either abstract objects or (complex) properties and relationships.

One special factor contributing much to (REF) is that there has been much debate about the treatment of natural language quantification within attitude-report complements — as in “*Mike believes that a spy is in the cupboard*”, where what is at issue is the reading of the quantificational phrase “*a spy*”. It is safe to say that quantification arises more naturally and frequently in noun phrases than elsewhere. This has encouraged a concentration on noun phrases.

The reason that (REF) contributes to (ATOM) is simply that researchers understandably tend to take the apparently-easiest possible approach to those problems to which they have given little attention. The apparently easiest way to deal with predicates is to use single predicate symbols for them.

Tendency (NO-META) is a result of the honorable practice of dealing with mundane natural language before going on to arcane language. According to this practice, consideration of arcane theory-laden sentences like (7), (9), (10), or (11), or the variants of them obtained by eliminating artificial language, is postponed. Unfortunately, this postponement does not, despite appearances, correspond to a natural division of the issue of propositional attitudes. The reason is given by the considerations of Section 2: arcane theory-laden sentences like those just listed are formally

rendered as expressions that are identical to (or in looser entailment relationships with) the formal renderings of perfectly *mundane* sentences like (6). Thus, there is a behind-the-scenes connection between those mundane sentences and the related theory-laden sentences, so that the naturalness of (NO-META)’s reduction of the natural language understanding problem by concentrating on the mundane sentences turns out to be illusory.

On tendency (IMP), I conjecture — and it is only a conjecture — that a researcher working on the representation of propositional attitudes can get so intensely embroiled in his/her own scheme that its underlying theory of propositional attitudes comes to seem quite commonsensical. Hence, the chances of the researcher noticing the fact that the scheme is leading to the imputation of this theory to ordinary agents are reduced.

There is a danger of such imputation occurring even if the researcher does not regard the scheme as commonsensical, and does give serious consideration to nested attitudes. The danger is that he/she will think about the view the outer agent (e.g. George) has of the inner attitudes *by looking at the world through that agent’s eyes*. This is tantamount to a mental identification not dissimilar from those one establishes in watching a film or reading a novel. A mental identification one establishes with another person involves, in effect, the creation of a new person some of whose mental characteristics are drawn from oneself and some of which are drawn from the other person. Therefore, some of each person’s mental characteristics are suppressed. In the sort of case under discussion, trouble arises if the theory of attitudes that survives the mental-identification process is the researcher’s rather than that of the outer agent in question.⁶

Finally, (IMP) returns us to the last question in the list at the start of Section 2.4: *Isn’t it reasonable to impute one’s own view of attitudes to other agents — so why shouldn’t an AI system act like this?* The answer to the first part is affirmative, as I stated earlier: and this seems to contravene what I have just said about (IMP). But the answer is only affirmative if the ambiguous phrase “one’s own view” is taken to mean “one’s own commonsense view, largely shared with other human agents”. But the answer is *negative* if the phrase “one’s own view” is taken to mean “one’s theoretical view on which a formal attitude representation scheme is based”. It is this latter reading that (IMP) is concerned with. I conjecture that it is possible for a propositional-attitude researcher to confuse the two versions of the question, and therefore to escape noticing the (IMP) danger.

Suppose for the sake of illustration that it is common for people in general to think of a situation in which some agent X believing something C as being a matter of X mentally entertaining a natural language statement of C. Let this view of belief be called the “naive internal speech view”

⁶ I have noticed a tendency of this nature in my own theorizing about attitude representation, but of course I am therefore perhaps guilty of imputing my own erroneous, imputational methods of thought to other researchers.

of belief. On this view, for Mike to believe that Voluptia is clever is for Mike to be saying to himself “Voluptia is clever” (say). Then it is certainly reasonable (and presumably typical) for a particular person Peter not only to subscribe to the theory himself but also to impute the theory to other believers: that is, to assume that when George believes that Mike believes that Voluptia is clever, George is taking the naive internal speech view of Mike’s belief. Thus, Peter assumes that George believes that Mike is saying “Voluptia is clever” to himself; moreover, because of Peter’s *own* use of the naive internal speech view, as applied to George, he actually assumes that George is saying to himself the sentence: “Mike is saying to himself ‘Voluptia is clever’ ”.

I do not object to *this* sort of imputation of a (hypothetically) commonsense “theory” of belief. What is objectionable is the analogous sort of imputation that arises in the case of a *non*-commonsensical theory of belief, such as the one underlying the quotational-logic approach of Section 2.1. (Notice that is very like the naive internal speech view, but using an abstract relationship to a formula rather than a mental entertainment of a natural language sentence). So, to go back to the fourth question at the start of Section 2.4, we can say that there is nothing wrong with an AI system imputing its own view of belief to other agents, if that view is a humanly-commonsensical one, because then the system will do no worse than a human agent would; but that there is something wrong if the AI system’s view is not humanly commonsensical.

Our affirmative answer to the “commonsensical” version of that question is linked to the new approach to attitude representation to be expounded at the end of the next Section.

4: THE CURE?

It is not the purpose of this paper to show in detail how the implausible-entailment problem can be cured. Therefore, in the present Section I concentrate on explaining why the problem appears to have no easy cure. I do this by rejecting modal logic and a well-known semantic-network scheme as cures, then rejecting an effective but uncomfortably complex cure instantiated in one of my own proposals, and finally outlining my current approach, which may be effective but which entails a major program of research running against the prevailing methodologies in the attitude representation subfields of AI and philosophy (but has much in common with some contemporary linguistics and philosophy).

Before going on, I should stress that I do not claim to have shown that there is no existing attitude-representation scheme that avoids our implausible-entailment problem while also being satisfactory in other respects. However, as I said earlier, the variegated nature of the schemes that I have shown to suffer from the problem, combined with various other difficulties that will appear below, does suggest that the “burden of proof” now lies with the designers of attitude

representation schemes. I claim that it is unlikely that a (non-modal) scheme will be free of the problem, unless explicit efforts are made to avoid it.

4.1: Why Not a Modal Cure?

We observed that the modal-logic approach to attitude representation was left unscathed by the arguments about implausible entailment.⁷ However, this does not mean that the modal approach does not have serious problems of its own, despite the genuine value it has had in providing a matrix within which to consider various fundamental issues. Some of the problems with the modal-logic approach have been discussed by other authors. Much attention has been given to the undesirable forms of logical consequence that classical approaches to the semantics of modal logic give rise to, the prime focus of study being a set of “logical omniscience” problems [e.g. Levesque 1984, Fagin & Halpern 1987, Drapkin & Perlis 1986]. Konolige (1986) shows that some recent attempts to remove these forms of undesirable consequence have undesirable effects of their own. Another problem is the difficulty of encompassing discourse about attitude objects (beliefs and so on) as objects, as in “These intentions of John’s are in conflict”, and in particular of dealing with quantification over attitude objects, as in “All of John’s beliefs are influenced by his paranoia” or “Mike hopes the opposite of everything Bill hopes” [see e.g. Asher & Kamp 1986, 1987; and the latter paper points out that the intensional logic of Montague (1970) corrects the problem but still suffers from undesirable forms of consequence]. Several other difficulties are noted in Perlis (1988).

One difficulty that is addressed relatively infrequently is that of encompassing certain types of reading of attitude reports that are in some sense intermediate between *de-re* and *de-dicto* readings [Barnden 1986a, Hellan 1981; also Kraut 1983, Saarinen 1981]. Consider for instance

((16)) *Mike believes that someone is a spy.*

Normally, only two readings for this are considered: the *de-re* one in which the existential quantification has outer scope, and the *de-dicto* one in which the existential quantification has inner scope. In the former case, the interpretation is that there is an actual, particular person whom Mike believes to be a spy, and that Mike is using in his belief a mental person-characterization which uniquely (in context) picks out that person. (The person-characterization could be, say, a perceptual description of a person, or a mental version of the description “Jim’s wife”.) In the *de-dicto* case, the interpretation is that Mike merely believes in the existence of a spy (much as if his belief were a matter of him saying to himself “There is a spy”). However, there is a third, “middle”

⁷ Actually, Barnden (1986a) points out that modal logics are susceptible to a form of nested-attitude implausible-entailment problem if they explicate some attitudes in terms of others — e.g. knowledge in terms of truth and justified belief.

reading which is like the de-re in casting Mike as entertaining a mental person-characterization (again unspecified); however, the reading is unlike the de-re in *not* assuming that this characterization picks out a particular person in the real world. For instance, it may be that Mike believes that Jim’s wife is a spy, but then the reading allows it to be the case that Jim does not have a wife. The lack of an assumption of a real-world referent makes the reading a little like the de-dicto. Although there are ways of approaching the problem of middle readings within modal logic (see above references), they involve major extensions of the logical tools typically assumed in work on propositional attitudes.

In case it should be argued that middle readings are of somewhat secondary importance, I should point out that they are raised to the status of *preferred* reading by sentences like the following, which are, I believe, of fundamental importance but which are hardly ever discussed:

((17)) *Mike believes that someone is a spy, but he has only a hazy idea of who it might be.*

This sentence is surely to be given some sort of middle reading, since there is no reference to a particular person in the real world (or indeed in Mike’s belief world), yet clearly Mike’s belief does involve a lot more than just an existential quantification (as used in the de-dicto case).

Importantly, this sentence is also an example of *explicit reference to psychological entities* in an attitude report. Modal logics are not directed at such cases. This is not to say that it is impossible to encompass them in a modal-logic framework: one suggestion⁸ is to use special modal operators. For example, the formula $\text{Hazy}(\text{Mike}, \text{John})$ could be used to state that Mike has a hazy idea of John. However, such operators do not fit naturally with possible-world semantics of modal operators. Even if this is not regarded as a major problem, psychological statements are much more naturally dealt with in an approach to propositional attitudes that *already* rests on terms that denote psychological entities such as “concepts”. Neo-Fregean approaches like those of Creary (1979) can easily and naturally cope with sentences like (17); in such a system the middle readings are easily expressed (using, in fact, slightly simpler formulae than are required for de-re readings), and the expression of haziness merely requires an ordinary predicate symbol **hazy** that applies to concepts.

Statements like (17) are quite mundane and by no means rare; and they highlight the fact that human communication and thought about the mental states of human beings is by no means confined to the use of verbs such as “believes”. This obvious point has been ignored in the bulk of research (in philosophy and AI) on propositional attitudes, yet it has fundamental consequences, which will appear in Section 4.4.

The neglect we have just asserted is worthy of being named:

⁸ by an anonymous referee

(NO-PSYCH) *Neglect of the need to deal with attitude reports that make explicit commonsensical, psychological statements about the attitudes reported.*⁹

It is a methodological tendency on a par with those listed at the beginning of Section 3 (although its contribution to the neglect of our implausible-entailment problem is unclear). Indeed, it is closely allied to tendency (NO-META). Tendency (NO-META) is a neglect of thought and language about the contents of propositional-attitude *theoreticians'* theories about attitudes; (NO-PSYCH) is a neglect of thought and language about the contents of *commonsensical* theories about attitudes. The seriousness of (NO-PSYCH) is by no means lessened by the possible incorrectness of commonsense-psychology theories — the fact that they are used in ordinary communication requires their content to be representable. The point is perhaps made more clearly by use of sentences like “*Susan believes that Mike fears that something terrible is going to happen, but that he has only a hazy idea of what it is*”. Here it is Susan who is casting Mike as having a hazy idea as part of his fear attitude, so that the *actual* correctness or otherwise of the commonsense psychology underlying the casting mitigates not at all against the need to represent the casting.

4.2: Why Not a Semantic-Network Cure?

Most of this paper explicitly addresses only logic-based representation schemes. However, semantic network schemes are so closely allied to logic-based schemes that similar issues are likely to arise. Probably the best-developed semantic network scheme that is explicitly concerned with propositional attitudes is the SNePS scheme [Maida & Shapiro 1982, Rapaport 1986, Shapiro & Rapaport 1986]. The question then arises of whether it suffers from our implausible-entailment problem, and, if not, whether it should therefore be regarded as a good cure.

In Barnden (1986a), I argue that SNePS does suffer from a form of implausible entailment. Nevertheless, these entailments are less objectionable (i.e. more plausible) than are those studied in Section 2; also, they arise only in rather special conditions (a point not made clear in Barnden 1986a). Therefore, the scheme could still be put forward as a candidate cure. However, I resist this move. In Barnden (1986b) I show that the authors of the scheme have not paid adequate attention to the need for differential levels of intensionality of nodes in the networks, thereby making certain

⁹ I refrain from using the adjective “folk-psychological” here, because it has sometimes been taken in a pejorative sense that I do not intend, and suggests contemporary philosophical issues that are beyond the scope of this paper. My emphasis on commonsense psychology (folk psychology) should not be construed as a claim that a mature scientific account of cognition should be based on folk psychological accounts (see Stich 1983 for a discussion of this issue). Rather, folk psychology is important to our concerns in this paper because of *people’s commonsensical (folk-psychological) views* of people — irrespective of the possible incorrectness or unscientific nature of those views.

types of statement awkward if not impossible to express. In essence, there appears to be no provision for making statements about intensions (concepts) themselves, except in certain very limited ways, despite the fact that intensions are what the network nodes represent. As a result, the scheme is not well suited to dealing with the “hazy idea” sentence (17) that was a problem for modal schemes, or to quantifying over intensions (cf. another criticism levelled at modal schemes).

4.3: An Effective but Expensive Cure

One could try to avoid the type of implausible entailments in question by trying to avoid all *unsound* entailments between sentences of forms (1) and (2). By a sound entailment I mean one where the truth (in a given context) of the entailing sentence absolutely guarantees the truth (in that context) of the entailed sentence. A sound entailment must be plausible, so that eliminating unsound entailments ensures that implausible ones are eliminated too.

There is in fact a way of eliminating unsound entailments of the sort in question. This can be done by including certain special facilities in an attitude-representation scheme — and using them very carefully. The main (but not the only) thing that appears to be needed is a facility for constructing or defining complex predicates to supplement the basic set of predicates provided in the scheme. Barnden (1986a) suggests how the unsound entailments can be eliminated in an extended quotational scheme that includes λ -abstraction as a predicate-construction facility; and Barnden (1987a,b) shows how it can be done in an extended quotational scheme that includes “expression templates” (incomplete expressions that can be filled in to produce proper expressions).¹⁰ It is probable that situation-based and concept-based schemes could be extended in analogous ways. It is likely also that modified versions of the schemes proposed by Church (1951, 1973, 1974), Cresswell (1985), Parsons (1980) and others, which include λ -abstraction or other forms of predicate construction/definition, could be deployed so as to avoid our unsound entailments.

However, it appears that these entailments can only be avoided in such schemes at the cost of considerable complexity in the representational expressions. This is suggested by the complexity of the formal renderings of attitude reports in Barnden (1987b). The complexity arises mainly from the nature of the representational task, not from idiosyncratic features of the representation scheme, so that it is to be expected that the other modified schemes alluded to in the previous paragraph would run into the same degree of complication.

¹⁰ The scheme is not an ad-hoc response to our implausible entailment problem, as one commentator has claimed. The scheme is geared also towards supporting a rich space of possible reading types for attitude reports, generalizing from the usual de-dicto/de-re pair of reading types, and to dealing with sentences like (17).

4.4: A Commonsense-Psychology Cure

The previous subsections suggest that attempts to adopt schemes that totally avoid *unsound* entailments arising in the way explored in Section 2 lead to unwelcome complication. But there is a compromise strategy that one can take, and that has independent justification in any case:

allow a range of commonsensical and potentially plausible, though still unsound, entailments, but ensure that the system can choose ones that are plausible in context, and can change its mind about which entailments to effect.

This strategy takes note of the fact that what is objectionable about the unsound entailments of Section 2 is not so much their sheer unsoundness, but more

- the *implausibility* of what is entailed about the outer attitude-holder's state of mind (e.g. the implausibility of George regarding Mike as being in some relationship to a formula), and
- the *inescapability* of those entailments.

Thus, we might be able to live with entailments that were unsound but, nevertheless, ascribed plausible states of mind to outer attitude-holders, and which, anyway, the system could escape by replacing them with different entailments if they turn out to be misleading.

As an example leading to the type of plausible ascription I have in mind, consider the sentence pair

((18)) *George hopes that Mike will come to realize that Voluptia is clever.*

((19)) *But George is afraid that the idea of Voluptia being clever is having an uphill struggle against Mike's sexist way of thinking.*

The second sentence brings in the metaphor of Mike's mind as a BATTLEGROUND in which forces (corresponding to ideas, habits, etc.) engage in struggles. Notice that there is a scope ambiguity with respect to this metaphor. The sentence could be taken to mean that George is himself thinking in terms of of the metaphor of an idea battling with a habit (the inner-scope reading of the metaphor). Alternatively, we could assume merely that the speaker is using the metaphor to describe Mike's hoped-for mental state (the outer-scope reading). I hold that the inner-scope reading is very likely in practice to be the more appropriate one, and that therefore we must be seriously concerned with treating it properly, although the outer-scope reading might sometimes be preferred.

I claim that the system could only interpret the two sentences *coherently*, assuming an inner-scope approach to the second, by giving the "realize" notion in the first sentence a *metaphorical*

explication in terms of the MIND-AS-BATTLEGROUND metaphor appealed to in the second sentence. Thus, the first sentence should be interpreted as if it had been something like

((20)) *But George hopes that the idea of Voluptia being clever will gain dominance in the battleground of Mike’s mind.*

That is, we are in effect taking the first sentence to entail this new version. The entailment is unsound, in the sense of not being definitely correct; yet, it is a plausible entailment. The strategy I am proposing, therefore, is to adopt unsound but plausible entailments of this commonsense sort as a replacement for the non-commonsensical sort of unsound entailments uncovered in Section 2.

Only by interpreting (18) as if it had been some such sentence as (20) can the system respect the plausible, inner-scope interpretation of (19) that George is actually thinking about Mike by means of the MIND-AS-BATTLEGROUND metaphor. And, hence, only in this way can the system be in a position to draw the full plausible implications of the discourse. One such implication is that there is a very good chance that George’s hope will not be satisfied (though this is already hinted by the “But”), and, further, that the reason for that possible failure is Mike’s sexist thinking. Another plausible implication is, on the other hand, that “part of” Mike *is* trying hard to achieve the realization.

It might be argued that instead of *introducing* a metaphor into the interpretation of the first sentence, the system should seek to *eliminate* the metaphorical aspect of the second sentence, by giving it a literal paraphrase. This argument, however, faces the well-known difficulty of eliminating metaphors for mental states/processes [see e.g. Fainsilber & Ortony 1987], and simply fails to respect the observation that *George* himself is conjectured to be thinking in terms of the metaphor.

The MIND-AS-BATTLEGROUND metaphor in the example is just one of the commonsensical ways in which people think about attitudes and which are apparent in ordinary discourse. Further, many of these ways are based on prevalent metaphors of mind. These observations lead to the main claim of this Section:

a system capable of interpreting attitude reports in realistic discourse should be able to cast out attitude-holders (e.g. George) in nested situations as thinking about the inner attitudes in any one of the variety of commonsensical, and largely-metaphorical, ways that discourse suggests people have for thinking about attitudes.

Commonsense, metaphorical views of mind that people entertain have received some close attention [see. e.g.: Johnson 1987; Lakoff & Johnson 1980; Larsen 1987; Reddy 1979; Sweetser, forthcoming; Tomlinson 1986]. Some metaphors for mind are closely related to common metaphors for arguments and reasoning processes — cf. the ARGUMENT-AS-WAR, ARGUMENT-AS-BUILDING,

and ARGUMENT-AS-JOURNEY metaphors [Johnson 1987; Lakoff & Johnson 1980]. Metaphorical relationships between understanding and seeing have been studied [e.g. Johnson 1987: p.108; Sweetser, forthcoming]. The metaphors used are also linked to commonsense metaphors for language, particularly the famous CONDUIT metaphor for communication [Reddy 1979] and bodily-force-based metaphors for moral action, alethic modalities (necessity, possibility), and reasoning processes [Johnson 1987: p.16, 48ff; Sweetser, forthcoming; Talmy 1985, 1988]. There are several variants of the MIND-AS-CONTAINER metaphor [Lakoff & Johnson 1980, Lakoff 1987, Johnson 1987] involving views of the mind or head as a container of ideas, internal speech, visual images, etc. This metaphor is related to the CONDUIT metaphor. Metaphors for ideas as plants, creatures to be hunted, stuff to be mined from the ground, food items, or archaeological finds have also been discussed [Larsen 1987, Tomlinson 1986].

Metaphorical explications of attitudes are not the only possible ones. For instance, Mike's realization might under other circumstances have been appropriately explicated as a disposition to act as if Voluptia were clever. Such an explication is perhaps partially metaphorical, but in any case it is not my purpose to argue that *all* commonsense explications of attitudes are metaphorical.

In sum, returning to sentence (18), a different context might have suggested the use of some plausible, commonsensical way of explicating Mike's realization, other than by means of the battleground metaphor. The system should ideally be able to detect and respect any available contextual pointers towards particular commonsensical explications. The system should be able to change its mind about what explication to use, if evidence accrues against its current choice. Further, the system should ideally be able to learn to use an unfamiliar mode of explication (an unfamiliar metaphor, for instance).

What if context does not suggest any particular way of explicating Mike's realization? In such a case I propose that a default explication be used, based on the MIND-AS-CONTAINER metaphor. I also propose that this explication be used for outermost attitudes. For example, to represent Mike's realizing that Voluptia is clever, as an outermost attitude, the system would use the formula

$$\text{contains}(\text{realization-part-of}(\text{mind-of}(\text{mike})), \sigma(\text{clever}(\text{voluptia}))).$$

Here σ is a modal (non-truth-functional), term-forming operator that delivers an intension (idea, concept) corresponding to an expression in the system's representation scheme.¹¹

¹¹ The formula shown is a simplification of what I propose, and should be buttressed by "counterpart" relationships analogous to those of Fauconnier (1985). See Barnden (1988b) for a fuller exposition.

Going back now to sentence (18), if context does not suggest a particular explication for Mike’s realization, the system will use the MIND-AS-CONTAINER explication for it (as well as for George’s state of hope, as that is the outermost attitude). In this way we get:

contains(hope-part-of(mind-of(george)),
 $\sigma(\text{contains}(\text{realization-part-of}(\text{mind-of}(\text{mike})), \sigma(\text{clever}(\text{voluptia}))))).$

When the system is just using the default MIND-AS-CONTAINER explications, as in this formula, the representational expressions are similar to the structures used in existing mental-space approaches to attitudes [Ballim 1987; Dinsmore 1987; Fauconnier 1985; Johnson-Laird 1983; Maida 1984; Wilks & Ballim 1987; Wilks, Ballim & Barnden 1988]. Therefore, not much extra complexity is being caused by my *wholesale* application of the idea that attitudes be commonsensically explicated.

The formula shown is also the preliminary representation the system would give to sentence (18) before encountering (19). On encountering the latter, it could revise its rendering of the first sentence in the light of the new information about George’s view of Mike, and derive the following replacement:

contains(hope-part-of(mind-of(george)),
 $\sigma(\text{dominate}(\sigma(\text{clever}(\text{voluptia})), \text{battleground-in}(\text{mind-of}(\text{mike}))))).$

The thesis is that *this* is the sort of representation that will be heuristically desirable in order to establish a proper degree of discourse coherence.

Notice that the effect our considerations are having on attitude representation comes down to an effect on the *content* of representational expressions. The arguments have not displayed a reason for thinking that traditional *general styles* of representation — e.g. quotational, concept-based, situation-based — are inadequate; rather, what is at issue is more the way they are applied. Indeed, the question of the choice of a general style becomes somewhat secondary: what is of primary importance is the means whereby a system can detect suggestions of commonsensical explications of attitudes, assess them, and apply them appropriately in the interpretation of attitude reports. It is this need to apply commonsense explications (or else indulge in the complex measures of Section 4.3) that constitutes the fundamental effect, advertised in this paper’s Abstract and Introduction, that a proper consideration of nested attitudes has on methods of attitude representation.

Not only are the general situation-based and concept-based styles not impugned, but we should also be happy to allow the system’s repertoire of commonsense explications to include some that are based on (commonsensical) notions of situation and concept; indeed, the example above relies on the σ operator that delivers ideas (concepts, intensions). Our criticisms in Section 2 were

of non-commonsensual, arcane aspects of the precise ways those concepts have been applied in attitude representation schemes, not of the notions in themselves.

It is not the purpose of this paper to give a detailed solution to the problems addressed in it, and in Barnden (1988b) I give more detail on the formal representation scheme I am proposing on the lines sketched above. I also place the argument within a more general framework concerning polysemous terms within attitude contexts. In Barnden (1988a) I place the argument in a considerably more general context to do with the interactions between attitude representation and commonsense reasoning. My approach is also related to the view of Green (1985) that attitude reports act merely as a rough but heuristically reasonable guide to thoughts.

5: PROPHYLAXIS AND MORALS

We have seen that the methodological tendencies discussed in Section 3 have led to the neglect of an implausible-entailment problem in the representation of nested attitudes. This problem has practical and theoretical importance, and is difficult to circumvent. My response to this difficulty is to say: “Well then, let’s allow unsound entailments, but let’s make sure that the system can base them on a set of plausible, commonsensual, and typically metaphorical, explications of attitudes.” The need to use such explications is a fundamental effect on attitude representation exerted by our consideration of nested attitudes.

Two sets of morals or lessons can be drawn, one for the propositional attitude subfield of AI, and another, much more speculative, for AI as a whole. The first set is substantially a matter of recommending the opposite of the tendencies listed at the beginning of Section 3 and of the related tendency (NO-PSYCH) displayed in Section 4.1. Two further injunctions, (FOLK) and (EMPIR), are included at the end of the following list.

(NEST) Pay attention to how nested attitudes are to be represented, before becoming too deeply committed to a particular way of representing non-nested attitudes. In particular, beware of the types of implausible entailment discussed in Section 2.

(NO-IMP) Beware of coming to regard your own theory of attitudes as being commonsensual, and of failing to discard it when putting yourself in the shoes of outer agents in nested-attitude situations.

(PRED) Give as much attention to the question of encompassing complex predications as to problems raised by referential constructs.¹²

¹² This moral is meant to include the opposite of tendency (ATOM) as well as that of (REF).

(META) Do not neglect thought and language *about* proposed theories of attitudes (recall the meta-level sentences relied on in Section 2)

(PSYCH) Do not neglect thought and language about the psychological nature of agents' attitudes (recall the "hazy idea" example of Section 4.1). And notice that such thought and language will rely on *commonsense* psychology, resting typically on a wide range of metaphors of mind.

(FOLK) Do not be seduced by the siren of formal simplicity into ignoring the charms of commonsense psychology (folk psychology).

(EMPIR) Propositional attitude research should occasionally leave its logical armchair to visit the psychological laboratory.

All of these except possibly (NO-IMP) are directed against forms of inappropriate problem reduction, and perhaps (NO-IMP) can be viewed in this light as well.

(FOLK) and (EMPIR) need some explanation. They are warnings against treating the problem of attitude representation as if it were an exercise in setting up elegant axiom sets (that owe more to traditions of formal logic than to the exigencies of the types of discourse and thought process in which attitudes actually figure). Propositional attitudes are just special elements in a complex array of views that people hold about the workings of people's minds, and therefore the truly adequate representation of attitudes is not likely to be achieved except by paying attention to those views. Specifically, if the commonsense-explication approach to attitudes recommended in Section 4.4 is valid, then we should look forward to a greater dependence of attitude research than heretofore on empirical studies in psychology — the point of (EMPIR) — or on systematic surveys of attitude-talk in real discourse. What is needed is more work aimed at elucidating just what are the ways, metaphorical ones especially, in which people think about other people's minds and attitudes.

As for lessons for AI generally, we can generalize from the specialized ones (other than NO-IMP and PRED) to get:

- When an issue has a natural meta-level, always consider the meta-level aspects before becoming too deeply committed to an approach at the base level.
- In particular, remember that the contents of both AI theories and commonsense theories are themselves important domains of thought and communication.
- And when commonsense theories are relevant, empirical studies might have something to say about them.

The validity and importance of these general guidelines remains to be determined — they are merely put forward as points of discussion.

6: CONCLUSION

We have considered an overlooked problem in the representation of propositional attitudes. The problem arises in a wide variety of (non-modal) schemes, and has important practical, theoretical, and methodological ramifications. The problem can roughly be viewed as that of a cognitive system that uses such a scheme implausibly imputing its own arcane (non-commonsensical), theoretical view of attitudes to ordinary people.

The focus of the paper has been not on the problem's fine detail, which has been reported elsewhere, but rather on the methodological practices that have led to its being overlooked. A set of lessons for propositional attitude research have thereby been presented. As a matter of historical fact, one of the methodological products of considering the problem has been the notion that in designing propositional attitude representation schemes we should attend to the variety of commonsensical views that people actually hold of mental matters, propositional attitudes in particular. It is *these* views, not arcane theoretical ones, that should be imputed to ordinary human attitude-holders. Since the views tend to be metaphorical, the approach leads to an integration of the study of attitudes with that of metaphor (although we do not have to discount the possibility of non-metaphorical commonsense views of attitudes).

This new approach amounts to a paradigm shift in the way that attitudes are normally treated in artificial intelligence.¹³ The shift is away from the prevailing formalist approach in AI (and philosophy) and towards an alliance with the concerns of the more cognitively-oriented linguists and philosophers. The work of Lakoff (1987), Johnson (1987) and Sweetser (1987 and forthcoming) on metaphor, including metaphors for mental states and processes and related matters, is especially relevant. The discourse coherence considerations that were appealed to are related to the work on discourse and metaphor within AI by Hobbs (1983a, 1983b, 1985a) and Carbonell (1980, 1982a,b).

There is no claim that the implausible-entailment problem is pernicious in the sense that, say, the Liar Paradox and related difficulties for belief representation are pernicious in actually leading to a danger of internal inconsistency within a logic [see e.g. Asher & Kamp 1987; Halpern 1986; Montague 1974; Thomason 1980; Perlis 1988]. However, this does not mean that the problem is not important, but just that the importance is of a different sort. The importance is primarily to do with *content and ontology*: the question of exactly *what* sorts of state of the world should be represented in propositional attitude representations, and, only secondarily, the question of what precise logical or other representational *language* should be available to ensure adequate adequate

¹³ But note that Maida (1986) presents an account of belief representation and reasoning in which an agent uses its own beliefs and reasoning abilities as an analogy for those of other agents. Maida's dependence on analogy is reminiscent of my dependence on metaphor.

representation of such states. (This point is underscored by the fact that the quotational-logic approach, for one, *can* be followed in such a way that the implausible entailment problem does not arise. What is needed is the representation of different sorts of objects and relations — i.e. different sorts of content — from those represented in standard quotational-logic approaches.) By contrast, classical problems such as the Liar Paradox are important primarily with regard to devising the representational language necessary to express a type of content that to a large extent is already *agreed upon*. In the case of the Liar Paradox for natural language sentences, the problem is to be able to express, without inconsistency, the meaning of sentences like “This sentence is false”. No one can seriously doubt that the meaning of such sentences must be accounted for. (There is an analogous problem of allowing formal representational structures that state their own falsity.) This is not to say that classical problems do not bring in issues of content and ontology, or that my approach to attitudes does not bring in issues of basic representational language (great care is still needed within that approach to avoid implausible entailments analogous to those mentioned in this paper). Rather, it is a question of relative emphasis.

One worry about the new approach that has been expressed to me is that it seems to require that the problem of how to deal computationally with metaphor needs to be solved first. Although it would certainly have been helpful if that problem had already been solved, the worry is not as serious as it might seem. First, it seems to me that headway can be made within the new approach by using preliminary, simplified computational approaches to metaphor, which is the most that we have at present. The formal representation scheme based on the approach and sketched in Barnden (1988b) is being incorporated in an extension of the Meta5 program of Fass (1987). This program establishes internal coherence of natural language sentences, a major focus being the treatment of metaphor and metonymy. The resulting program will enable useful research on the new approach. In particular, initial investigations using the program will not attempt to account for the learning of new metaphors by the system, although of course this is an important part of the total picture (and, in line with this paper’s concern with the dangers of problem reduction, there is no guarantee that the postponement of the learning issue is safe.) Second, attitude representation is one perfectly good application area on which to concentrate on in the very enterprise of developing a general account of metaphor. Indeed, because of its high level of abstraction and the rich web of abstract concepts and categories into which it is integrated [cf. the cited work of Lakoff, Johnson and Sweetser], attitude representation may be an especially fruitful locus for metaphor research. Last, the worry ignores the point that proper accounts of metaphor should have a fundamental partial dependence on accounts of propositional attitudes: metaphor has much to do with *people’s views* of things, not with how things “really are”. To make this observation more precise, an account of the fact that people can entertain partially conflicting metaphorical views of the same subject matter should presumably be integrated with an account of the more general fact that people can

entertain inconsistent beliefs, hopes, and so on. The dependence of metaphor on belief is given flesh in a preliminary way in recent work by Wilks, Ballim & Barnden (1988).

An important issue that I have not addressed is the extent to which my considerations should be refined to account for different degrees of conscious awareness people may have of the commonsense views they are taking of attitudes. In the MIND-AS-BATTLEGROUND example we considered (see sentences (18) and (19)), in which we assumed that George himself was thinking in terms of that metaphor, to what extent do we need to consider whether George is conscious of doing so or not? Certainly, I did not intend any presumption of consciousness — I claim that my analysis applies even if we take George to be unconsciously viewing Mike in terms of the metaphor. However, the issue does need to be looked at more closely.

Finally, a caveat: nothing I have said suggests that progress on attitude representation can *only* be made by adopting my approach or by heeding the “morals” drawn in the previous section. I claim that a *full* account will ultimately need to do so, and that the morals should not be far beneath one’s consciousness, but this does not mean that progress on specialized issues cannot be made using existing techniques.

ACKNOWLEDGMENTS

I am grateful for encouragement, stimulation, and constructive criticism from Afzal Ballim, J. Michael Dunn, Dan Fass, Sylvia Candelaria del Ram, and Yorick Wilks. The comments of the anonymous reviewers led to significant improvements in the paper. One reviewer in particular supplied an exceptionally detailed and thought-provoking review.

REFERENCES

- Asher, N. & Kamp, J. (1986). The Knower’s Paradox and representational theories of attitudes. In Halpern (1986).
- Asher, N. & Kamp, J. (1987). Self-reference, attitudes and paradox. In *Procs. 1986 Conf. on Property Theory*, Univ. of Massachusetts, Amherst.
- Ballim, A. (1987). The subjective ascription of belief to agents. In J. Hallam & C. Mellish (Eds), *Advances in artificial intelligence*. Chichester, UK: Ellis Horwood.
- Barnden, J. A. (1983). Intensions as such: an outline. *Procs. 8th Int. Joint Conf. on Artificial Intelligence*, Karlsruhe, W. Germany.
- Barnden, J.A. (1986a). Imputations and explications: representational problems in treatments of propositional attitudes. *Cognitive Science*, 10 (3), 319–364.

- Barnden, J.A. (1986b). A viewpoint distinction in the representation of propositional attitudes. In *Proceedings of the 5th Nat. Conf. on Artificial Intelligence (AAAI86)*, Philadelphia, Penn.
- Barnden, J. A. (1987a). Interpreting propositional attitude reports: towards greater freedom and control. In B. du Boulay, D. Hogg & L. Steels (Eds), *Advances in artificial intelligence – II*, Elsevier (North–Holland): Amsterdam.
- Barnden, J.A. (1987b). Avoiding some unwarranted entailments among nested attitude reports. *Memoranda in Computer and Cognitive Science*, No. MCCS–87–113, Computing Research Laboratory, New Mexico State University.
- Barnden, J.A. (1988a). Propositional attitudes, commonsense reasoning, and metaphor. In *Procs. 10th Annual Conf. of the Cognitive Science Soc.*, Hillsdale, N.J.: Lawrence Erlbaum, 1988.
- Barnden, J.A. (1988b). Propositional attitudes, polysemy, and metaphor. *Memoranda in Computer and Cognitive Science*, No. MCCS–88–139, Computing Research Laboratory, New Mexico State University.
- Barwise, J. & Perry, J. (1983). *Situations and attitudes*. Cambridge, Mass.: MIT Press.
- Carbonell, J.G. (1980). Metaphor — a key to extensible semantic analysis. In *Procs. 18th Annual Meeting of the Association for Computational Linguistics*.
- Carbonell, J.G. (1982a). Towards a computational model of metaphor in commonsense reasoning. In *Procs. 4th Annual Conference of the Cognitive Science Society*.
- Carbonell, J.G. (1982b). Metaphor: an inescapable phenomenon in natural-language comprehension. In W. Lehnert & M. Ringle (eds), *Strategies for natural language processing*. Hillsdale, N.J.: Lawrence Erlbaum.
- Chellas, B.F. (1980). *Modal logic*. Cambridge University Press.
- Church, A. (1951). A formulation of the logic of sense and denotation. In P. Henle (Ed.), *Structure, method and meaning: essays in honor of Henry M. Sheffer*. New York: Liberal Arts Press.
- Church, A. (1973). Outline of a revised formulation of the logic of sense and denotation (part I). *Noûs*, 7, 24–33.
- Church, A. (1974). Outline of a revised formulation of the logic of sense and denotation (part II). *Noûs*, 8, 135–156.
- Creary, L. G. (1979). Propositional attitudes: Fregean representation and simulative reasoning. *Procs. 6th. Int. Joint Conf. on Artificial Intelligence*, Tokyo.
- Creary, L.G. & Pollard, C.J. (1985). A computational semantics for natural language. *Procs. 23rd Ann. Meeting of the Association for Computational Linguistics*, Univ. of Chicago.

- Cresswell, M.J. (1985). *Structured meanings: the semantics of propositional attitudes*. MIT Press: Cambridge, Mass.
- Dinsmore, J. (1987). Mental spaces from a functional perspective. *Cognitive Science*, 11 (1), 1–21.
- Drapkin, J. and Perlis, D. (1986). Step logics: an alternative approach to limited reasoning. In *Procs. 7th Int. European Conf. on Artificial Intelligence, Vol. II*, Brighton, U.K.
- Fagin, R. & Halpern, Y.J. (1987). Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34 (1), 39–76.
- Fainsilber, L. & Ortony, A. (1987). Metaphorical uses of language in the expression of emotions. *Metaphor and Symbolic Activity*, 2 (4), 239–250.
- Fass, D. C. (1987). Collative Semantics: an overview of the current Meta5 program. *Memoranda in Computer and Cognitive Science*, No. MCCS-87-112, Computing Research Laboratory, New Mexico State University.
- Fauconnier, G. (1985) *Mental spaces: aspects of meaning construction in natural language*. MIT Press: Cambridge, Mass.
- Green, K. (1985). Is a logic for belief sentences possible? *Phil. Studies*, 47, 29–55.
- Haas, A.R. (1986). A syntactic theory of belief and action. *Artificial Intelligence*, 28, 245–292.
- Halpern, J. Y. (ed.) (1986). *Theoretical aspects of reasoning about knowledge: proceedings of the 1986 conference*. Los Altos, CA: Morgan Kaufmann.
- Hellan, L. (1981). On semantic scope. In F. Heny (ed.), *Ambiguities in Intensional Contexts*, Dordrecht: D. Reidel, 1981.
- Hobbs, J. R. (1983a). Metaphor interpretation as selective inferencing: cognitive processes in understanding metaphor (Part 1). *Empirical Studies of the Arts*, 1 (1), 17–33.
- Hobbs, J. R. (1983b). Metaphor interpretation as selective inferencing: cognitive processes in understanding metaphor (Part 2). *Empirical Studies of the Arts*, 1 (2), 125–141.
- Hobbs, J. R. (1985a). On the coherence and structure of discourse. Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University, Calif.
- Hobbs, J.R. (1985b). Ontological promiscuity. *Procs. 23rd Ann. Meeting of the Association for Computational Linguistics*, Univ. of Chicago.
- Johnson, M. (1987). *The body in the mind*. Chicago: Chicago University Press.
- Johnson-Laird, P. N. (1983). *Mental models*. Harvard University Press: Cambridge, Mass.

- Konolige, K. (1983). A deductive model of belief. *Procs. 8th Int. Joint Conf. on Artificial Intelligence*, Karlsruhe, W. Germany.
- Konolige, K. (1986). What awareness isn't: a sentential view of implicit and explicit belief. In Halpern (1986).
- Kraut, R. (1983). There are no *de dicto* attitudes. *Synthese*, 54, 275–294.
- Lakoff, G. (1987). *Women, fire and dangerous things*. Chicago: University of Chicago Press.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: U. Chicago Press.
- Langford, C. H. (1942). The notion of analysis in Moore's philosophy. In P.A. Schilpp (Ed.), *The Philosophy of G.E. Moore*, Northwestern Univ.: Evanston and Chicago.
- Larsen, S. F. (1987). Remembering and the archaeology metaphor. *Metaphor and Symbolic Activity*, 2 (3), 187–199.
- Levesque, H. J. (1984). A logic of implicit and explicit belief. *Procs. Natl. Conf. on Artificial Intelligence*, Univ. of Texas at Austin.
- Linsky, L. (1983). *Oblique contexts*. Chicago: U. Chicago Press.
- Maida, A. S. (1984). Belief spaces: foundations of a computational theory of belief. Tech. Rep. CS-84-22, Dept. of Comp. Sci., The Pennsylvania State University.
- Maida, A. S. (1986). Introspection and reasoning about the beliefs of other agents. *Procs. 8th Conference of the Cognitive Science Society*.
- Maida, A. S. & Shapiro, S. C. (1982). Intensional concepts in propositional semantic networks. *Cognitive Science*, 6, 291–330.
- Mates, B. (1950). Synonymity. *Univ. of California Publications in Philosophy*, 25 201–226. [Reprinted in L. Linsky (ed.), *Semantics and the philosophy of language*, Urbana: U. Illinois Press, 1952.]
- McCarthy, J. (1979). First order theories of individual concepts and propositions. In J. E. Hayes, D. Michie & L. I. Mikulich (Eds.), *Machine Intelligence 9*. Chichester: Ellis Horwood.
- Montague, R. (1970). Pragmatics and intensional logic. *Synthese*, 22, pp. 68–94.
- Montague, R. (1974). Syntactic treatments of modality, with corollaries on reflexion principles and finite axiomatizability. In R.H. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague*, Yale University Press, 1974.
- Moore, G. E. (1942). A reply to my critics. In P.A. Schilpp (Ed.), *The Philosophy of G.E. Moore*, Northwestern Univ.: Evanston and Chicago.

- Morgenstern, L. (1986). A first-order theory of planning, knowledge, and action. In Halpern (1986).
- Parsons, T. (1980). *Nonexistent objects*. New Haven: Yale Univ. Press.
- Partee, B. H. (1982). Belief sentences and the limits of semantics. In S. Peters & E. Saarinen (Eds.), *Processes, beliefs and questions*. Dordrecht: Reidel.
- Perlis, D. (1985). Languages with self-reference I: Foundations. *Artificial Intelligence*, 25, 301–322.
- Perlis, D. (1988). Languages with self-reference II: knowledge, belief, and modality. *Artificial Intelligence*, 34 (2), 179–212.
- Quine, W.V.O. (1981). Intensions revisited. In W.V. Quine, *Theories and things*. Cambridge, Mass: Harvard U. Press.
- Rapaport, W. J. (1986). Logical foundations of belief representation. *Cognitive Science*, 10: 371–422.
- Reddy, M. J. (1979). The conduit metaphor — a case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and Thought*, Cambridge, UK: Cambridge University Press.
- Saarinen, E. (1981). Quantifier phrases are (at least) five ways ambiguous in intensional contexts. In F. Heny (ed.), *Ambiguities in Intensional Contexts*, Dordrecht: D. Reidel, 1981.
- Schank, R.C. (1973). Identification of conceptualizations underlying natural language. In R. C. Schank & K. M. Colby (eds), *Computer models of thought and language*, San Francisco: Freeman.
- Shapiro, S. C. & Rapaport, W. J. (1986). SNePS considered as a fully intensional propositional semantic network. *Procs. 5th National Conf. on Artificial Intelligence (AAAI-86)*,
- Stich, S. (1983). *From folk psychology to cognitive science: the case against belief*. Cambridge, Mass.: MIT Press.
- Sweetser, E.E. (1987). Metaphorical models of thought and speech: a comparison of historical directions and metaphorical mappings in the two domains. In L. Michaelis, J. Aske & H. Filip (Eds), *Procs. 13th Annual Meeting of the Berkeley Linguistics Society*.
- Sweetser, E.E. (forthcoming). *From etymology to pragmatics*. Cambridge, Cambridge University Press.
- Talmy, L. (1985). Force dynamics in language and thought. *Chicago Linguistics Society 21, Part 2: Parasession on causatives and agentivity*, 293–337.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12, 49–100.

- Thomason, R. H. (1980). A note on syntactical treatments of modality. *Synthese*, 44, 391–395.
- Tomlinson, B. (1986). Cooking, mining, gardening, hunting: metaphorical stories writers tell about their composing processes. *Metaphor and Symbolic Activity*, 1 (1), 57–79.
- Wilks, Y. & Ballim, A. (1987). Multiple agents and the heuristic ascription of belief. *Procs. 10th Int. Joint Conf. on Artificial Intelligence*, Milan, Italy.
- Wilks, Y., Ballim, A. & Barnden, J. A. (1988). Belief ascription, metaphor and intensional identification. *Memoranda in Computer and Cognitive Science*, No. MCCS-88-138, Computing Research Laboratory, New Mexico State University.
- Wilks, Y. & Bien, J. Beliefs, points of view and multiple environments. *Cognitive Science*, 7(2), 95–119.