The Knower's Paradox and the Logics of Attitudes

Nicholas Asher and Hans Kamp

AI TR 85-15 August 1985

This work supported in part by the Center for Cognitive Science, and Army Research Office grant #DAAG29-84-K-0060 to the Artificial Intelligence Laboratory.

Table of Contents

The Knower's Paradox and Logics of the Attitudes

As Montague and Kaplan (Kaplan & Montague, 1960, Montague, 1963) pointed out long ago, a syntactic treatment of propositional attitudes has a fundamental weakness: virtually all of epistemic logic must be sacrificed, if it, like ordinary logic, is to be applicable to arbitrary subject matter. Thomason (Thomason, 1980) extended these results to include a syntactic treatment of the logic of belief, and showed how they also apply to analyses of the attitudes that, while not strictly "syntactic," use representations or structured propositions as the objects of the attitudes. In what follows, we shall call all such treatments "representational." The Hangman's Paradox as presented by Montague and Kaplan is another example of how even a minimal amount of epistemic logic in conjunction with a self-referential attitude can lead to disaster. A more abstract formulation of the same self-referential attitude is exemplified in the "Knower's Paradox" and is perhaps best expressed in colloquial English by 'my negation is known' (we shall call this sentence the "Knower Sentence").

The usual moral drawn from these results is that a formally serious analysis of the attitudes should treat expressions of propositional attitudes, like modality, as predicates of propositions qua intensions, not as predicates of sentences or sentence-like entities. It is well-known how to provide a coherent logic of propositional attitudes with this approach. One of our main points, however, is that this "solution" is bought at far too dear a price: it leads neither to a satisfactory analysis of attitude reports nor to a general theory of attitudes. Since we think that these are desiderata of any theory of attitudes, we think that the standard "solution" is no solution at all, but rather refusal to take a theory of the attitudes seriously. A real solution to the Knower Paradox, one developed within a framework that leads to these desiderata, demands a thorough reanalysis of the logic of attitudes. We have both been developing such frameworks in other papers using discourse representation theory (Asher, 1984a, Asher, 1984b, Kamp, 1985a, Kamp, 1985b). In this paper, however, we shall only make use of very general features of those frameworks that are relevant to the discussion of the Knower Paradox.

1. Why the Knower Paradox is Still a Paradox

To review briefly, Kaplan and Montague (Kaplan & Montague, 1960) showed originally how the hangman paradox was in fact a real paradox: quite intuitively acceptable and limited principles of epistemic logic, together with a treatment of knowledge as a predicate of sentences in a language with sufficiently powerful syntactic resources, will yield an inconsistency. Later, Montague (Montague, 1963) showed the following disturbing result. Let T be an extension of Q^{β} (Robinson arithmetic relativized to the formula β whose only free variable is u), such that for any sentences ϕ and ψ whose standard names are $<\phi>$ and $<\psi>$ and some one place predicate of expressions 'K': (i) $\vdash_T K(<\phi>) \rightarrow \phi$, (ii) if ϕ is a theorem of logic then $\vdash_T K(<\phi>)$, (iii) $(K(<\phi>) \ \& \ K(<\phi \rightarrow \psi>)) \rightarrow K(<\psi>)$. Then T is inconsistent. Thus, the addition of the minimal rules and axioms for the knowledge predicate to weak arithmetic produce inconsistency.

Later again, Thomason (Thomason, 1980) showed that Montague's results could be extended to axiomatizations of knowledge (or any other operator which adopts (i) -(iii)) within a framework in which attitudes were not explicitly treated as predicates of sentences but as predicates of representations or, as current philosophical fashion would have it, structured propositions. Thomason also extended Montague's results to axiomatizations of belief. From a standard axiomatization of belief, he derives within T (which is defined as above, except that instead of the axioms for knowledge we add those for belief): $\vdash_T$ Bel('Q') --> Bel('S'), where S is any sentence whatsoever and Q is a single axiom of T. T becomes inconsistent if Bel('Q') and ¬Bel('S') is added to T. As Thomason says, "this seems enough to show a coherent theory of idealized belief as a syntactic predicate to be problematic ((Thomason, 1980), p. 393)."

It is important to understand why one can generalize the Knower Paradox to representational theories of attitudes. Higher order intensional logic itself becomes subject to epistemic paradoxes, when one adds to the theory an expression relation E with the axiom schema (1) below. We use here Montague's notation, letting S be a variable for sentences, while ^S stands for the proposition expressed by S (i.e., intension

of S) and <S> for the standard name of S.

(1) $\forall G$ (E(G, <S>) <--> G = ^S).

(1) seems innocuous but is in effect very powerful: it allows us to define a sentence predicate correlate for any sentential operator that is construed as a property of a proposition or intension (e.g., all modal and propositional attitude operators in intensional logic), such that we can prove that the axioms that hold for the operator hold in their transcribed form for the newly defined sentence predicates. This is all that is needed to get into trouble with the Knower Paradox. (1) is also sufficient to get into trouble in other ways as well: for instance, using (1) one can define within intensional logic the truth predicate for the language in which the logic is expressed.[1]

Intensional logic is thus subject to a subtle limitation, insofar as the expression relation is not definable within and cannot be consistently added to higher order intensional logic as it stands. But nothing in intensional logic demands that it be added either. Is the analogous relation between representations and sentences definable within a representational theory? Typically, representational theories must supply descriptive resources of the syntactic structure of the representational system sufficient to define some expression relation by means of which one can construct representations from sentences of natural language. Let us call this relation R. If R is recursive (as most representational theories suppose if we restrict the representational theory to indexical free sentences), then if T is an extension of arithmetic that incorporates also a standard axiomatization for a knowledge predicate defined on representations, there is a formula of T, E(x, y), such that $\vdash_T \forall x$ (E(x, a) <--> x = b) just in case R(a, b). But this is precisely the analogue of the expression relation in (1) and that we need to define the sentential predicate correlates for any epistemic predicates that apply to representations. Representational theories of the attitudes are thus in general open to the difficulties that Montague and Kaplan first made clear. Intensional logic is not, since the relation between intensions, which are functions from worlds to truth values, and sentences is not recursive.[2]

If the status of idealized belief or knowledge in representational theories is

problematic, however, there are very good reasons, we feel, for abandoning the treatment of the attitudes within intensional logic or "possible worlds semantics." First, the difficulties of treating attitude reports within the framework of intensional logic are well known. For intensional logic predicts that substitution of logical equivalents within belief contexts should preserve truth value. But this prediction is patently at odds with the linguistic facts concerning belief and other attitude reports.

Further, intensional logic must treat all propositional attitudes as closed under logical equivalence. We think that is a mistake even for the attitudes of belief and knowledge, since a theory burdened by such a requirement cannot lead to a satisfactory theory of attitude reports. But proponents of intensional treatments might reply that they are providing a semantics for idealized or rational belief and knowledge. The trouble with this reply is that there are attitudes like considering, entertaining or being aware of (the proposition that) that have no sensible, idealized counterparts as belief or knowledge may. The fundamental feature of these attitudes is that they are not closed under logical equivalence. There is nothing indefensible or odd (and lot that is right) in supposing that even ideally rational (though not omniscient) agents consider that A but not consider that B, even though A and B are logically equivalent. The arguments concerning the epistemic or doxastic indefensibility of the analogous situation with rationally ideal knowers and believers carry little weight with attitudes like considering, entertaining, or being aware of. But because within intensional logic $\hat{}A = \hat{}B$, one cannot distinguish between considering or entertaining that A and considering or entertaining that B. This gets the logic of these attitudes wrong, regardless of whether the agents are rationally idealized or not.

Yet another source of dissatisfaction with the possible worlds approach to the attitudes is that it is parochial. It cannot deal with the full range of attitudes that apply to human or other cognitive agents. Some of these are explicitly "sentential" attitudes. One can, we suppose, treat certain sentential attitudes within intensional logic, as long as their logic is not that of knowledge, belief or metaphysical modalities. But it also seems to us that there are sentential attitudes with such a logic, and they

merit treatment in a comprehensive theory of attitudes. For instance consider the attitudes, "rationally accepting S" or "justifiably accepting S", where S is a sentence. This attitude seems to have much in common with belief; in particular, they seem to share the same logic. The objects of acceptance, however, are sentences by hypothesis, *not* propositions construed as sets of possible worlds. A framework for attitudes that does not recognize cognitive agents as having attitudes (with a well-defined logic) toward sentences simply cannot say anything about the attitudes of rational acceptance or justifiable acceptance toward sentences.

Finally, all attitudes, even those that are standardly considered from the possible worlds framework like idealized belief and knowledge, have dimensions that one cannot even address, let alone investigate, within that framework. One such dimension is the possibility of having "self-referential" attitudes. Kaplan and Montague's reformulation of the hangman paradox is a concrete example of one. The example is perhaps somewhat artificial, but so too are liar sentences: the point is that any satisfactory theory of the attitudes must take account of these possibilities. Self-referential attitudes also have an honorable though not as hoary a history as that of Liar sentences. One could faithfully reformulate Descartes's cogito as: I believe (think) this belief (thought). Therefore, I exist. The notions of common knowledge and common belief furnish everyday examples of self-referential attitudes, as well as the basis of successful communication (Clark & Miller, 1981, Barwise, 1985). The resources of intensional logic simply do not permit the study of self-reference and self-referential attitudes. Since these exist within natural language, it is incumbent upon us to adopt a framework within which such attitudes can be studied.

The moral of the Knower Paradox is, we feel, that it remains an unsolved problem that any general theory of the attitudes must tackle. Representational theories arose in an attempt to answer the deficiencies of intensional logic's treatment of attitude reports. A particular kind of representational theory, we feel, provides a general framework for the analysis of the attitudes and their logic. Yet none of the proponents of such theories, as far as we know, have addressed in a systematic fashion the problems

involved with the Knower Paradox. We turn to this task in the next section.

2. Implications of the Knower Paradox for Representational Theories

The implications of the Knower Paradox for representational theories are complicated. In order to spell them out, we need to say a little bit more about what we think representational theories of attitudes consist of. In representational theories, attitudes are predicates of representations or sentence-like entities (and individuals too, though we shall ignore this second argument place here).[3] These representations are typically related to sentences by means of some (let us suppose recursive) expression or derivation relation, and they are also assigned an interpretation in a model by means of a truth or correctness definition. The rules for constructing representations from sentences also presuppose a "representational language" with well-formedness conditions. Such a "language" may, like other languages, may contain a variety of syntactical devices that permit various forms of self-reference.

Within such a theory, there are two distinct tasks concerning the attitudes. One is the semantics of attitude reports and the other is the logic, or perhaps better called the *dynamics*, of the attitudes themselves. The semantics of attitude reports determines truth conditions for attributions of attitudes or attitude reports. Within a representational theory, the semantics of attitude attrreports is carried out by first constructing, from the report, a complex representation R_1 which contains an attitude attribution, namely an attitude predicate taking another representation R_2 as an argument. The correctness of the report will depend upon whether R_2 matches some representation R_3 that is in the extension of the attitude predicate.[4] In determining the truth conditions for attitude reports, however, the semantics of attitude reports determines the structure and logic of the attitudes that subjects of such reports actually have. To avoid confusion, we shall call this logic the *logic of the attitude reports*; it is extremely weak in comparison with standard epistemic or doxastic logic.

The theory of attitude dynamics investigates what attitudes a rational agent could or ought to have, given some initial set of attitudes. That is, in studying the dynamics

of the attitudes, we are studying how a set of attitudes ought to evolve under the processes of rational inquiry and reflection. The standard principles of epistemic logic provide one way of spelling out in the form of closure principles the rational consequences of an agent's epistemic attitudes.

With this sketch we can now begin to detail some of the consequences of the Knower Paradox for a representational theory of the attitudes. First the paradoxes do not affect either the semantic theory or the logic of attitude reports. The reason why is very simple. Montague's results on the syntactic analysis of modalities require that the modalities be closed under logical consequence in T. But a major goal of an adequate semantics of attitude reports is precisely to rule out substitution of logically equivalent expressions (as well as many other deductively valid inferences) within attitude contexts in attitude reports and this in turn requires that the logic of attitude reports not be closed under logical consequence.[5] Thus, *regardless* of the syntactic resources of the representational language or the recursiveness of the expression relation (other factors crucial in deriving paradoxes within a representational theory), Montague's results do not affect the logic of the attitude reports or their semantics.

The second consequence of the Knower Paradox is that there are various ways of getting the paradox within a representational theory of attitude dynamics. By Thomason's general argument, a representational theory T that adopts standard axiomatizations for knowledge or belief will be open to the paradoxes, whenever the expression relation is representable within T. When, however, must that occur? Suppose that Q and T'-- the latter containing the standard axiomatization for a knowledge predicate in a representational language L-- are subtheories of T, and that there exists a representational theory T* that recursively defines the expression relation E. Then E must be representable within T and T is inconsistent. But if L is the language of T, then the facts about E are irrelevant, since L is a language capable of expressing arithmetic and thus contains enough syntactic devices to generate the paradoxes on its own. There are, however, other possibilities. Suppose Q is a subtheory of T* which is defined in a language L'. Since T* is a representational theory and

defines truth conditions for L expressions, T' is interpretable in T*. Call the image or translation of T' in L', I(T'): E is expressible as a formula of T*, and it follows that I(T') U T* is inconsistent, *regardless* of the syntactical resources available in L. This is a disturbing result, since if one believes T' to be an accurate characterization of attitude dynamics, the representationalist seems to be committed to the truth (or at least consistency) of I(T') U T*. Moreover, this result shows that the representationalist cannot escape the epistemic paradoxes in a wholly satisfactory way simply by restricting the range of syntactic devices in the representational language.

One could still avoid any paradoxes by restricting on the one hand the syntax of L and on the other the strength of T*. As long as T* were not an extension of arithmetic and the syntactic resources of L sufficiently weak, one could have a consistent theory. Moreover, L might still be resourceful enough to represent beliefs about arithmetic. But trying to escape the epistemic paradoxes in these ways is a mistake in our view. The limitation of the resources of the representational language to avoid paradoxes restricts the applicability of the expression relation. This is unsatisfactory, since such restrictions lead once again to a parochial analysis of the attitudes, which we think is as undesirable a consequence for a representational theory as it is for the intensional treatment. If the representational theory is designed to treat the full panoply of attitudes that can be expressed within natural language, then it must inevitably treat self-referential attitudes, including the paradoxical ones. Once we resolve to strengthen the representational theory so that we can treat paradoxical attitudes, however, the general argument of Thomason's as an argument against representational theories ceases to carry much weight. On our view, a representational theory of attitude dynamics is committed to a representational language strong enough to express paradoxical attitudes.

3. Representational Languages, their Syntactical Resources and Semantics

In this section we give one way of treating self-referential and paradoxical attitudes within a representational theory of attitude dynamics. We show how the

intension of an epistemic predicate can be built up or suitably revised from an initial possibly empty intension or a completely undefined one, modeling within the theory the processes of rational inquiry and reflection upon what one already knows or believes. These processes become less straightforward as the restrictions on the language are lifted. The approach we have adopted follows Gupta's and Herzberger's work on type-free semantics and the Liar paradox (Gupta, 1982, Herzberger, 1982a, Herzberger, 1982b). We feel that some type free approach is more promising than other alternatives for dealing with epistemic paradoxes, but we lack the space here to discuss these issues fully.[6]

Let us consider first a very simple representational language for the attitudes. Without too much loss of generality we may identify this with a standard first order language L augmented by an epistemic predicate K. Since K is a predicate of formulas, the language will also have names for formulas and perhaps for predicates and other types of L expressions. We will define an (intensional) *initial model* M for L to be an ordered quadruple $<\mathbf{W}, \mathbf{D}, \mathbf{R}, [\![]\!]>$, where **W** is a set of worlds, **D** is a nonempty set of objects (the domain of the model), **R** is a reflexive and transitive alternativeness relation defined on **W** and $[\![]\!]$ is an assignment to non-logical constants of the appropriate intensions. We will suppose that S, a set of sentences of L, is a subset of D and that the intensions of names of expressions are constant functions from worlds into particular expressions. We shall ignore the arguments of such functions and simply identify the semantic values of names of L-expressions with the appropriate L-expressions.

For the moment let us consider a simplified version of L in which the only names of L expressions are quotation names. By imposing the condition on models of L that all nonlogical predicates and function symbols of L be "sentence-neutral,"[7] we can show the existence of models for L, in which the standard axioms for knowledge are valid. The sentence-neutrality of a model M's interpretation function $[\![]\!]$ is defined, following Gupta, as follows:

(2)

1. $[\![\ ]\!]$ assigns to a quotation name '$\langle A \rangle$' the sentence A.

2. If b is not a quotation name then $[\![b]\!]$ does not belong to S.

3. If P is an n-ary predicate of L and $[\![b_i]\!]$ is in S, then $\langle [\![b_1]\!], \ldots, [\![b_i]\!], \ldots, [\![b_n]\!] \rangle \in [\![P]\!]$ iff $\forall x \in S \langle [\![b_1]\!], \ldots, x, \ldots, [\![b_n]\!] \rangle \in [\![P]\!]$.

4. If f is an n-ary function symbol of L, then the range of $[\![f]\!]$ does not contain sentences and further if $[\![b_i]\!], [\![b'_i]\!] \in S$, $[\![f]\!](\langle [\![b_1]\!], \ldots, [\![b_i]\!], \ldots, [\![b_n]\!] \rangle) = [\![f]\!](\langle [\![b_1]\!], \ldots, [\![b'_i]\!], \ldots, [\![b_n]\!] \rangle)$.

We shall call L-models with sentence neutral interpretation functions *S-acceptable.*

There are several ways to construct full L-models from initial L-models. The method we have adopted preserves all classical validities. First, we will extend the interpretation function of an initial model M with familiar recursive clauses to cover the logical symbols of L including K. We shall call the result a *standard model* for L. A standard model for L will assign a possibly empty set of formulas $U_i \subseteq S$ to 'K' at each world w_i. Let $\mathcal{U}$ be the set of ordered pairs $\langle w_i, U_i \rangle$; $\mathcal{U}$ is an intension for 'K'. We shall write $M + \mathcal{U}$ to denote a standard model that extends an initial model M and whose assignment to 'K' is the intension $\mathcal{U}$. It is obvious that not just any assignment to 'K' will do intuitively. If we are concerned solely with knowledge, we might wish to place certain constraints on $\mathcal{U}$-- at least that the formulas in the extension of 'K' at a world w always form a consistent set, ideally that they are all true and justified at w. Other attitudes would impose different constraints on $\mathcal{U}$; we shall leave out any constraints, however, to make our treatment more general.

There is another sort of constraint on what counts as an intuitively good assignment to an epistemic predicates like 'K'. We are attempting to cpature what a rational agent ought to know, given some initial set of propositions that he knows. It is in rationally reflecting upon some initial set of propositions known that the rational agent (in the limit) comes to know all that he ought to know. We shall model this process by a process of revision; beginning with from an initial model for L, M_0 and some arbitrary assignment $\mathcal{U}$ to 'K', we construct a series of models that each provides a slightly better approximation of the "rationally ideal" intension of K. Let M_0 be an

initial model for L and $M_0 + \mathcal{U}$ a standard model for L, where $\mathcal{U}$ is defined as above. Then, we define by transfinite recursion the *state of knowledge* at ordinal level α and at a world w with respect to an initial extension U at w as follows:

(3)

(i) $\mathbf{K}^0(w, U) = U$.

(ii) If $\alpha = \beta + 1$ is a successor ordinal,
$\mathbf{K}^\alpha(w, U) = \{s \in S: [\![s]\!]_w^{M_0 + \mathcal{K}^\beta(\mathcal{U})} = 1 \text{ for all } w'\mathbf{R}w\}$.

(iii) If α is a limit ordinal,
$\mathbf{K}^\alpha(w, U) = \{s \in S: \exists \beta < \alpha \cap_{\beta \leq \gamma < \alpha} \mathbf{K}^\gamma(w, U)\}$.

(iv) $\mathcal{K}^\alpha(\mathcal{U}) = \{<w_i, \mathbf{K}^\alpha(w_i, U_i)> : <w_i, U_i> \in \mathcal{U}\}$

$\mathbf{K}(w, U)$ furnishes an extension for 'K' From the definition, it is obvious that $\mathcal{K}^\alpha(\mathcal{U})$ furnishes a revised intension for 'K'. So each $M_0 + \mathcal{K}^\alpha(\mathcal{U})$ is a standard model for L.

The standard model incorporating the state of knowledge at α is at least as good a candidate for the extension of K as one incorporating the state of knowledge at β for $\beta < \alpha$. At each stage the intension for K is revised based on a reflection of what is already known at the previous stage. We can see these revisions at work in a situation where the original assignment to K is arbitrary (reflecting, obviously, not what is initially known but perhaps what the agent might have been told he or she knew). Now suppose that at a world w the original assignment is one that contains $\neg S$, but that $M_0 \models_w S$, for all w'Rw. This original assignment is deficient in terms of what we would normally take as the extension of knowledge; the principles of reflection we have adopted revise this assignment and make it better. The initial extension of M_0, $M_0 + \mathcal{K}^0(\mathcal{U})$, verifies $K(<\neg S>)$ at w, but by the first stage of reflection in w, $\neg S$ is no longer in the extension of K at w. $\mathbf{K}^1(w, U)$ contains S and so $M_0 + \mathcal{K}^1(\mathcal{U}) \models_w K(<S>)$. Supposing that $M_0 + \mathcal{K}^0(\mathcal{U}) \models_w K(<\neg S>)$ for all w' such that w'Rw, then $\mathbf{K}^1(w, U)$ will contain $K(<\neg S>)$ and $M_0 + \mathcal{K}^1(\mathcal{U}) \models_w K(<\neg S>)$. But this deficiency will be corrected at the next stage.

In any S-acceptable model, every formula in S has a unique "quotation degree"--

i.e., a unique depth of embedding of quote names within quote names. This implies that every formula in S is grounded in the sense of (Kripke, 1975). So in every standard extension of M_0 and at any world w, the extension of K will contain only "grounded" sentences. For S-acceptable models, we can prove that **K** is "semi-monotonic"-- that is:

Theorem 1: $\forall n\ \forall \alpha > n + 1$ (if the degree of $\phi = n$, then $\forall w\ \forall U, V \subseteq S$ ($\phi \in \mathbf{K}^{n+2}(w, U)$ iff $\phi \in \mathbf{K}^{\alpha}(w, V)$)).

The proof of this theorem essentially follows Gupta's proof of the "semi-monotonicity" of his truth revision scheme. In S-acceptable models **K** has a unique fixed point at ω for any world w, regardless of the character of the initial extension at w. We shall with Gupta call models with unique fixed points regardless of initial starting point for the intension of 'K' Thomasonian models.

In fact if $M_0 + \mathcal{U}$ is an S-acceptable L-model, then

Theorem 2: $M_0 + K^{\omega}(\mathcal{U})$ is a classical model for L in which the standard axioms for knowledge are true.

Before sketching the proof, let us list what we shall take to be standard axioms for knowledge:

- (K1) $K(\langle\phi\rangle) \dashrightarrow \phi$
- (K2) if ϕ is a theorem of logic, then $K(\langle\phi\rangle)$ is a theorem
- (K3) $(K(\langle\phi \dashrightarrow \psi\rangle)\ \&\ K(\langle\phi\rangle)) \dashrightarrow K(\langle\psi\rangle)$
- (K4) $K(\langle\phi\rangle) \dashrightarrow K(\langle K(\langle\phi\rangle)\rangle)$
- (K5) $K(\langle K(\langle\phi\rangle) \dashrightarrow \phi\rangle)$.

Proof: To show (K1) and suppose that for some w $[\![K(\langle\phi\rangle)]\!]_w^{M_0 + K^{\omega}(\mathcal{U})} = 1$. Then for some $\beta < \omega$, '$K(\langle\phi\rangle)$' $\in \mathbf{K}^{\gamma}(w, U)$, for $\beta \leq \gamma < \omega$. The only case of interest here is where β is a successor ordinal, i.e. $\beta = \delta + 1$. By our construction, $[\![\phi]\!]_{w'}^{M_0 + K^{\eta}(\mathcal{U})} = 1$ for all $\delta \leq \eta < \omega$ and for all w'Rw. Since R is reflexive, $[\![\phi]\!]_w^{M_0 + K^{\omega}(\mathcal{U})} = 1$. To prove (K4) suppose that for some w $[\![K(\langle\phi\rangle)]\!]_w^{M_0 + K^{\omega}(\mathcal{U})} = 1$. Again for some $\beta < \omega$, '$K(\langle\phi\rangle)$' $\in$

$\mathbf{K}^{\gamma}$(w, U), for $\beta \leq \gamma < \omega$. Suppose again that $\beta = \delta + 1$. Then $\phi \in \mathbf{K}^{\eta}$(w', U), for all $\delta \leq \eta < \omega$ and for all w'Rw. By our construction, however, 'K(<K(<ϕ>)>)' $\in \mathbf{K}^{\eta + 2}$(w', U), for all $\delta \leq \eta < \omega$ and for all w'Rw, and we are done. From the proofs of (K1) and (K4), it is now easy to give a proof of (K5). To prove (K2), it is sufficient to note that since $\mathcal{K}^{\alpha}(\mathcal{U})$ is total at every α, every theorem of logic ϕ is already in the extension of K at level 1 at any world w. Consequently, at the next stage K(<ϕ>) is in the extension of K at w. Finally, to prove (K3) note that if K(<ϕ --> ψ>) and K(<ϕ>) are in the extension of K at w, ϕ --> ψ and ϕ are true at all w' such that w'Rw at the previous stage and so ψ must be true at all w' at that stage too.

Slight modifications of this construction yield a Kripke style approach to a type free semantics for the attitudes. To do this we redefine a standard extension of an initial L-model as a partial model in which K is assigned an initially empty pair of sets of formulas at each world (an antiextension and an extension) and the valuation rules for the other logical vocabulary follow the strong Kleene valuation scheme. To obtain a *monotonic* instead of semi-monotonic definition of **K**, we simply redefine the clause in (3) for limit ordinals: if α is a limit ordinal then $\mathbf{K}^{\alpha}$(w, U) = $\bigcup_{\beta < \alpha} \mathbf{K}^{\beta}$(w, U). By the fixed model theorem (Feferman 1984), we conclude that there is a least partial model M_{inf} that is a fixed point of **K**. In any such model, all the standard axioms of epistemic logic except (K2) of theorem 3 will hold; not every truth of classical logic need be in the extension of the knowledge predicate in an L-model obtained by a Kripke-style construction. The revision procedure we sketched first seems preferable to us in dealing with the attitudes. It seems unsatisfactory to exclude certain instances of classical theorems from the extension of an attitude predicate. Any such instance supplies all the justification the rational agent could want for its being known or believed.[8]

How far can one generalize this construction from S-acceptable models in (2) without losing the result of theorem 1? By relaxing the conditions of sentence-neutrality so that M_0 interpretations can distinguish between grounded sentences but not ungrounded ones, we can show analogues for theorems 1 and 2. Let us call, following Gupta, models in which such restricted interpretations occur *K-acceptable* models. So if M_0 is a K-acceptable model, then the fixed point of K will also verify all the axioms of traditional epistemic logic.[9] One can also add multiple knowers, thus distinguishing

multiple knowledge predicates. Further, one can also add to L a logical predicate 'B', which stands for a belief predicate and construct successive approximations for both the intensions of 'B' and 'K' by means of **K** and an analogous operation **B** to get a model for belief and knowledge. One would need to complicate the notion of an initial L-model by adding an additional weakly reflexive and transitive alternativeness relation **R'** for belief and postulating that **R** is a subset of **R'**. These modifications produce some complications but no real difficulty. Models with multiple knowers and believers that are either S- or K-acceptable (with respect to all nonlogical predicates and function symbols) will be Thomasonian, and the standard axiomatizations for both knowledge and belief hold in them.

Where we do see an obstacle is if we add a truth predicate or some suitably concocted attitude predicates to the logical vocablulary of L. Following our strategy, we will revise an initial assignment of intensions to the truth predicate by means of an operation **T**, analogous to **K** and **B**. But the simultaneous revision of the intension for the truth predicate and the intension for some epistemic attitude predicate may produce well-known instabilities in the intension of the truth predicate. Consider the pair of sentences,

(5) Everything Tom believes is false

(6) Everything Doug believes is true.

Suppose Doug believes (5) at a particular world w and nothing else and that Tom believes (6) and nothing else at w. One can easily check that a model in which these conditions hold is one that will not yield a fixed extension for truth at w. In this case, **T** is not even semi-monotonic. **T** has not one fixed point but a "fixed cycle" (a sequence of several stages through which it cycles endlessly). Similar instabilities will be created if we admit as attitude predicates expressions like 'verify' and 'falsify' that would stand proxy for truth and falsity in (5) and (6).

Instabilities proper to the epistemic predicate K arise when we relax the conditions of sentence neutrality. It suffices to contravene 2.1 and 2.4 to construct the Knower Sentence within L. We need first to define quotation names in terms of quotation names of basic expressions of the language and a well-founded theory of concatenation (for a

definition of the latter see (Kamp, 1974)). Second, we need to allow function symbols to have an interpretation such that their ranges can contain sentences. We can then define a function in an L-model, sub(<A>, a, x) = the result of substituting x for every occurrence of singular term a in A. With these tools, we can construct the knower sentence and derive the paradox. These resources are stronger than what Gupta needs to derive liar sentences using simply sub and an ordinary theory of quotation.[10] Nevertheless, sub + ordinary quotation or sub + a relation of concatenation are still much weaker theories (they are decidable) than full arithmetic.

Obviously, models that permit the formulation of the Knower Sentence are neither S-acceptable nor K-acceptable. We will call such models *paradoxical* L models. In a paradoxical L-model, we can see a pattern of instability in the intension of K, analogous to that of the Liar sentence in Gupta's and Herzberger's work. Let us symbolize the Knower Sentence by means of a constant of L 'b' with the following interpretation in M_0: $[\![b]\!] = K(neg(b))$, where 'neg(b)' is the complex name that results from concatenating the name of sentential negation together with b (with our theory of quotation alluded to above we can do this). Suppose that the Knower sentence is not in the extension of K initially at any world w. So $M_0 + \mathcal{U} \models_w \neg K(neg(b))$, for all w'**R**w. By (3) and the interpretation of 'b' then, $K(neg(b)) \in \mathbf{K}^1(w, U)$. Since **R** is reflexive, there is a w'**R**w such that $[\![\neg K(neg(b))]\!]_w^{M_0 + \mathcal{K}^1(\mathcal{U})} = 0$ and so $\neg K(neg(b))$ does not belong to $\mathbf{K}^2(w, U)$, and so $M_0 + \mathcal{K}^2(U) \models_w \neg K(neg(b))$. But given that every world in M_0 has the same status with respect to this argument as does w, we have $M_0 + \mathcal{K}^2(U) \models_w \neg K(neg(b))$ for all w'**R**w. Consequently, $\neg K(neg(b)) \in \mathbf{K}^3(w, U)$ as we saw with $\mathbf{K}^1(w, U)$. What we see is that the Knower Sentence shuttles endlessly back and forth from being true in a model M_α to being false in $M_{\alpha+1}$. In paradoxical models, the intension for 'K' never stabilizes to a fixed point. Eventually, though, it settles down to a fixed cycle of possible values.

To get a general axiomatization for knowledge, we need to stipulate that some stage in the revision process constitutes the preferred intension of 'K' arrived at by rational reflection. We will choose a maximal limit ordinal stage of the revision process

of the original intension once the cyclic pattern has been achieved when dealing with paradoxical models or the maximal fixed point of the revision process should it exist. Given the behavior of sentences like the Knower Sentence, they will never show up in the extension of 'K' at some world w at limit ordinal stages. Logical truths will always be in the extension of 'K' at every world w in such maximal limit stages. Maximal limit ordinal stages will also verify (K1), (K3) and (K4) for the same reasons as in theorem 2. However, maximal limit stages do not verify axiom (K5) of theorem 3. Consider the Knower Sentence as formulated above and let $\mathcal{K}^{\alpha}(\mathcal{U})$ be a maximal limit ordinal stage of the revision process of $\mathcal{U}$. Although $[\![\mathrm{K}(\mathrm{neg}(\mathrm{b})) \dashrightarrow \neg\mathrm{K}(\mathrm{neg}(\mathrm{b}))]\!]_{w}^{M_0 + \mathcal{K}^{\alpha}(\mathcal{U})} = 1$ (veridicality), 'K(neg(b)) --> ¬K(neg(b))' does not belong to $\mathbf{K}^{\alpha}$(w', U) for any w' since 'K(neg(b)) is not stable in truth value for any final segment of α. Summing up then, we have:

> **Theorem 3:** Let $M_0 + \mathcal{K}^{\alpha}(\mathcal{U})$ be an L-model where $\mathcal{K}^{\alpha}(\mathcal{U})$ is a maximal limit ordinal stage of the revision process of $\mathcal{U}$. Then $M_0 + \mathcal{K}^{\alpha}(\mathcal{U})$ verifies axioms (K1) - (K4).

Thus, we get a consistent and nontrivial epistemic logic even if we allow paradoxical attitudes to be expressible. The derivation of Montague's original theorem is blocked insofar as that depends upon axiom (K5). To go back to the original difficulty that the representationalist was presented with, we can now consistently suppose that our epistemic theory T is an extension of Q^{β}. From this supposition it follows that $\vdash_T \phi \dashleftarrow\dashrightarrow \mathrm{K}(\langle\neg\phi\rangle)$. Pick some L-model such that the theorems of Q^{β} are true at every world w. Then presumably that $\phi \dashleftarrow\dashrightarrow \mathrm{K}\langle\neg\phi\rangle$ is in the extension of K at every world w after the first reflection. However, the formulas $\mathrm{K}\langle(\neg\phi)\rangle$ are unstable. There is no unique fixed point for **K** again here; for every world w there are two fixed point extensions for K-- one containing $\mathrm{K}(\langle\neg\phi\rangle)$ but where ϕ is true at w, the other not containing $\mathrm{K}(\langle\neg\phi\rangle)$ and where ϕ is false at w. At maximal limit ordinal stages neither ϕ nor $\neg\phi$ will be in the extension of 'K' at any world.

Theorem 3 still yields pretty nice results for epistemic logic. But this is only because at limit ordinal stages of the construction we have presented, nothing

paradoxical is ever known. This conclusion makes a certain amount of sense in the case of abstract formulation of the Knower sentence and of the "Godel sentence" for knowledge just examined in the previous paragraph. But in other cases, this conclusion seems too hasty. If one returns to Montague and Kaplan's formulation of the hangman paradox, it is very difficult to persuade oneself that the prisoner does not in fact know what the judge has decreed. But that of course is another paradoxical attitude. One might be tempted to conclude that in some instances agents really do have paradoxical knowledge. If we revise our definition of limit ordinal stages so that paradoxical sentences may still be in the extension of 'K' at some world w at maximal limit ordinal stages of the revision process, however, then the resulting epistemic logic is trivialized. The Knower Paradox then seems to point to a real limitation of logics of attitudes: one must give up either the contention that rational agents can have paradoxical attitudes or the applicability of any interesting axiomatization to these paradoxical attitudes.

To show how paradoxical attitudes affect the logic of knowledge, we first revise the definition in (3) of the limit ordinal stages of **K**.

(7) If α is a limit ordinal, then $\mathbf{K}^{\alpha}(w, U) = X \cup ((S \setminus Y) \cap U)$, where $X = \{s: (\exists \beta < \alpha)\ s \in \bigcap_{b \leq \gamma < \alpha} \mathbf{K}^{\gamma}(w, U)$ and $Y = \{s: (\exists \beta < \alpha)\ \neg s \in U_{\beta \leq \gamma < \alpha}\ (\mathbf{K}^{\gamma}(w, U))$."

We shall say that a formula ϕ is stably known (unknown) at α iff $\phi \in X(Y)$. The result of the limit ordinal stage of **K** is to subtract from U those things that are stably unknown at α and add to it the things that are stably known at α. We can show that a revised definition of **K** employing (7) yields a procedure for revising S- or K-acceptable models and turning them into Thomasonian ones (models with unique fixed points for the intension of 'K'). Let us call the results of applying the revision procedure based on (7) to initial L-models *Gupta models* and those resulting from applying (3) *Herzberger models*.

Gupta models that are not S- or K-acceptable can produce quite different results from non S- or K-acceptable Herzberger models. For instance, if the Knower Sentence or its negation are originally in U at some world w, they will be also in $\mathbf{K}^{\alpha}(w, U)$, where α is a limit ordinal. From this it follows that if $K^{\alpha}(U)$ is a maximal ordinal stage of the

revision process, $M_0 + \mathcal{K}^{\alpha}(\mathcal{U})$ will not always verify veridicality. For suppose that the negation of the Knower Sentence belongs to U at w. Then it must belong to $\mathbf{K}^{\alpha}$(w, U) and so $[\![\mathrm{K}(\mathrm{neg}(\mathrm{b}))]\!]_{w}^{M_0 + \mathcal{K}^{\alpha}(\mathcal{U})} = 1$. But if veridicality holds, then K(neg(b)) must be false at w. All theorems of logic will still be in $\mathbf{K}^{\alpha}$(w, U) and so (K2) will still hold in all maximal limit ordinal stages of Gupta models. But we can produce countermodels for (K3), (K4) and (K5). Let us for instance consider (K3), and let ϕ and ϕ --> $\psi \in$ U at world w. Suppose further that ϕ and ψ are unstable but "out of phase;" i.e. if $[\![\phi]\!]_{w}^{M_0 + \mathcal{K}^{\beta}(\mathcal{U})} = 1$, $[\![\psi]\!]_{w}^{M_0 + \mathcal{K}^{\beta}(\mathcal{U})} = 0$. Then ϕ --> ψ is also unstable (i.e., not stably false) at w. Given the definition in (7) ϕ and ϕ --> ψ belong to $\mathbf{K}^{\alpha}$(w, U), but ψ does not, which suffices to falsify (K3). Summing up these observations, we conclude:

> **Proposition 4:** Let $M_0 + \mathcal{K}^{\alpha}(\mathcal{U})$ be any model in which unstable sentences occur in the extension of 'K' at some world w. Then there is only a trivial axiomatization for the knowledge predicate.

References

C. Anderson: 1983, 'The Paradox of the Knower,' *Journal of Philosophy* 80, pp. 338-356.

N. Asher: 1984, 'Belief in Discourse Representation Theory,' forthcoming in *Journal of Philosophical Logic.*

N. Asher: 1984, 'A Typology for Cognitive Verbs,' forthcoming.

J. Barwise: 1985, 'The Situation in Logic III: Non Well-Founded Situations,' CSLI Report.

H. Clark & K. Miller: 1981, 'Definite Reference and Mutual Knowledge,' in A. Joshi *et al.*, ed., *Elements of Discourse Meaning*, Cambridge University Press, Cambridge, pp. 1-45.

S. Feferman: 1984, 'Toward Useful Type-Free Theories I,' *Journal of Symbolic Logic* 49, pp. 75-111.

A. Gupta: 1982, 'Truth and Paradox,' *Journal of Philosophical Logic* 11, pp. 1-60.

H. Herzberger: 1982, 'Notes on Naive Semantics,' *Journal of Philosophical Logic* 11, pp. 61-102.

H. Herzberger: 1982, 'Naive Semantics and the Liar Paradox,' *Journal of Philosophy* 79, pp. 479-497.

H. Kamp: 1974, 'The Formal Properties of 'Now',' *Theoria* , pp. 227-273.

H. Kamp: 1985, 'Fixation of Belief and Belief Reports,' Forthcoming.

H. Kamp: 1985, 'Context Thought and Communication,' *Proceedings of the Aristotelian Society* 85, pp. 239-261.

D. Kaplan & R. Montague: 1960, 'A Paradox Regained,' *Notre Dame Journal of Formal Logic* 1, pp. 79-90.

S. Kripke: 1975, 'Outline of a New Theory of Truth,' *Journal of Philosophy* 72, pp. 690-715.

R. Montague: 1963, 'Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability,' *Acta Philosophica Fennica* 16, pp. 153-167.

R. Thomason: 1980, 'A Note on Syntactical Treatments of Modality,' *Synthese* 44, pp. 391-395.

Notes

[1]Using Montague's notation again, we would define "true" as,
$\text{true}(\langle S \rangle) \leftrightarrow_{\text{def}} \exists G(G[] \ \& \ E(G, \langle S \rangle))$.

[2]Otherwise, at the very least, the theorems of first order logic would form a recursive set, which contradicts Church's theorem.

[3]By sentence-like entities we mean to include structured propositions-- sequences of properties and objects.

[4]The matching relation is in fact quite complex for the attitude predicates we have studied. See (Asher, 1984a, Asher, 1984b, Kamp, 1985a).

[5]We have each developed analyses of belief reports and begun to extend them to other attitudes, where substitution of logical equivalents is blocked for all "cognitive" type attitudes. In fact the empirical facts about believers make the principle of logical closure false.

[6]Several authors (most notably Kripke, Gupta and Herzberger) have offered what we think are persuasive reasons for choosing the type-free approach in dealing with the

Liar paradox. We feel that those reasons are equally persuasive in the context of the epistemic paradoxes, which leads us to advocate our approach over something like what Anderson (Anderson, 1984) proposes. Anderson distinguishes hierarchies of types within attitudes to solve the paradoxes. We find this approach to be objectionable for the same reasons as others have found Tarski's hierarchical approach to truth objectionable for natural languages.

[7]This term is due to Gupta.

[8]For a further discussion and comparison, see (Gupta, 1982, Herzberger, 1982a).

[9]In K-acceptable models not every member of S necessarily has a unique quotation degree; the fixed point of **K** may hence not be at ω but at a larger limit ordinal. But this does not affect the axiomatization for knowledge.

[10]As was noted by Montague, however, Tarski's truth axiom is considerably stronger than the axioms for various attitude predicates, and the syntactical resources needed to construct Liar sentences are intuitively weaker than those required to construct the sentences needed to derive the Knower Paradox.