COMPUTATIONAL MODELS OF BELIEF AND THE SEMANTICS OF BELIEF SENTENCES

Technical Note 187

SRI Project 7910/5844

June 1979

By: Robert C. Moore, Computer Scientist
Gary G. Hendrix, Program Manager

Artificial Intelligence Center
Computer Science and Technology

The work reported herein was supported by the National Science Foundation under Grant No. MCS76-22004, and by the Defense Advanced Research Projects Agency under Contract N00039-79-C-0118 with the Naval Electronic Systems Command.

ABSTRACT

This paper considers a number of problems in the semantics of belief sentences from the perspective of computational models of the psychology of belief. We present a semantic interpretation for belief sentences that is suggested by a computational model of belief, and show how this interpretation overcomes some of the difficulties of alternative approaches, especially those based on possible-world semantics. Finally, we argue that these difficulties arise from a mistaken attempt to identify the truth conditions of a sentence with what a competent speaker knows about the meaning of the sentence.

I COMPUTATIONAL THEORIES AND COMPUTATIONAL MODELS

Over the years the psychology of belief and the semantics of belief sentences have provided a seemingly endless series of fascinating problems for linguists, psychologists, and philosophers. Despite all the attention that has been paid to these problems, however, there is little agreement on proposed solutions, or even on what form solutions should take. We believe that a great deal of light can be shed on the problems of belief by studying them from the viewpoint of computational models of the psychological processes and states associated with belief. The role of computational theories and computational models in the cognitive sciences always seems to be a matter of controversy. When such theories and models are discussed by non-computer scientists, they are frequently presented in a rather apologetic tone, with assurances and caveats that, of course, this is all oversimplified and things couldn't really be like this, but...

This may be the result of an unwarranted inference that anyone who takes a computational approach in one of these disciplines thereby endorses what is sometimes called the thesis of "mechanism" (Lucas, 1961): that minds can be completely explained in terms of machines, which in contemporary discussions are usually taken to be computers. When the metaphysical doctrine of dualism was more widely held, the mechanism thesis could be rejected on the grounds that minds were nonphysical. It appears to be more fashionable to adhere to a materialistic metaphysics nowadays, but to hold that the way minds are embodied in brains is so complex as to be beyond all human understanding, or at least too complex to be represented by Turing machines or computer programs. On the basis of current knowledge, these questions appear to us to be completely open. The existing evidence may well have as little relevance to future discoveries as the arguments of the Greek philosophers about atomism have to modern atomic theory.

We wish to argue, however, that the usefulness of computational approaches in the cognitive sciences does not depend on how (or even

whether) these questions are eventually answered. In elaborating this view, it will be helpful to make a distinction between computational theories and computational models. We will say that a theory of a cognitive process is a computational theory if it claims that the process is a computational process. The mechanism thesis can be viewed as the claim that every cognitive process is a computational process. Obviously, one can hold that certain cognitive processes are computational without claiming that all are, so having a particular computational theory of some cognitive process still leaves the mechanism thesis an open question.

The construction of computational theories is the most obvious use of computational ideas in the cognitive sciences. However, even without computational theories, computational models can be extremely useful. By a computational model of a cognitive process we mean a computational system whose behavior is similar to the behavior of the process in some interesting way. The important point is that one can make use of computational models without making any claims about the nature of the process being modeled. For example, the use of computational models in weather forecasting does not commit one to the claim that meteorological processes are computational.

What makes computational models in meteorology interesting is the fact that they can make useful predictions about the behavior of the system being modeled. In the cognitive sciences few models, computational or otherwise, have such predictive power, and we are hard pressed to think of any cases in which the predictions that are made can be considered useful. Thus at our current level of understanding, prediction of behavior does not appear to be the most productive role for computational models of cognitive processes.

What computational models do seem to be good for is clarification of conceptual problems. Many of the most vexing problems in the cognitive sciences are questions as to how any physical system could have the properties that cognitive systems apparently possess. Computational models can often supply answers to questions of this kind

independently of empirical considerations regarding the way human (or other) cognitive systems actually function. The point is that conceptual arguments often proceed from general observations about some cognitive process to specific conclusions as to what the process must be like. One way of testing such an argument is to construct a computational model that satisfies the premises of the argument and then to see whether the conclusions apply to the model. When used in this way, a computational model may be best thought of not so much as a model of a process, but rather as a model (in the sense of "model theory" in formal logic) of the theory in which the argument is made. That is, a conceptual argument ought to be valid for all possible models that satisfy its premises, so it had better be valid for a particular computational model, independently of how closely that model resembles the cognitive process that is the "intended model."

In the remainder of this paper we will try to apply computational models in this way to investigate some of the problems about belief and the semantics of belief sentences. First we will present a model of belief that seems to satisfy most of our pretheoretical notions. Then we will ask what implications it would have for the semantics of belief sentences, if human belief were analogous to our model. As will be seen below, this leads us to some conclusions quite different from those drawn by other authors.

II INTERNAL LANGUAGES

Before going into the details of what a computational model of belief might look like, we need to deal with a set of objections that have been raised to one of the basic assumptions we will make. The assumption is that beliefs are to be explained in terms of expressions in some sort of internal language that is not the language used externally--a "language of thought" to use Fodor's (1975) term. To posess a particular belief is to bear a certain computational relation to the appropriate expression in this internal language. This sort of

explanation has frequently been attacked by philosophers, particularly Ryle (1949) and Wittgenstein (1953), as incomprehensible, but this is surely a case of a conceptual argument that fails when applied to computational models. Many computer systems have been built that have internal languages in this sense, and we are unable to find appeals to any features of human cognition in the usual arguments against internal languages that would make these arguments inapplicable to those systems. In particular, the internal language used in one of these computer systems always has a well-defined syntax, and usually a clear notion of inference defined in terms of manipulations of the formulas of the language. Whether these languages have truth-conditional semantics is more problematical, but for purposes of psychological explanation this may well be unnecessary. After all, if truth-conditional semantics cannot be given for the internal language the machine uses, although we know how to explain the behavior of the machine (since it was specifically designed to have that behavior), then such semantics cannot be required for the explanation. But if truth-conditional semantics is not required to perform "psychological explanation" for machines, why should it be required for humans?

Even if we accept the existence of computer systems that use such an internal language as a "model-theoretic" demonstration that the arguments against internal languages are misguided, just where they go wrong remains an interesting question. It would clearly be impossible to examine all such arguments in this brief paper, (and we confess that we are not scholars of that literature,) but it may be instructive to look at at least one example. One familiar type of argument used by behaviorists against any number of concepts in cognitive psychology runs something like this:

> The only evidence admissible in psychology is behavioral evidence. There will always be many hypotheses, equally compatible with any possible behavioral evidence, about what X an organism has. Therefore, there is no empirical content to the claim that an organism has one X rather than another. Therefore the notion of X is unintelligible.

Quine has used this type of argument repeatedly in his discussions of the indeterminacy of translation (1960), ontological relativity (1971), and knowlege of grammatical rules (1972). "Set of expressions in an internal language" is one of the concepts frequently substituted for "X" in this schema. The argument has some plausibility when applied to the human mind, where we have very little idea of how expressions in an internal language might be physically represented. It loses that plausibility when applied to computational models. If we recast the argument we can see why:

> The only evidence admissible for analyzing computer systems is behavioral evidence. There will always be many hypotheses, equally compatible with any possible behavioral evidence, about what set of expressions in an internal language form the basis of a computer system's "beliefs." Therefore, there is no empirical content to the claim that a computer system has one set of expressions rather than another. Therefore the notion of a set of expressions in an internal language in a computer system is unintelligible.

Where this argument breaks down depends on what is taken to be behavioral evidence. If we take behavioral evidence to be simply the input/output behavior of the system when it is running normally, then there is certainly more than behavioral evidence to draw upon. With a computer system we can do the equivalent of mapping out the entire "nervous system," and so understand its internal operations as well. On the other hand, if behavioral evidence includes internal behavior, it becomes much less plausible to say that there will be no way to tell which set of expressions the system possesses.

At this point, a computer scientist might be tempted to shout, "Of course! To find out what internal expressions the systems has, all you have to do is to print them out and look at them!"--but there is more to be said for the Quinean argument than this. What are directly observable, after all, are the physical states of the machine and their causal connections. There are many levels of interpretation between them and the print-out containing the set of expressions we wish to

attribute to the machine. A Quinean might argue that there will be other interpretations that will lead to a different set of expressions, perhaps in a different internal language. With computer systems, however, the fact that they are designed to be interpreted in a certain way makes it extremely likely that any alternative interpretation would be far less natural, and so could be rejected on general grounds of simplicity and elegance. If this were not the case, it would be like discovering that the score of Beethoven's Ninth is actually the score of Bach's Mass in B Minor under a different, but no more complex, interpretation of the usual system of musical notation.

The Quinean argument fares somewhat better when applied to humans because there is no a priori reason to assume that human brains are designed to be interpreted in any particular way. Thus it is more plausible that there might be multiple descriptions of the operation of the brain in terms of internal languages, and that these descriptions, while incompatible with one another, are nevertheless equally compatible with all the evidence, including neurological evidence. But as the example of the computer system shows, and contrary to the Quinean argument, there is also no a priori reason to assume that this must be the case. It is, as the saying goes, an empirical question. It should be clear that one of the empirical commitments of any theory in cognitive psychology is that there be a preferred interpretation of the physical system in terms of the entities postulated by the theory. If this commitment is recognized, then failure to find a preferred interpretation makes the theory not incoherent or unintelligble, but simply false.

In view of all this, the best that can be said for the Quinean argument is that it points out the possibility that there will be more than one theory compatible with any evidence that can be obtained. But this is always the case in science. Surely no one would suggest that atomic theory is incoherent because there might be some as yet undiscovered alternative that is equally compatible with the evidence. Thus our consideration of computational models leads us to agree with

Chomsky (1975, p. 182) that Quine's indeterminacy doctrine comes to no more than the observation that nontrivial empirical theories are underdetermined by evidence.

III A COMPUTATIONAL MODEL OF BELIEF

The basic outlines of the computational model of belief presented below should be familiar to anyone acquainted with developments in artificial intelligence or cognitive simulation over the past few years. In calling this a model of belief, however, we must be careful to distinguish between psychology and semantics. Our model is intended to be a psychologically plausible account of what might be going on in an organism or system that could usefully be said to have beliefs. Even if we assume that the model does describe what is going on, the semantic question remains of how the English word "believe" relates to the model. We will put off adressing that question until Section IV.

As we said in the preceding section, belief will be explained in our model in terms of a system's being in a certain computational relation to expressions in an internal language. We will call the set of expressions to which a system is so related the belief set of the system. The exact relationship between the expressions in this set and what we would intuitively call the beliefs of the system will be left unspecified until we discuss the semantics of belief sentences in Section IV. We will also be somewhat vague as to just what computational relation defines the belief set, but we can name some of the constraints it must satisfy. First of all, we will stipulate that, to be in the belief set of a system, an expression must be explicitly stored in the system's memory. It may turn out that we want to say the system has beliefs that would correspond to expressions that are not explicitly stored, but can be derived from stored expressions. In that case, the relationship between the system's beliefs and its belief set will be more complicated, but it will still be important to single out the expressions that are explicitly stored.

The fact that an expression is stored in the memory of the system cannot be sufficient, however, for that expression to be in the system's belief set. If the system is to be even a crude model of an intelligent organism, it will need to have propositional attitudes besides belief, which we would also presumably explain in terms of expressions in its internal language stored in its memory. We can account for this by treating the memory of the system as being logically partitioned into different spaces--one space for the expressions corresponding to beliefs, another for desires, another for fears, etc. These various spaces will be functionally differentiated by the processes that operate on them and connect them to the system's sensors and effectors. For example, perceiving that there is a red block on the table might directly give rise to a belief that there is a red block on the table, but probably not the desire or fear that there is a red block on the table. Similarly, wanting to pick up a red block might be one of the immediate causes of trying to pick up a red block, but imagining picking up a red block would presumably not.

This is a bit oversimplified, but not too much. Although it is true that perceiving a red block on the table could cause a fear that there is a red block on the table, this would need to be explained by, say, a belief that red blocks are explosive. In going from perception to belief, no such additional explanation is necessary. It seems completely compatible with our pretheoretical notions (which is what our model is supposed to reflect) to assume that we are simply built in such a way that we automatically accept our perceptions as beliefs unless they conflict with existing beliefs. (Anyone who does not think we are built this way should look out his window and try to disbelieve that what he sees is actually there.)

As to the internal language itself, we will again leave the details somewhat sketchy. For the purposes of this discussion, it will be sufficient to assume that the language is that of ordinary predicate logic augmented by intensional operators for propositional attitudes. The expressions in a belief set would be well-formed formulas in this

language. The basic inference procedures should certainly be inclusive enough so that there is some way of applying them to generate any valid inference, but they could include procedures for generating plausible inferences as well. The important point is that, to interpret a set of formulas as a belief set, there had better be a well-defined notion of inference for them, since people clearly draw inferences from their beliefs. It is equally important, moreover, that there be a notion of an inference process in the model. The basic inference procedures merely define what inferences are possible, not what inferences will actually be drawn. There must be a global inference process that applies specific inference procedures to the formulas in the belief set and adds the resulting formulas to the belief set.

As simple as this model is, it seems to account fairly well for the obvious facts about belief. For example, it explains how "one-shot" learning can occur when one is told something. The explanation is that the hearer of a natural-language utterence decodes it into a formula in his internal language and adds the formula to his belief set. This idea, which seems to be almost universally accepted in generative linguistics and cognitive psychology, would hardly be worth mentioning if it were not for the fact that it differs so radically from the view presented in behaviorist psychology. According to standard behaviorist assumptions, we would expect that repeated trials and reinforcements would be necessary for learning to occur. This has some plausibility in the case of complex skills or large bodies of information, but a moment's reflection will show that very little learning fits this picture. Most "learning" consists of acquiring commonplace information such as where the laundry was put and what time dinner will be ready. Our model seems to explain this type of learning much better than does reinforcement of responses to a given stimulus.

A slightly less trivial, but still fairly obvious comment is that this model has no difficulty explaining how the system could accept one belief, yet reject another that is its logical equivalent. Suppose that beliefs are individuated more or less as are formulas in the internal

language. Suppose further that the system has a particular formula P in its belief set that is logically equivalent to another formula Q, in the sense there is some way of applying the basic inference procedures of the system to infer Q from P and vice versa. The system may not put Q in its belief set, however, because it never tries to derive Q, or because its heuristics for applying its inference procedures are not sufficient to find the derivation of Q, or because the derivation of Q is so long that it exhausts the system's resources of memory and time. We raise this point because the possibility that "A believes P" is true and "A believes Q" is false, even though P and Q are logically equivalent, is currently considered to be a major problem in the semantics of belief sentences, especially for theories based on possible-world semantics. In view of the voluminous literature this problem has generated (Montague, 1970) (Partee, 1973, 1978) (Stalnaker, 1976) (Cresswell, 1979), it is striking to note that, if reality is even vaguely like our a computational model, this is no problem at all for the psychology of belief. This suggests to us that the problem is artificial, a point we will return to in Section V.

A more serious problem that can be handled rather nicely in this model is the question of what beliefs are expressed by sentences containing indexicals such as "I," "now," and "here." This is particularly troublesome for theories that take the language of thought to be identical to the external natural language. To take an example suggested by the work of Perry (1977, 1979), suppose that Jones has a belief he would express by saying "I am sitting down." We would take Jones's use of the word "I" to be a reference to Jones himself and take Jones's belief to be about himself. What is it that makes Jones's belief a belief about himself? It can't be simply that he has used the word "I" to express it, because he might not be using "I" as it is normally used in English; he must also believe or intend that in using "I" he refers to himself. But if this belief or intention consists in having certain English sentences stored in the appropriate space in his memory, it is hard to see how the explanation can avoid being circular. It is certainly not sufficient for Jones to believe "When I use 'I,' it

refers to me," because this doesn't express the right belief or intention unless it has already been established that Jones uses "me" and "I" to refer to himself.

One way to try to get out of this problem is to say that Jones has some nonindexical description of himself and that his use of "I" is shorthand for this description. But, as Perry points out, having such a description is neither necessary nor sufficient to account for his use of "I." To see that it is not necessary, suppose Jones is the official biographer of Jimmy Carter, but he becomes insane and begins to believe that he actually is Jimmy Carter. Thus his beliefs include things he would express by "My name is 'Jimmy Carter,'" "I am President of the United States," "My daughter is Amy Carter," and so forth, in great detail. It does not seem to be logically impossible that <u>all</u> the nonindexical descriptions he attributes to himself are in fact true of Jimmy Carter and not true of him. On the description theory of indexicals, this should mean that Jones uses "I" to refer to Jimmy Carter and that his beliefs are all true. But it is intuitively clear that he still uses "I" to refer to himself, and that his beliefs are all delusional and false. On the other hand, suppose he is not insane, but uses "I" as a shorthand for some true description of himself such as "Jimmy Carter's biographer." Hence, when he says "I am sitting down" he expresses the belief "Jimmy Carter's biographer is sitting down." This does not explain his belief that he is sitting down, however, unless he also believes that <u>he</u> is Jimmy Carter's biographer.

In view of Kripke's (1972) critique of the description theory of proper names, it is not surprising that the description theory of indexicals doesn't work either. Nevertheless, it is interesting that Kripke's alternative, which does seem to work for proper names, still does not work for indexicals. Kripke's theory is essentially that when someone uses a proper name, it derives its reference from the occasion on which he acquired the use of the name, and that this creates a causal chain extending back to the original "dubbing" of the individual with that name. Thus, our use of "Kripke" refers to Kripke because we have

acquired the name from occasions on which it was used to refer to Kripke. But this can't explain the use of the word "I," because no one ever acquires the use of "I" from an occasion on which it was used to refer to him.

In our computational model we can explain the use of "I" by assuming that the system has an individual constant in its internal language--call it SYS--that intrinsically refers to the system itself, and that the system uses "I" to express in English formulas of its internal language that involve this individual constant. This may seem to be no progress, since we are left with the task of explaining how SYS refers to the system. This is an easier task, however. A substantial part of the problem posed by "I" is that it is part of a natural language, and natural languages are acquired. The problem about beliefs being English sentences in the mind is that the person might have acquired a nonstandard understanding of them. Similarly, Kripke's causal-chain theory fails to explain the reference of "I" because "I" doesn't fit the assumptions the theory makes about how terms are acquired. As Fodor (1975) points out, however, if the internal language of thought is in fact not an external natural language, then we can assume that it is innate, and we are relieved of the problem of explaining how the expressions in it are acquired.

We can explain how SYS refers intrinsically to the system in terms of the functional role it plays. The system can be so constructed that, when it seems to see a red block, a formula roughly equivalent to "SYS seems to see a red block" is automatically added to the belief set, or at least becomes derivable in the belief space. Similarly, wanting to pick up a red block is intrinsically connected to "SYS wants to pick up a red block," and so forth. If the meaning of SYS is "hard-wired" in this way, then learning the appropriate use of "I" requires only learning something like "Use 'I' to refer to SYS." This type of explanation cannot be given in terms of the word "I" alone, because people are not hard-wired to use "I" in any way at all.

IV THE SEMANTICS OF BELIEF SENTENCES

We hope the picture that we have presented so far is plausible as a model of the psychology of belief. If it is, then we have solved a number of interesting conceptual problems. That is, we have given at least a partial answer to the original question of how any physical system could have the properties that cognitive systems appear to have. Of course, solving conceptual problems is different from solving empirical problems; we have very little evidence that human cognitive systems actually work this way. On the other hand, we tend to agree with Fodor (1975, p. 27) that the only current theories in psychology that are even remotely plausible are computational theories, and that having remotely plausible theories is better than having no theories at all.

In light of the foregoing, there is a truly remarkable fact: although the psychology of belief is relatively clear conceptually, the semantics of belief sentences is widely held to suffer from serious conceptual problems. This might be less remarkable if the authors who find difficulties with the semantics of belief sentences rejected our conceptual picture, but that is not necessarily the case. For instance Cresswell (1979, p. 8) acknowledges that "it is probably true that what makes someone believe something is indeed standing in an certain relation to an internal representation of a proposition," and it appears that Partee (1978) would also be favorably inclined towards this kind of approach.

It seems to us that, if we have a clear picture of what the psychology of belief is like, it ought to go a long way towards telling us under what condition attributions of belief are true. That is, it ought to give us a basis for stating the truth-conditional semantics of belief sentences. Our general view should be clear by now: if our computational model of belief is roughly the way people work, then "A believes that S" is true if and only if the individual denoted by "A" has the formula of his internal language that corresponds appropriately

to "S" in his belief set, or can perhaps be derived in his belief set with limited effort. This latter qualification can be included or excluded, according to whether one wants to say that a person believes things he may never have thought about but that are trivial inferences from his explicit beliefs, such as the fact that 98742 is an even number, or that Anwar Sadat is a creature with a brain.

To complete this view we have to specify the relation between an attributed belief and the coresponding formula in the belief set. As a first approximation, we could say that "A believes that S" is true if and only if the individual denoted by "A" has in his belief set a formula he would express by uttering "S." For example, "John believes that Venus is the morning star" would be true if and only if the person denoted by "John" has a formula in his belief set that he would express as "Venus is the morning star." We believe this formulation is on the right track, but it has a number of difficulties that need to be repaired. For one thing, it is obviously not right for de re belief reports, such as "John believes Bill's mistress is Bill's wife." On its most likely reading, "Bill's mistress" is a description used by the speaker of the sentence, not John. We would not expect John to express his belief as "Bill's mistress is Bill's wife." We will return to the issue of de re belief reports later, but for now we will confine ourselves to de dicto readings.

Another apparent problem is the notion of a sentence in an external language expressing a formula of an internal language, but this can be dealt with by the same sort of functional explanation that we used initially to justify the notion of a belief set. A sentence expresses the internal formula that has the right causal connection with an utterence of the sentence. That causal connection may be complicated, but it is basically like the one between the contents of a computer's memory and a print-out of those contents. We will therefore assume that, given a causal account of how the production of utterances depends on the cognitive state of the speaker, there is a best interpretation of which formula in the internal language is expressed by a sentence in the external language.

A genuine problem in our current formulation is the fact that a person cannot be counted on to express his belief that Venus is the morning star as "Venus is the morning star," unless he is a competent speaker of English. A possible way around this would be to say that A believes that P if A has in his belief set a formula of his internal language that a competent speaker of English would express by uttering "S." This would be plausible, however, only if we assume that every person has the same internal language, and that expressions in the language can be identified across individuals. It might well be true that the internal language has the same syntax for all persons, since this would presumably be genetically determined, but that is not enough. We would have to further assume that a formula in the internal language means the same thing for every person.

This is clearly not the case, however, as many examples by Putnam (1973, 1975), Kaplan (1977), and Perry (1977, 1979) demonstrate. What these examples show is that two persons can be in exactly the same mental state (which, on our view, would require having the same belief sets), yet have different beliefs, because their beliefs are about different things. This should not be surprising, since there is nothing in our computational model to suggest that the reference or semantic interpretation of every expression in the language of thought is innate. Some expressions can be considered to have an innate interpretation because of the functional role they play in the model. Logical connectives and quantifiers in the internal language might have an a priori interpretation because of the way they are treated by innate inference procedures, and we have already discussed the idea that a cognitive system could have a constant symbol that intrinsically refers to the system. Predicates and relations for perceptual qualities, such as shapes and colors, would also seem to have a fixed interpretation based on the functional role they play in perception.

For most other expressions, including most individual constants and nonperceptual functions, predicates, and relations, there seems no reason to suppose that the interpretations are innately given. In fact,

"concept learning" seems to be best accounted for by assuming that the internal language has an abitrarily large number of "unused" symbols on which information can be pegged. Acquiring a natural-kind concept might begin by noticing regularities in the perceptual properties of certain objects and deciding to "assign" one of the unused predicate symbols to that type of object. Then one could proceed to investigate the properties of these objects, adding more and more formulas involving this predicate to his belief set. Note that there is no reason to assume that the formulas added to the belief set constitute a biconditional definition of the concept; hence this picture is completely compatible with Wittgenstein's (1953) observation that we typically do not know necessary and sufficient condtions for application of the concepts we possess. Furthermore, since it is the acquisition process that gives the predicate symbol its interpretation, we can accomodate Putnam's (1973, 1975) point that a concept's extension can be partly determined by unobserved properties of the exemplars involved in its acquisition. This also demonstrates, contrary to Fodor (1975), that concept learning can be explained in terms of an internal language, without assuming that the language already contains an expression for the concept.

It appears that concept acquisition processes like the one suggested above could provide the symbols of the internal language with a semantic interpretation via the sort of causal chain that Kripke and Putnam discuss in connection with the semantics of proper names and natural-kind terms. Assuming the details can be worked out, we can use this semantic interpretation to try to define sameness of meaning across persons for expressions in the internal language. We can do this along lines suggested by Lewis's (1972) definition of meaning for natural languages: an expression P has the same meaning for A as Q has for B if P and Q have the same syntactic structure and each primitive symbol in P has the same intension for A as the corresponding symbol in Q has for B. We take an intension to be a function from possible worlds to extensions, and we assume that the intension of a primitive symbol is either innate, because of the functional role of the symbol, or is acquired in accordance with the causal-chain theory.

The problem with this definition is that two primitive symbols can have the same intension, but differ in what we would intuitively call meaning. Suppose John believes that Tully and Cicero are two different people. He might have in his belief set expressions corresponding to:

NAME(PERSON3453) = "TULLY"
NAME(PESRON9876) = "CICERO"
NOT(PERSON3453 = PERSON9876)

The best that the causal-chain theory can do for us is to provide the same intension for both PERSON3453 and PERSON9876, a function that picks out Cicero in all possible worlds. But clearly, these two symbols do not have the same meaning for John. In general, we probably would want to say that two symbols differ in meaning for an individual unless they have the same intension and are treated as such in the person's belief set (e.g., by having a formula asserting that they are necessarily equivalent).

To accomodate this observation, we will say that if the primitive symbol P has the same intension for A that the primitive symbol Q has for B, then P has the same meaning for A that Q has for B, providing either that these are the only symbols having that intension for A and B, or that the same expression in a common external language expresses P for A and Q for B. This latter condition may seem arbitrary, but it will allow us to say that if Bill and John both believe that Cicero denounced Catiline and Tully did not, then they both believe the same things. To use Quine's (1971, p. 153) phrase, this amounts to "acquiescing in our [or in this case, Bill and John's] mother tongue."

These criteria obviously do not guarantee that, if two persons possess symbols with the same intension, there is some way to determine which ones have the same meaning. There may be other conditions that would allow us to do this that we have not thought of, but there will undoubtedly be residual cases. Suppose a language has two terms, P and Q, that, unknown to the speakers of the language, are rigid designators for the same natural kind, and so have the same intension. In a language that has only one term for this natural kind, it might well be

impossible to express the belief that these speakers express when they say, "Some P's are not Q's." Imagine a culture in which the idea of the relativity of motion was so deeply embedded that they had no concept of X going around Y rather than Y going around X, but only X and Y being in relative circular motion. How would we go about explaining to them what it was that got Galileo into trouble?

We are finally in a position to state the truth conditions for de dicto belief reports that seem to follow from our computational model. First, we will say that an English expression "S" expresses the meaning of an internal expression P for an individual A just in case, for any competent English speaker B, there is an internal expression Q that has the same meaning for B as P has for A, and "S" expresses Q for B. Then a de dicto belief report of the form "A believes that S" is true if and only if the individual denoted by "A" has in his belief set a formula P such that "S" expresses the meaning of P for him.

To modify this theory to account for de re belief reports we will essentially reconstruct Kaplan's (1969) approach to apply to the internal language. According to our computational model, having a belief comes down to having the right formula in in one's belief set, and a belief report tells us something about that formula. A de dicto belief report, such as "John believes Venus is the morning star," provides us with a sentence that expresses the meaning of the formula in the belief set. In a de re belief report, such as "John believes Bill's mistress is Bill's wife," part of the sentence, in this case "Bill's wife," need not express the meaning or intension of any part of the corresponding formula. Instead, it expresses the reference of part of the formula. Suppose that the relevant formulas in John's belief set are something like:

```
NAME(PERSON55443) = "BILL"
WIFE(PERSON55443) = PERSON12345
```

If these formulas are the basis for the assertion that John believes Bill's mistress is Bill's wife, then it must at least be the case that the occurences of PERSON12345 in John's belief set refer to

Bill's mistress. Otherwise, if John's belief is about anybody at all, then it is that person rather than Bill's mistress whom John believes to be Bill's wife. Something more than this is required, though. De re belief reports are generally held to support existential generalization. That is, from the fact that John believes Bill's mistress is Bill's wife we can infer that there is someone whom John believes to be Bill's wife. Phrasing it this way, however, we seem to be saying that John not only believes Bill is married, but he can pick out the person he thinks Bill is married to. If John has merely been told that Bill has been seen around lately with a beautiful woman and he has inferred that she must be his wife, then we could not really say that there is some specific person that he believes to be Bill's wife. There seems to be a certain amount of identifying information that John must have about PERSON12345 for his belief set to justify a de re belief report, although it is not always clear exactly what this information would be.

Now we can fully state our theory of the semantics of belief sentences. A sentence of the form "A believes S" is true if and only if the individual denoted by "A" has in his belief set a formula P that meets the following two conditions: first, the subexpressions of "S" that are interpreted de dicto must express the meaning for him of the corresponding subexpressions of P; second, the subexpressions of "S" that are interpreted de re must have the reference for him of the corresponding subexpressions of P, and he must be able to pick out the reference of those subexpressions of P.

V CONCLUSION

The truth-conditional semantics for belief sentences presented above is a fairly complicated theory, but that really should not count against it. Most of its complexity was introduced to explain how a belief report in English could be true of someone who is not a competent speaker of that language. Most alternative theories of belief ignore this question entirely. All the formulations of possible-world

semantics for belief that we know of, for instance, assume an unanalyzed accessibility relation between a person and the possible worlds compatible with his beliefs. That relation must surely be mediated somehow by his psychological state or his language, but no explanation of this is given. Furthermore, the most serious problem that plagues possible-world theories, the problem of distinguishing among logically equivalent beliefs of the same person, is no problem at all in our theory.

The really interesting question for us, though, is not whether one particular semantic theory is superior to another, but why so little effort has been made thus far to develop an account of the truth conditions of belief sentences in terms of psychological states and processes. Since belief is a psychological state, it seems that this would be the most natural approach to follow. Almost all the recent work on the semantics of belief sentences, however, appears to strive for independence from psychology. Most of this work tries to define belief in terms of a relation between persons and some sort of nonpsychological entities, with the relation either left unanalyzed or analyzed in nonpsychological terms [e.g., (Hintikka, 1962, 1969) (Montague, 1974) (Partee, 1973, 1978) (Stalnaker, 1976) (Cresswell, 1979) (Quine, 1956, 1960)]. We can only speculate as to why this is the case, but we can think of at least two probable motivations.

One motivation is what Cresswell calls "the autonomy of semantics"--the idea that the goal of semantics is to characterize the conditions under which a sentence of a language is true, and that this can be done independently of any considerations as to how someone could know what the sentence means or believe that what the sentence says is true. Thus we can say that "The cat is on the mat," is true if and only if the object refered to by "the cat" bears the relation named by "is on" to the object refered to by "the mat", without raising or answering any psychological questions. The point that the truth conditions of sentences do not in general involve psychological notions seems well taken, but it surely does not follow that they <u>never</u> do. No one seems

to object to giving the truth conditions of sentences about physical states in terms of physical relations and physical objects, as in the example above. Why then, should there be any objection to giving the truth conditions for sentences about psychological states in terms of psychological relations and psychological objects?

To look at the matter a little more closely, the possible-world theories attempt to give the semantics of belief sentences in terms of semantic rather than psychological objects. That is, these theories claim that the objects of belief are built out of the constructs of the semantic theory itself. This would be a very interesting claim if it were true, but the failure up to now to make such a theory work suggests that it is probably not. If this assessment is correct, it seems natural to assume that the truth conditions of sentences about belief and other psychological states will involve the objects described by true psychological theories. If a true theory of the psychology of belief turns out to require the notion of an internal language, then it is probable that the truth conditions for belief sentences will involve expressions of that language.

The other motivation for seeking a nonpsychological semantics for belief sentences is the desire to unify the kind of truth-conditional semantics that we have been discussing with what is sometimes called "linguistic semantics," the task of characterizing what competent speakers know about the "meaning" of the sentences of their language. The most straightforward way to make this unification is to assume that the semantic knowledge that competent speakers of a language have is knowledge of the truth conditions of the language's sentences--a view that is, in fact, widely endorsed (Davidson, 1967) (Moravcsik, 1973) (Partee, 1978) (Woods, 1978) (Cresswell, 1979). It is quite implausible, however, that the kind of theory we have been sketching is what people know about belief or belief sentences. The root of the problem is our claim that the truth conditions for belief sentences can ultimately be stated only in terms of a true theory of the psychology of belief. But it is no more plausible that all speakers know such a

theory than that all speakers know true theories of physics, chemistry, or any other science.

Our answer to this objection is that the idea that the semantic knowledge of speakers amounts to knowledge of truth conditions is simply mistaken. This is a general point that applies not only to sentences about psychological states, but to many other kinds of sentences as well. As we mentioned in Section IV, Putnam has convincingly argued that the extension of natural-kind terms generally depends not simply on what speakers of a language know or believe about the extension of the term, but also on what properties the objects that the term is intended to describe actually possess. But this means that speakers do not, in fact, know the truth-conditions of sentences that involve natural-kind terms. The properties that speakers believe characterize the extension of a natural-kind term may turn out to be incomplete or even wrong. When it was discovered that whales are mammals, what was discovered was just that. It was not discovered that whales did not exist, even if being a fish was previously central to what speakers of English believed about the truth conditions of "X is a whale." In general, the truth conditions for a natural-kind term depend not so much on the knowledge of competent speakers as on true scientific theories about the natural kind in question. Viewed from this perspective, the truth conditions of belief sentences depend on what turns out to be true in psychology because belief states form natural kinds in the domain of psychology.

According to our computational model, what a competent speaker of a language needs to know about the meaning of a sentence is not its truth conditions, but what formula in his internal language the sentence expresses in a given context. Of course, as we discussed in Section IV, this formula has truth conditions, and it seems plausible to say that the truth conditions of a sentence in a context are the same as those of the formula it expresses in that context. Now, knowing a formula in the internal language that has the same truth conditions as the sentence is something like knowing the truth conditions of that sentence, but not very much like it. In particular, it is nothing like knowing the

statement of those truth conditions in any of the semantic theories we have discussed.

In the case of a belief sentence, the corresponding formula in the internal language might be thought of as an expression in a first-order language with a belief operator. If the hearer of "John believes that snow is white," takes "John" to refer to the same person as his internal symbol PERSON98765, and takes "snow is white" to express WHITE(SNOW), then the whole sentence might express for him the formula BELIEVE(PERSON98765,WHITE(SNOW)). The functional roles and causal connections of the symbols in this formula determine its truth conditions, and those must be right for this formula to actually have the meaning for the hearer that John believes snow is white. Otherwise the hearer has not understood the sentence. To get those truth conditions right the hearer might have to have a lot of knowledge about belief, such as that people generally believe what they say, that they often draw inferences from their beliefs, and that they usually know what they believe. Knowing these properties of belief would help pin down the fact that belief is the psychological state denoted by BELIEVE, yet these properties do not by any means constitute necessary and sufficient truth conditions for formulas involving BELIEVE. But it is only required that these formulas have such truth conditions, not that the hearer <u>know</u> them.

The mistaken attempt to identify truth conditions with what speakers know about the meaning of sentences in their language has led to many pseudoproblems. For instance, Partee (1978) raises the question of whether for possible-world semantics to be correct, an infinte number of possible world models would have to exist in our heads. She concludes that they would not because "performance limitations" could let us get by with a finite number of finite models. This whole issue seems to be pretty much beside the point, however. Even for notions for which possible-world semantics appears to be adequate, such as the concept of necessity, nothing approximating possible worlds needs to be in our heads, although something like modal logic might.

Another example of the confusion that results from trying to unify these two notions of semantics is Woods's (1978) attempt to base a theory of meaning on "procedural semantics." Woods tries to identify the meaning of a sentence with some sort of ideal procedure for verifying its truth, saying that this procedure is what someone knows when he knows the meaning of the sentence. This has an advantage over possible-world semantics in that it can provide distinct meanings for logically equivalent sentences, since two different procedures could compute the same truth value in all possible worlds. The "procedures" that Woods is forced to invent, however, are not computable in the usual sense, even in principle. For example, to account for quantification over infinite sets he proposes infinite computations, while for propositional attitudes he suggests something like running our procedures in someone else's head. The sense in which these nonexecutable procedures are procedures at all is left obscure.

Partee starts from a particular notion of truth conditions, that of Montague semantics, and asks how such conditions could be represented in the head of a speaker. Woods starts from something that could be in the head of a speaker, i.e., procedures, and tries to make them yield truth conditions. In both cases, unlikely theories result from trying to say that it is truth conditions that are in the head, when all that is required is that what is in the head <u>have</u> truth conditions.

In this paper we have examined a wide range of issues from the perspective of computational models of psychological processes and states. These issues include the legitimacy of psychological models based on internal languages, the problem of distinguishing logically equivalent beliefs, the psychology of having beliefs about oneself, belief reports about a nonspeaker of the language of the report, and the relation between truth-conditional and linguistic semantics. We do not claim to know whether the computational models we have proposed provide a correct account of all the phenomena we have discussed. What we do claim, however, is that many abstract arguments as to how things must be can be shown to be incorrect, and that many confusing conceptual

problems can be clarified when approached from the standpoint of the concrete examples that computational models can provide.

REFERENCES

Chomsky, N. (1975) Reflections on Language, Pantheon Books, New York.

Cresswell, M. J. (1979) "The Autonomy of Semantics," unpublished manuscript, Victoria University of Wellington, March 1979.

Davidson, D. (1967) "Truth and Meaning," Synthese, No. 17, pp. 304-323.

Fodor, J. A. (1975) The Language of Thought, Thomas Y. Crowell Company, New York.

Hintikka, J. (1962) Knowledge and Belief: An Introduction to the Logic of the Two Notions, Cornell University Press, Ithaca, New York.

Hintikka, J. (1969) "Semantics for Propositional Attitudes," in Reference and Modality, L. Linsky (ed.) pp. 112-144, Oxford University Press, London, England, 1971.

Kaplan, D. (1969) "Quantifying In," in Reference and Modality, L. Linsky (ed.) pp. 112-144, Oxford University Press, London, England, 1971.

Kaplan, K. (1977) "Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals," unpublished manuscript, March 1977.

Kripke, S. A. (1972) "Naming and Necessity," in Semantics of Natural Language, D. Davidson and G. Harmon (eds.) pp. 253-355, D. Reidel Publishing Co., Dordrecht, Holland, 1972.

Lewis, D. (1972) "General Semantics," Semantics of Natural Language, D. Davidson and G. Harmon (eds.) pp. 169-218, D. Reidel Publishing Co., Dordrecht, Holland, 1972.

Lucas, J. R. (1961) "Minds, Machines and Goedel," in Minds and Machines, A. R. Anderson (ed.) pp. 43-59, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964.

Montague, R. (1970) "Pragmatics and Intensional Logic," Synthese No. 22, pp. 68-94.

Montague, R. (1974) Formal Philosophy: Selected Papers of Richard Montague, R. H. Thomason (ed.) Yale University Press, New Haven, Connecticut.

Moravcsik, J. (1973) "Comments on Partee's Paper," in Approaches to Natural Language, J. K. K. Hintikka et al. (eds.) pp. 349-369, D. Reidel Publishing Co., Dordrecht, Holland, 1973.

Partee, B. H. (1973) "The Semantics of Belief-Sentences," in Approaches to Natural Language, J. K. K. Hintikka et al. (eds.) pp. 309-336, D. Reidel Publishing Co., Dordrecht, Holland, 1973.

Partee, B. H. (1978) "Semantics--Mathematics or Psychology?", paper presented at University of Konstanz Colloquium, September 1978.

Perry, J. (1977) "Frege on Demonstratives," The Philosophical Review, Vol. 86, No. 4, October 1977.

Perry, J. (1979) "The Problem of the Essential Indexical," Nous 13, Indiana University.

Putnam, H. (1973) "Meaning and Reference," in Naming, Necessity, and Natural Kinds, S. P. Schwartz (ed.) pp. 118-132, Cornell University Press, Ithaca, New York, 1977.

Putnam, H. (1975) "The Meaning of Meaning," in Minnesota Studies in the Philosophy of Science, Vol. VII, Language, Mind, and Knowledge, K. Gunderson (ed.) pp. 131-193 University of Minnesota Press, Minneapolis, Minnesota, 1975.

Quine, W.V.O. (1956) "Quantifiers and Propositional Attitudes," in Reference and Modality, L. Linsky (ed.) pp. 112-144, Oxford University Press, London, England, 1971.

Quine, W.V.O. (1960) Word and Object, The M.I.T. Press, Cambridge, Massachusetts.

Quine, W.V.O. (1971) "The Inscrutability of Reference," in Semantics, D. D. Steinberg and L. A. Jakobovits (eds.) Cambridge University Press, London, England, 1971.

Quine, W.V.O. (1972) "Methodological Reflections on Current Linguistic Theory," in Semantics of Natural Language, D. Davidson and G. Harman (eds.) pp. 442-454, D. Reidel Publishing Co., Dordrecht, Holland, 1972.

Ryle, G. (1949) The Concept of Mind, Barnes and Noble, Inc., New York.

Stalnaker, R. C. (1976) "Propositions," in Issues in the Philosophy of Language, A. F. Mackay and D. D. Merrill (eds.) pp. 79-91, Yale University Press, New Haven, Connecticut, 1976.

Wittgenstein, L. (1953) Philosophical Investigations, Blackwell, Oxford, England.

Woods, W. A. (1978) "Procedural Semantics and a Theory of Meaning," unpublished manuscript, Bolt Beranek & Newman, Inc., May 1978.