Syntactical Treatments of Propositional Attitudes ⋆

Michael Morreau [a,c] Sarit Kraus [b,c]

[a] *Philosophy Department, University of Maryland, College Park, MD 20742, USA mimo@umiacs.umd.edu*

[b] *Department of Mathematics and Computer Science Bar-Ilan University, 52900 Ramat-Gan, Israel sarit@umiacs.umd.edu*

[c] *Institute for Advanced Computer Studies, University of Maryland, College Park, MD, 20742*

Abstract

Syntactical treatments of propositional attitudes are attractive to artificial intelligence researchers. But results of Montague [12] and Thomason [16] seem to show that syntactical treatments are not viable. They show that if representation languages are sufficiently expressive, then axiom schemes characterizing knowledge and belief give rise to paradox. De Rivières and Levesque [6] characterized a class of sentences within which these schemes can safely be instantiated. These sentences do not quantify over the propositional objects of knowledge and belief. We argue that their solution is incomplete, and extend it by characterizing a more inclusive class of sentences over which the axiom schemes can safely range. Our sentences do quantify over propositional objects.

1 Syntactical treatments of the attitudes

Knowledge, belief, desire, intention and other propositional attitudes have *propositional objects*. Concerning the nature of these objects there seem to be two main possibilities. One is that they are representations of some kind: perhaps sentences of a language, or pictures, or other devices for representing. Call the view that propositional attitudes are relations between individuals and sentences the *syntactical* view. The other possibility is that attitudes have as their objects not the representations themselves, but instead the represented states of affairs, meanings, or propositions. This is the *semantical* view.

⋆ This research was supported in part by the National Science Foundation grant number IIS9724937.

The syntactical view has seemed attractive to researchers in artificial intelligence (see chapters 2 and 8 of [5]). For one thing, it is natural to represent the knowledge and beliefs of a computer program by writing sentences representing facts that are known or believed. For another, syntactical treatments can be formalized in the classical predicate calculus, which is the *lingua franca* of knowledge representation [9,13,14]; since there is good theorem-proving technology for this language (e.g. [1–3]), such treatments lend themselves directly to implementations. Attractive as they may be, though, syntactical treatments face a serious theoretical difficulty. As we shall see, Richard Montague [12] shows that if the language in which known facts are represented is sufficiently expressive, then familiar axiom schemes characterizing knowledge are inconsistent. He argues that knowledge can be given a syntactical treatment only on pain of triviality. Richmond Thomason [16] argues under similar assumptions that a syntactical treatment also trivializes belief.

Montague and Thomason obtain their limiting results by instantiating their axiom schemes with self-referential formulas. In response, Jim de Rivières and Hector Levesque [6] show that syntactical treatments are viable if these axiom schemes are instantiated only with what they call *regular* formulas. This result is a good start. But syntactical treatments will be more useful if it can be extended, since regular formulas are rather restricted in their expressive power. For one thing, they do not quantify over the sentences that, on a syntactical treatment, are the propositional objects of the attitudes. Such quantification is needed to express matters that it may be useful for automated agent to know. To take just one example from natural language processing, it is widely recognized that participants in cooperative dialogue will say only what they believe to be true, and will make their conversational contributions as informative as possible [8]. That a speaker conforms to these maxims of conversation is naturally expressed, assuming a syntactical treatment, by quantifying over the sentences he utters. It is not expressed by any regular formula.

Our result, which extends that of de Rivières and Levesque, shows that such quantification does not pose a special problem for syntactical treatments. We show that the schemes of Montague and Thomason can without paradox be instantiated with formulas that quantify over the propositional objects of the attitudes. We shall proceed as follows. First, in section 2, we discuss the difficulty with syntactical treatments of propositional attitudes. After introducing a language with self-reference, and the axiom schemes of Montague and Thomason, we reproduce their negative results. In section 3, we present the limited positive result of de Rivières and Levesque concerning regular formulas. We also show, by elaborating on the example from natural language processing, that it can be useful to instantiate the axiom schemes with formulas that quantify over propositional objects. In section 4, we show how to extend the result of section 3 to cover such formulas. Finally, in section 5, we present a further extension.

2 Knowledge and Belief

To bring out the difficulties with syntactical treatments, the language for representing attitudes must have resources for self-reference. These are provided here by arithmetization. [1] We introduce the following language of arithmetic.

Definition 1 (The language of arithmetic) $\mathcal{L}_0$ *is a first-order language which includes a set of individual variables* $(x, y, ...)$*; a set of individual constants* $(a, b, c, c_1, c_2, ...)$ *including the special constant symbol* 0 *(zero); a set of predicate symbols* $(P, R, ...)$ *including the special unary predicate* $I\!N$*; and function symbols* S *(the successor function),* $+$ *(addition) and* $\times$ *(multiplication). The terms and atomic formulas of* $\mathcal{L}_0$ *are defined in the usual way, as are the formulas formed with the connectives* $\wedge$, $\vee$, $\rightarrow$, $\neg$, *and quantifiers* $\exists$ *and* $\forall$.

In the following, we shall consider only languages which extend $\mathcal{L}_0$. The following definition concerns what can be regarded as terms denoting formulas:

Definition 2 (Gödel Numerals) *Let* $\mathcal{L}$ *be a language and* $\varphi \in \mathcal{L}$*. Letting* $g(\varphi)$ *be the Gödel number of* φ *(according to some fixed scheme),* $\lceil \varphi \rceil$ *is the* $\mathcal{L}_0$*-numeral* $\overbrace{S...}^{g(\varphi) times}$ 0.

We can regard $\lceil \varphi \rceil$ as a term standing for the formula φ. Now the following definition introduces the relation symbol with which we shall express propositional attitudes as relations to formulas.

Definition 3 (The language of the attitudes) $\mathcal{L}_\alpha$ *is the language whose vocabulary is that of* $\mathcal{L}_0$*, together with the special dyadic relation symbol* α.

We shall refer to the formulas of $\mathcal{L}_\alpha$ using lower case Greek letters: ϕ, ψ, χ etc. Below, α will represent either knowledge or belief. Thus, $\alpha_a(\lceil \varphi \rceil)$ expresses that the agent a has attitude α toward sentence φ; just which attitude α represents will be clear from the context. When we are not interested in them we shall suppress the agents, writing simply $\alpha(\lceil \varphi \rceil)$. [2] Now we can characterize knowledge and belief as follows:

Knowledge Axioms

[1] Nothing depends essentially on this. The paradoxes discovered by Montague and Thomason and our responses to them can be discussed while obtaining self reference using other techniques.

[2] In $\mathcal{L}_\alpha$ we can, for example, write $\forall x(utter(speaker, x) \rightarrow \alpha_{speaker}(x))$. The intended interpretation of this sentence is that the speaker utters only what he believes to be true.

Interpreting $\alpha(\ulcorner\varphi\urcorner)$ as the claim that the agent knows φ, the following axiom scheme expresses the *facticity* that distinguishes knowledge from belief and other propositional attitudes:

i $\alpha(\ulcorner\varphi\urcorner) \rightarrow \varphi$.

Other schemes express that the agent knows certain things, and that its knowledge is closed under certain logical operations. Thus:

ii $\alpha(\ulcorner\alpha(\ulcorner\varphi\urcorner) \rightarrow \varphi\urcorner)$ expresses that the agent knows that what is known is true. The axiom scheme:

iii $\alpha(\ulcorner\varphi\urcorner)$, for each φ that is a logically valid formula of the first-order predicate calculus, expresses that the agent knows logic. Finally:

iv $\alpha(\ulcorner\varphi \rightarrow \psi\urcorner) \rightarrow [\alpha(\ulcorner\varphi\urcorner) \rightarrow \alpha(\ulcorner\psi\urcorner)]$ expresses the assumption that the agent's knowledge is closed under *modus ponens.*

To facilitate the coming discussion, for any set S of formulas let $\mathcal{K}(S)$ be the set of all instantiations of schemes (i)–(iv) with formulas φ and ψ of S.

Belief Axioms

Interpreting $\alpha(\ulcorner\varphi\urcorner)$ as the claim that the agent *believes* φ, scheme (i) is unacceptable. Beliefs can be false. Belief is characterized by axiom schemes (ii)-(iv) above together, perhaps, with *positive introspection*:

v $\alpha(\ulcorner\varphi\urcorner) \rightarrow \alpha(\ulcorner\alpha(\ulcorner\varphi\urcorner)\urcorner)$.

This scheme expresses that the agent is cognizant of his own state of belief: if he believes something, then he believes that he believes it. Let $\mathcal{B}(S)$ be the set of all instantiations of schemes (ii)-(v) with formulas of S. With this background in place, we are in a position to see the difficulty with syntactical treatments.

2.1 The difficulty with syntactical treatments

Generalizing Tarski's [15] demonstration of the undefinability of truth, first Montague [12] and then Thomason [16] proved results that seem to show that propositional attitudes can be given syntactical treatments only on pain of triviality. Let Q be a single axiom for Robinson's arithmetic (see, for example, [11]). For any (perhaps complex) monadic predicate β of $\mathcal{L}_\alpha$, let Q^β be the result of relativizing quantifiers to β.[3] For any set Γ of formulas, let Γ^s be the set of all closed formulas, or *sentences*, of Γ. Montague proved the following theorem about syntactical treatments of knowledge:

Theorem 4 (Montague[4]) *Let Γ be a first-order theory such that $\mathcal{K}(\mathcal{L}^s_\alpha) \subseteq \Gamma$, and such that for some monadic predicate β, $Q^\beta \in \Gamma$. Then Γ is inconsistent.*

Montague's result seems to show that someone who supposes arithmetic is true cannot coherently suppose an agent's knowledge is characterized by the relevant schemes, if knowledge is given a syntactical treatment. Thomason extended Montague's result to the case of belief:

Theorem 5 (Thomason) *Let Γ be a first-order theory, $\mathcal{B}(\mathcal{L}^s_\alpha) \subseteq \Gamma$ and, for some β, $Q^\beta \in \Gamma$. Then for all ψ, $\Gamma \vdash \alpha(\ulcorner Q^\beta \urcorner) \rightarrow \alpha(\ulcorner \psi \urcorner)$.*

Thomason's result seems to show that someone who supposes arithmetic is true cannot suppose that an agent believes arithmetic, without supposing that the agent is perfectly credulous. He must suppose that such an agent believes *everything*, if belief is given a syntactical treatment.

The upshot for knowledge representation is that, on a syntactical treatment of propositional attitudes, one cannot add to a knowledge base all instantiations of axiom schemes (i)–(iv), or of (ii)-(v). To do so will, under unobjectionable further assumptions, make the knowledge base inconsistent. Our response to these results will be to guarantee consistency by limiting the scope of the axiom schemes (i)-(v). We shall allow them to be instantiated only within a limited part of the object language. That is, we shall characterize a somewhat limited set of unproblematic sentences, within which these schemes can safely be instantiated.

3 A Partial Solution

The first response of this kind came from de Revières and Levesque [6]. They showed that the schemes characterizing knowledge and belief can safely range over what they called *regular* sentences. We shall in this section define this set and state their result. To this end, we shall first slightly extend $\mathcal{L}_\alpha$.

[3] If $\beta(w)$ is a monadic predicate, the *relativization of φ to β*, φ^β, is defined for arbitrary φ by recursion: if φ is atomic, φ^β is φ; $(\neg\varphi)^\beta$ is $\neg\varphi^\beta$; $(\varphi \rightarrow \psi)^\beta$ is $\varphi^\beta \rightarrow \psi^\beta$, and so on; finally, replacing β, if need be, by an alphabetic variant that does not contain the variable x, $(\forall x \varphi)^\beta$ is $\forall x(\beta\{x/w\} \rightarrow \varphi^\beta)$ and $(\exists x \varphi)^\beta$ is $\exists x(\beta\{x/w\} \wedge \varphi^\beta)$.

Re-definition 1 (The language of the attitudes) *We add to the earlier* $\mathcal{L}_\alpha$, *for each* $i > 1$, *a relation symbol* α^i *of arity* $i + 1$. *(We now write "*α^1*" for "*α*.")*

Informally, one argument place is for the agent, one is for (the Gödel numeral of) a formula, and the rest are for the parameters of the formula. The intended interpretation of α^i_a is best illustrated by example: informally speaking, $\alpha^1_{user}(\ulcorner \textit{departure-time(train,3:15)} \urcorner)$ means that the user believes that the train departs at 3:15. The intended interpretation of $\exists x \alpha^2_{user}(\ulcorner \textit{departure-time(train,x)} \urcorner, x)$ is that the user believes that the train departs at some time.

Definition 6 (Regular formulas) *Let* $\mathcal{L}_\alpha$ *be a first-order language as defined above. Then the regular formulas,* R*, are the smallest subset of* $\mathcal{L}_\alpha$ *such that:*

(i) *Any atomic formula of* $\mathcal{L}_0$ *is a regular formula.*
(ii) *If* φ *and* ψ *are regular formulas and* x *is a variable, then* $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \rightarrow \psi$, $\neg\varphi$, $\forall x \varphi$ *and* $\exists x \varphi$ *are regular formulas.*
(iii) *If* φ *is a regular formula whose free variables are just* $x_1, ..., x_k$,[5] *and* a *is a constant symbol, then* $\alpha^{k+1}_a(\ulcorner \varphi \urcorner, x_1, ..., x_k)$ *is a regular formula.*

We can now state the result of des Rivières and Levesque [6].

Theorem 7

(i) *For some consistent theory* Γ *of* $\mathcal{L}_\alpha$, $\{Q^{\mathbb{N}}\} \cup \mathcal{K}(R^s) \subseteq \Gamma$.
(ii) *For some theory* Γ *of* $\mathcal{L}_\alpha$ *and sentence* ψ *of* $\mathcal{L}_\alpha$, $\Gamma \not\vdash \alpha(\ulcorner Q^{\mathbb{N}} \urcorner) \rightarrow \alpha(\ulcorner \psi \urcorner)$ *and* $\{Q^{\mathbb{N}}\} \cup \mathcal{B}(R^s) \subseteq \Gamma$.

Thus, the problems discovered by Montague and Thomason arise if their axiom schemes are instantiated with *all* sentences of $\mathcal{L}_\alpha$, but they do not arise if these schemes are instantiated with just the *regular* sentences.

As far as it goes this result is very nice, but one might hope that it can be extended from the regular sentences to some more inclusive set. This is because quite straightforward matters cannot be expressed using regular sentences. Sentences that quantify over the propositional objects of knowledge and belief are a case in point. For example, to provide a query-answering system with the knowledge that the user believes that the system believes what the system says, we might wish to write in the knowledge base some sentence such as: $\alpha_{user}(\ulcorner \forall x(utter(system, x) \rightarrow \alpha_{system}(x) \urcorner)$. While the syntax of $\mathcal{L}_\alpha$ allows us to write this sentence, though, as things stand there is no guarantee that the system will be able to use it to make desirable inferences. This is because the sentence $\forall x(utter(system, x) \rightarrow \alpha_{system}(x))$ quantifies over propo-

[5] In this case we write '$\varphi(x_1, ..., x_k)$'.

sitional objects, and is therefore not regular. The result of des Rivières and Levesque does not allow us to instantiate the axioms schemes of Montague and Thomason with it, or with any sentence of which it is a subformula. For a simple example of an inference that for this reason cannot be made, suppose that in response to a query the system utters: *departure-time(train,3:15)* (or an equivalent in English). Afterwards, the system ought to be able to infer that the user believes that the system believes what it has just said: $\alpha_{user}(\ulcorner\alpha_{system}(\ulcorner \textit{departure-time(train,3:15)}\urcorner)\urcorner)$. Now, it can be seen that this sentence can be inferred from $\alpha_{user}(\ulcorner\forall x(utter(system,x)\rightarrow\alpha_{system}(x))\urcorner)$ together with the fact that the user believes that the system has uttered the sentence in question, $\alpha_{user}(\ulcorner utter(system,\ulcorner \textit{departure-time(train,3:15)}\urcorner)\urcorner)$, provided the schemes of Thomason are instantiated in the system's knowledge base with a sentence of which $\forall x(utter(system,x)\rightarrow\alpha_{system}(x))$ is a subformula. (Which instantiations are needed is left as an exercise.) But the theorem of de Revières and Levesque offers no guarantee that the necessary instantiations can be written in the knowledge base without introducing inconsistency.

In the following section 4 we characterize a set, more inclusive than the regular sentences, which includes sentences that quantify over propositional objects of knowledge and belief. Then we generalize Theorem 7, showing that the schemes characterizing knowledge and belief can without paradox be instantiated within this set. This generalization, together with a further generalization in section 5, does guarantee that the instantiations needed for the above inference can consistently be written in a knowledge base.

4 A less partial solution

To begin we must once again slightly extend $\mathcal{L}_\alpha$. We add to $\mathcal{L}_\alpha$ the new monadic predicate symbols T and $\mathcal{P}$. Also, in addition to the variables $x, x_1, x_2, ..., y, ...$, we add a countable stock of new variables $X, X_1, X_2, ..., Y, ...$

Intuitively speaking, the symbol $\mathcal{P}$ will pick out (the Gödel numbers of) sentences of $\mathcal{L}_\alpha$; T will pick out the true sentences. The following definition concerns the promised set, to which we shall generalize Theorem 7.

Definition 8 (Regular formulas with Propositional Quantification (RPQ)) *The RPQ formulas are the smallest set of $\mathcal{L}_\alpha$ such that:*

(i) Any atomic formula of $\mathcal{L}_0$ is an RPQ formula, as is $T(X)$, for any (new) variable X.
(ii) If φ and ψ are RPQ formulas and x is an (old) variable, then $\varphi\wedge\psi$, $\varphi\vee\psi$, $\varphi\rightarrow\psi$, $\neg\varphi$, $\forall x\varphi$ and $\exists x\varphi$ are RPQ formulas.
(iii) If $\varphi(x_1,...,x_k,X_1,...,X_l)$ is an RPQ formula, and a is a constant, then

$\alpha_a^{k+l+1}(\lceil\varphi\rceil, x_1, ..., x_k, X_1, ..., X_l)$ is an RPQ formula.

(iv) If φ is an RPQ formula then $\forall X(\mathcal{P}(X) \rightarrow \varphi)$ and $\exists X(\mathcal{P}(X) \wedge \varphi)$ are also RPQ formulas, where X is a (new) variable.

It is easy to see that all regular formulas are *RPQ*. The *RPQ* part of $\mathcal{L}_\alpha$ is more expressive than the regular part, though, since there are *RPQ* formulas that, intuitively speaking, quantify over propositional objects. For example:

$$\forall X(\mathcal{P}(X) \rightarrow (utter(system, X) \rightarrow \alpha^2_{system}(\lceil T(X)\rceil, X)))$$

is an *RPQ* representation of the claim, discussed in the previous section, that the system only says what it takes to be true. This sentence can be further embedded. Thus:

$$\alpha^1_{user}(\lceil\forall X(\mathcal{P}(X) \rightarrow (utter(system, X) \rightarrow \alpha^2_{system}(\lceil T(X)\rceil, X)))\rceil)$$

is an *RPQ* representation of the claim that the user believes that the system does this. Again:

$$\exists X(\mathcal{P}(X) \wedge \alpha^2_{user}(\lceil T(X)\rceil, X) \wedge \neg T(X))$$

is an *RPQ* representation of the claim that there are among the user's beliefs some that are false. This too can be further embedded. Thus:

$$\alpha^1_{user}(\lceil\exists X(\mathcal{P}(X) \wedge \alpha^2_{user}(\lceil T(X)\rceil, X) \wedge \neg T(X))\rceil)$$

represents the user's belief in his doxastic fallibility.

We shall now show that the axiom schemes characterizing knowledge and belief can safely be instantiated with *RPQ* sentences.

Theorem 9

(i) For some consistent theory Γ of $\mathcal{L}_\alpha$, $\{Q^{\mathbb{N}}\} \cup \mathcal{K}(RPQ^s) \subseteq \Gamma$.

(ii) For some theory Γ of $\mathcal{L}_\alpha$ and sentence ψ of $\mathcal{L}_\alpha$, $\Gamma \nvdash \alpha(\lceil Q^{\mathbb{N}}\rceil) \rightarrow \alpha(\lceil\psi\rceil)$ and $\{Q^{\mathbb{N}}\} \cup \mathcal{B}(RPQ^s) \subseteq \Gamma$.

Theorem 9 extends Theorem 7 from the regular part of $\mathcal{L}_\alpha$ to the more inclusive *RPQ* part of $\mathcal{L}_\alpha$. The upshot is that although according to Theorem 1 and 2 we cannot unrestrictedly instantiate axiom schemes (i)-(v) within a knowledge base, we can safely instantiate these schemes with *RPQ* formulas, which quantify over propositional objects. This is not sanctioned by de Revières and Levesque's earlier result. It allows an automated agent to reason about knowledge and belief in useful ways, as illustrated at the end of section 3. Notice that the consistent theory Γ of theorem 4 can contain, besides instantiations of the axiom schemes with *RPQ* sentences, other sentences that are not *RPQ*. The point is not that the *knowledge base* should contain only *RPQ* sentences. It is just the instantiations of the axiom schemes that should be restricted.

We shall prove Theorem 9 by first establishing a correspondence between the RPQ formulas and a language for second-order modal logic [7]. First, though, we need several more definitions. The first of these introduces the relevant modal language.

Definition 10 (The language of $\Box$ with propositional quantification) We extend $\mathcal{L}_0$ to a multi-modal language $\mathcal{L}_\Box$ by adding, to the vocabulary of $\mathcal{L}_0$, for each individual constant a of $\mathcal{L}_0$, a modal operator $\Box_a$. We also add the same new variables $X, Y, ...$ that we included in $\mathcal{L}_\alpha$. The atomic formulas of $\mathcal{L}_\Box$ are just those of $\mathcal{L}_0$, together with the variables $X, Y, ...$ Complex formulas are formed using $\neg$, $\wedge$, $\Box$, $\forall x$, $\exists x$, $\forall X$ and $\exists X$ in the usual way.

We shall refer to the formulas of $\mathcal{L}_\Box$ by p, q, r etc. This is to distinguish them from those of $\mathcal{L}_\alpha$, to which we refer using φ, ψ, χ etc. Examples of formulas of $\mathcal{L}_\Box$ are: $\Box P(x)$, $\forall X \forall x (P(x) \wedge X)$, $\exists X \Box X$. We shall now turn to the interpretation of $\mathcal{L}_\Box$. Note, first, that whereas $X, Y, ...$ in interpretations of $\mathcal{L}_\alpha$ were allowed to range over the same domain as the variables $x, y, ...$, in interpretations of $\mathcal{L}_\Box$ the variables $X, Y, ...$ will range over *propositions*. These will not be elements of the domain over which the variables $x, y, ...$ range. Instead they will be sets of possible worlds of the models within which we shall interpret $\mathcal{L}_\Box$.

Definition 11 (Models of $\mathcal{L}_\Box$) A model for $\mathcal{L}_\Box$ is a tuple $< \mathcal{W}, \mathcal{R}, \Delta, \mathcal{I} >$ where $\mathcal{W}$ is a non-empty set of possible worlds, $\mathcal{R}$ is a function assigning to each individual constant of $\mathcal{L}_0$ a binary relation on $\mathcal{W}$, and Δ is an infinite domain of individuals. $\mathcal{I}$ is an interpretation function which provides each nonlogical symbol of $\mathcal{L}_\Box$ with a suitable interpretation. Thus $\mathcal{I}$ associates with each world and individual constant of $\mathcal{L}_\Box$ an element of Δ, and with each world and relation symbol a suitable relation on Δ.

We assume that the individual constants are rigid designators: for each $v, w \in \mathcal{W}$, $\mathcal{I}(w, c) = \mathcal{I}(v, c)$. Therefore, we can simply write: "$\mathcal{I}(c)$". We assume furthermore that for each a, $\mathcal{R}_a$ is an equivalence relation. Thus, all of the $\mathcal{L}_\Box$ models considered below are S5 models [4]. We now turn to the evaluation of formulas. A *variable assignment* h into a model $< \mathcal{W}, \mathcal{R}, \Delta, \mathcal{I} >$ associates with each lower case variable an element of Δ (i.e., $h(x) \in \Delta$), and with each upper case variable a set of possible worlds (i.e., $h(X) \subseteq \mathcal{W}$). The truth conditions of formulas at possible worlds relative to variable assignments are, leaving out the standard clauses, as follows:

Definition 12 (Truth conditions in $\mathcal{L}_\Box$ models)

- $\mathcal{M}, w \models \Box_a p[h]$ *iff* $\forall v \in \mathcal{W}$ *such that* $R_a(w, v)$, $\mathcal{M}, v \models p[h]$;
- $\mathcal{M}, w \models X[h]$ *iff* $w \in h(X)$;
- $\mathcal{M}, w \models \forall X p[h]$ *iff for all* $W' \subseteq \mathcal{W}$, $\mathcal{M}, w \models p[h(X/W')]$; *here* $h(X/W')$ *is the variable assignment which differs from* h*, if at all, in that* $h(X/W')(X) =$

W'.

Propositions, we said, are sets of possible worlds. Letting $\mathcal{M}$ be a model of $\mathcal{L}_\Box$ and letting p be a sentence of $\mathcal{L}_\Box$, the *proposition expressed by* p in $\mathcal{M}$, written $\|p\|_\mathcal{M}$, is $\{w|w \in \mathcal{W},\ \mathcal{M}, w \models p\}$. We now single out those models in which *every* proposition is expressed by some or other sentence. An $\mathcal{L}_\Box$ model $< \mathcal{W}, \mathcal{R}, \Delta, \mathcal{I} >$ is said to be *definite* just in case for every $W' \subseteq \mathcal{W}$, W' is expressed by some sentence. Notice that there are a lot of definite models. For example, any model with just a single possible world is definite. Indeed, any $\mathcal{L}_\Box$ model $< \mathcal{W}, \mathcal{R}, \Delta, \mathcal{I} >$ that is not definite can be extended into a definite model of a more expressive language. We simply add to $\mathcal{L}_\Box$ a set of new propositional constants whose cardinality is at least that of $2^\mathcal{W}$, and interpret these constants in such a way as to distinguish between each of the propositions in this set.

The following translation function now establishes the correspondence between $\mathcal{L}_\alpha$ and $\mathcal{L}_\Box$ that we shall exploit in our result. It maps $\mathcal{L}_\Box$ formulas onto those $\mathcal{L}_\alpha$ formulas that are, modulo substitution of sentences for propositions, their equivalents.

Definition 13 (Translation function) *The translation function* * *from* $\mathcal{L}_\Box$ *to* $\mathcal{L}_\alpha$ *is defined, for each* $p, q \in \mathcal{L}_\Box$*, as follows:*

- $p^* = p$, *where* p *is an atomic formula of* $\mathcal{L}_0$.
- $X^* = T(X)$.
- $(p \wedge q)^* = p^* \wedge q^*$, $(p \vee q)^* = p^* \vee q^*$, $(p \rightarrow q)^* = p^* \rightarrow q^*$, $(\neg p)^* = \neg p^*$.
- $(\forall x p)^* = \forall x p^*$ *and* $(\exists x p)^* = \exists x p^*$.
- $(\forall X p)^* = \forall X(\mathcal{P}(X) \rightarrow p^*)$ *and* $(\exists X p)^* = \exists X(\mathcal{P}(X) \wedge p^*)$.
- $(\Box_a p(x_1, ..., x_k, X_1, ..., X_l))^* = \alpha_a^{k+l+1}(\lceil p^* \rceil, x_1, ..., x_k, X_1, ..., X_l)$.

Now, the *RPQ* formulas are just translations of formulas of $\mathcal{L}_\Box$:

Lemma 14 $\varphi \in \mathcal{L}_\alpha$ *is an RPQ formula iff for some* $p \in \mathcal{L}_\Box$, $\varphi = p^*$.

Proof: Straightforward induction on the length of φ. ∎

To state the main lemma we use to prove Theorem 9, we first define, for any variable assignment h into an $\mathcal{L}_\Box$ model, an associated set of variable assignments into models of $\mathcal{L}_\alpha$.

Definition 15 (The variable assignments associated with h) *Let* h *be a variable assignment of* $\mathcal{L}_\Box$ *into a definite* $\mathcal{L}_\Box$ *model* $\mathcal{M}$. *The set* A_h *of variable* *assignments (into a certain $\mathcal{L}_\alpha$ model to be defined below) is:*

$$\left\{ f \;\middle|\; \begin{array}{l} \textit{For any lower case variable } x: \;\; f(x) = h(x) \\ \textit{For any upper case variable } X: f(X) = g(p^*), \textit{for some sentence } p \in \mathcal{L}_\Box \\ \hfill \textit{such that } \|p\|_\mathcal{M} = h(X) \end{array} \right\}$$

Now we can state and prove the main lemma needed for the proof of Theorem 9. It concerns the construction of an $\mathcal{L}_\alpha$ model which is equivalent, modulo *, to any given possible world of a given definite $\mathcal{L}_\Box$ model.

Lemma 16 *For any definite $\mathcal{L}_\Box$ model, $\mathcal{M} =< \mathcal{W}, \mathcal{R}, \Delta, \mathcal{I} >$, and for any world $w \in \mathcal{W}$, there is a model $\mathcal{N}$ of $\mathcal{L}_\alpha$ such that for any variable assignment h into $\mathcal{M}$, and for any $p \in \mathcal{L}_\Box$, $\mathcal{M}, w \models p[h] \Longleftrightarrow \forall f \in A_h, \mathcal{N} \models p^*[f]$.*

Using our earlier observation that any model can be expanded into a definite model of a more expressive language, it is possible to strengthen this lemma by dropping the requirement that $\mathcal{M}$ be definite. We shall not go further into this since it is not needed for our present purpose.

Proof of Lemma 16 Consider a given definite model $\mathcal{M} =< \mathcal{W}, \mathcal{R}, \Delta, \mathcal{I} >$, and world $w \in \mathcal{W}$. Let the domain of $\mathcal{N}$ be Δ. The interpretation function I of $\mathcal{N}$ is as follows:

(i) Where σ is any predicate symbol other than $\mathcal{P}, T$ or α^i, and where σ is an individual constant, $I(\sigma) = \mathcal{I}(w, \sigma)$.
(ii) $I(T) = \{g(p^*) \mid p$ is a sentence of $\mathcal{L}_\Box$ and $\mathcal{M}, w \models p\}$.
(iii) $I(\mathcal{P}) = \{g(p^*) \mid p$ is a sentence of $\mathcal{L}_\Box\}$.
(iv) $I(\alpha^{k+l+1}) = \{(g(p^*), \delta_1, ..., \delta_k, g(q_1^*), ..., g(q_l^*)) \mid$ (letting $x_1, ..., x_k, X_1, ..., X_l$ be all the variables of p) $\mathcal{M}, w \models \Box p[(x_1/\delta_1), ..., (x_k/\delta_k), (X_1/\|q_1\|), ..., (X_l/\|q_l\|)]\}$.

The proof can now proceed with a straightforward induction on the length of p. We illustrate by considering the cases $p = X$, $p = \forall X q$ and $p = \Box q$.

$p = X$: We show that $\mathcal{M}, w \models X[h] \Longleftrightarrow \forall f \in A_h, \mathcal{N} \models T(X)[f]$.
$\Rightarrow$: Suppose, $\mathcal{M}, w \models X[h]$. Consider $f \in A_h$. By definition 15, $f(X) = g(q^*)$, for some sentence $q \in \mathcal{L}_\Box$ such that $\|q\|_\mathcal{M} = h(X)$. Since $\mathcal{M}, w \models X[h]$, $w \in \|q\|_\mathcal{M}$, i.e., $M, w \models q$. Thus by the definition of I, $\mathcal{N} \models T(X)[f]$.
$\Leftarrow$: Suppose $\forall f \in A_h, \mathcal{N} \models T(X)[f]$. Since $\mathcal{M}$ is a definite model, there is a sentence $q \in \mathcal{L}_\Box$ such that $\|q\|_\mathcal{M} = h(X)$. Choose $f \in A_h$ such that $f(x) = g(q^*)$. Since $\mathcal{N} \models T(X)[f]$, $g(q^*) \in I(T)$. From the definition of I (since formulas are recoverable from their Gödel numbers), $\mathcal{M}, w \models q$. Thus, $w \in \|q\|_\mathcal{M} = h(X)$ and we can conclude that $\mathcal{M}, w \models X[h]$.
$p = \forall X q$: We show that $\mathcal{M}, w \models \forall X q[h] \Longleftrightarrow \forall f \in A_h, \mathcal{N} \models \forall X(\mathcal{P}(X) \rightarrow q^*)[f]$.

$\Rightarrow$: Suppose $\mathcal{M}, w \models \forall X q[h]$. Let $f \in A_h$ and let $\delta \in \Delta$. Suppose $\mathcal{N} \models \mathcal{P}(X)[f(X/\delta)]$. Then, by the definition of I, there is a sentence $r \in \mathcal{L}_\Box$ such that $\delta = g(r^*)$. $\mathcal{M}, w \models q[h(X/\|r\|_\mathcal{M})]$. Also, $f(X/\delta) \in A_{h(X/\|r\|_\mathcal{M})}$, since $f \in A_h$. Thus, by the induction hypothesis, $\mathcal{N} \models q^*[f(X/\delta)]$. Since δ is arbitrary, $\mathcal{N} \models \forall X(\mathcal{P}(X) \to q^*)[f]$. Since f is arbitrary, for each $f \in A_h$, $\mathcal{N} \models \forall X(\mathcal{P}(X) \to q^*)[f]$.

$\Leftarrow$: Suppose $\forall f \in A_h$, $\mathcal{N} \models \forall X(\mathcal{P}(X) \to q^*)[f]$. Let $P \subseteq \mathcal{W}$. Let $f' \in A_{h(X/P)}$. There is a sentence $r \in \mathcal{L}_\Box$ such that $\|r\|_\mathcal{M} = P$, and such that for some $f \in A_h$, $f' = f(X/g(r^*))$. By the definition of $I(\mathcal{P})$, $\mathcal{N} \models \mathcal{P}(X)[f(X/g(r^*))]$, so by the assumption, $\mathcal{N} \models q^*[f']$. Thus, since f' was chosen arbitrarily, for each $f' \in A_{h(X/P)}$, $\mathcal{N} \models q^*[f']$. By the induction hypothesis, $\mathcal{M}, w \models q[h(X/P)]$. Finally, since P was chosen arbitrarily, $\mathcal{M}, w \models \forall X q[h]$.

$p = \Box q$: We assume, without loss of generality, that the free variables of q are $x_1, ..., x_k$, $X_1, ...X_l$. We show that
$\mathcal{M}, w \models \Box q[h] \iff \forall f \in A_h, \mathcal{N} \models \alpha^{k+l+1}(\ulcorner q^* \urcorner, x_1, ..., x_k, X_1, ...X_j)[f]$.

$\Rightarrow$: Suppose $\mathcal{M}, w \models \Box q[h]$. Choose any $f \in A_h$. By definition 15, $f(x_i) = h(x_i)$, and we can choose $r_1, ..., r_l \in \mathcal{L}_\Box$ such that $f(X_i) = g(r_i^*)$ and $\|r_i\|_\mathcal{M} = h(X_i)$. Then, $\mathcal{M}, w \models \Box q[(x_1/f(x_1)), ..., (x_k/f(x_k)), (X_1/\|r_1\|), ..., (X_l/\|r_l)\|]$. By the definition of I, $(g(q^*), f(x_1), ..., f(x_l), g(r_1^*), ..., g(r_l^*)) \in I(\alpha^{k+l+1})$, so $\mathcal{N} \models \alpha^{k+l+1}(\ulcorner q^* \urcorner, x_1, ..., x_k, X_1, ...X_l)[f]$. Since f was chosen arbitrarily, $\forall f \in A_h, \mathcal{N} \models \alpha^{k+l+1}(\ulcorner q^* \urcorner, x_1, ..., x_k, X_1, ...X_l)[f]$.

$\Leftarrow$: Suppose $\forall f \in A_h, \mathcal{N} \models \alpha^{k+l+1}(\ulcorner q^* \urcorner, x_1, ..., x_k, X_1, ...X_l)[f]$. Choose any $f \in A_h$. Then $(g(q^*), f(x_1), ..., f(x_k), f(X_1), ..., f(X_l)) \in I(\alpha^{k+l+1})$. By Definition 15, choose $r_1, ..., r_l$ such that $\|r_i\|_\mathcal{M} = h(X_i)$ and $f(X_i) = g(r_i^*)$. By the definition of I, $\mathcal{M}, w \models \Box q[(x_1/f(x_1)), ..., (x_k/f(x_k)), (X_1/\|r_1\|), ..., (X_l/\|r_l\|)]$. By definition 15, $h(x_i) = f(x_i)$, and by choice of r_i, $h(X_i) = \|r_i\|$. Thus $\mathcal{M}, w \models \Box q[h]$. This completes the proof of Lemma 16. ∎

We are now in a position to prove Theorem 9.

Proof of Theorem 9. We shall show, for any definite S5 model $\mathcal{M}$ in each of whose possible worlds $Q^{\mathbb{N}}$ holds true, and for any world w of $\mathcal{M}$, that $Th(\mathcal{N})$ is a suitable theory Γ. Here, $\mathcal{N}$ is the $\mathcal{L}_\alpha$ model gotten from $\mathcal{M}$ and w by the construction of Lemma 16. $Th(\mathcal{N})$ is the set of all sentences that are true in this model.

Clearly $Th(\mathcal{N})$ is consistent. Furthermore, since $(Q^{\mathbb{N}})^* = Q^{\mathbb{N}}$, it is clear by Lemma 16 that $Q^{\mathbb{N}} \in Th(\mathcal{N})$. To complete the proof it remains to be shown that $\mathcal{K}(RPQ^s) \subseteq Th(\mathcal{N})$, that $\mathcal{B}(RPQ^s) \subseteq Th(\mathcal{N})$, and that, say, $\alpha(\ulcorner Q^{\mathbb{N}} \urcorner) \to \alpha(\ulcorner 0 = 1 \urcorner) \notin Th(\mathcal{N})$.

The first axiom scheme of $\mathcal{K}$ is $\alpha(\ulcorner \varphi \urcorner) \to \varphi$. We shall show, for each $\varphi \in RPQ^s$, that $\alpha(\ulcorner \varphi \urcorner) \to \varphi \in Th(\mathcal{N})$. The proofs for the remaining schemes of $\mathcal{K}$ and the schemes of $\mathcal{B}$ are analogous. Let $\varphi \in RPQ^s$. By Lemma 14 there is some sentence $p \in \mathcal{L}_\Box$ such that $\varphi = p^*$. Since $\mathcal{M}$ is an S5 model, $\mathcal{M}, w \models \Box p \rightarrow p$. So by definition 13 and Lemma 16, $\alpha\lceil p^* \rceil \rightarrow \lceil p^* \rceil \in Th(\mathcal{N})$.

Finally, we show that $\alpha(\lceil Q^{I\!N} \rceil) \rightarrow \alpha(\lceil 0 = 1 \rceil) \notin Th(\mathcal{N})$. This is a direct consequence of Lemma 16, together with the fact that while $\mathcal{M}, w \not\models \Box(Q^{I\!N}) \rightarrow \Box(0 = 1)$, $(\Box(Q^{I\!N}) \rightarrow \Box(0 = 1))^* = \alpha(\lceil Q^{I\!N} \rceil) \rightarrow \alpha(\lceil 0 = 1 \rceil)$. This completes the proof of Theorem 9. ∎

Our result can be extended to languages with modal operators other than $\Box$, with modal conditional connectives such as that of Lewis [10], and other constructions that can be analyzed within possible-worlds models. In the following section, we shall extend it instead to a larger fragment of the same language $\mathcal{L}_\alpha$.

5 An Extension

In the previous sections we characterized sentences that are well behaved, in the sense that the axiom schemes corresponding to knowledge and belief can without paradox range over them. Regular sentences are well-behaved in this sense and so are *RPQ* sentences. But these characterizations are partial, insofar as there are more inclusive sets of sentences within which these schemes can safely be instantiated. In this section we shall extend Theorem 9 to one such set. The effect is to extend even further the range of sentences with which axiom schemes (i)–(v) can be instantiated within a knowledge base.

One motivation for this extension is that useful inferences appear to rely on instantiating the schemes of Montague and Thomason with certain sentences that are not *RPQ*. For example, the inference discussed at the end of section 3 relies on instantiating with $\alpha^2(\lceil T(X) \rceil, \lceil \textit{departure-time(train,3:15)} \rceil)$, if the fact that the system utters only sentences that it believes is given the *RPQ* representation of section 4. Another motivation lies in the observation that some *RPQ* sentences have the same informal interpretations as non-RPQ sentences. Both $\alpha(\lceil \textit{departure-time(train,3:15)} \rceil)$, which is *RPQ*, and $\alpha^2(\lceil T(X) \rceil, \lceil \textit{departure-time(train,3:15)} \rceil)$, which is not, mean, intuitively speaking, that the agent believes that the train leaves at 3:15. Now, whereas Theorem 4 sanctions the instantiation of the axiom schemes of Montague and Thomason with the former sentence, it does not sanction their instantiation with the latter. Our extension of Theorem 4 sanctions both. It respects such informal equivalences more fully, and it illustrates a technique that, we expect, can be used to obtain further extensions of our main result. A first step is to generalize the translation function of Definition 9 by introducing the following mapping:

Definition 17 (Translation map) *The translation mapping* ^{e} *from* $\mathcal{L}_\Box$ *to* $\mathcal{L}_\alpha$ *is defined, for each* $p, q \in \mathcal{L}_\Box$*, as follows:*

- $p^e = \{p\}$, *where* p *is an atomic formula of* $\mathcal{L}_0$.
- $X^e = \{T(X)\}$.
- $(p \wedge q)^e = \{\hat{p} \wedge \hat{q} \mid \hat{p} \in p^e, \hat{q} \in q^e\}$, $(p \vee q)^e = \{\hat{p} \vee \hat{q} \mid \hat{p} \in p^e, \hat{q} \in q^e\}$, $(p \rightarrow q)^e = \{\hat{p} \rightarrow \hat{q} \mid \hat{p} \in p^e, \hat{q} \in q^e\}$, $(\neg p)^e = \{\neg \hat{p} \mid \hat{p} \in p^e\}$.
- $(\forall x p)^e = \{\forall x \hat{p} \mid \hat{p} \in p^e\}$ *and* $(\exists x p)^e = \{\exists x \hat{p} \mid \hat{p} \in p^e\}$.
- $(\forall X p)^e = \{\forall X(\mathcal{P}(X) \rightarrow \hat{p}) \mid \hat{p} \in p^e\}$ *and* $(\exists X p)^e = \{\exists X(\mathcal{P}(X) \wedge \hat{p}) \mid \hat{p} \in p^e\}$.
- $(\Box_a(p))^e = \{\alpha_a^1(\lceil \hat{p} \rceil) \mid \hat{p} \in p^e\} \cup \{\alpha_a^2(\lceil T(X) \rceil, \lceil \hat{p} \rceil) \mid \hat{p} \in p^e\}$.
- $(\Box_a(p(x_1, ..., x_k, X_1, ..., X_l)))^e = \{\alpha_a^{k+l+1}(\lceil \hat{p} \rceil, x_1, ..., x_k, X_1, ..., X_l) \mid \hat{p} \in p^e\}$.

It is easy to see that for any $p \in \mathcal{L}_\Box$, $p^* \in p^e$. That is, the mapping ^{e} is a generalization of the function *. We next show that it is well behaved.

Lemma 18 *For any* $p, q \in \mathcal{L}_\Box$*, if* $p \neq q$*, then* $p^e \cap q^e = \emptyset$*, i.e., a formula is recoverable from each of its translations.*

Proof: The proof is by induction on the length of p and q.

Base case: If p and q are different atomic formulas, then $p^e \cap q^e = \{p\} \cap \{q\} = \emptyset$.

Inductive case: If the logical connectives of p and q are different, the claim is clear from the definition of ^{e}, with respect to the connectives.

Suppose the hypothesis is true for $p_1, p_2, q_1, q_2 \in \mathcal{L}_\Box$.

- $p = p_1 \wedge p_2$ and $q = q_1 \wedge q_2$: if $p \neq q$, then either (i) $p_1 \neq q_1$, or (ii) $p_2 \neq q_2$. Letting $\varphi \in p^e$ and $\psi \in q^e$, we have $\varphi = \varphi_1 \wedge \varphi_2$ and $\psi = \psi_1 \wedge \psi_2$, for some $\varphi_1 \in p_1^e$, $\varphi_2 \in p_2^e$, $\psi_1 \in q_1^e$ and $\psi_2 \in q_2^e$. If (i), then by the induction hypothesis $p_1^e \cap q_1^e = \emptyset$, so $\varphi_1 \neq \psi_1$. If (ii), then by the induction hypothesis $p_2^e \cap q_2^e = \emptyset$, so $\varphi_2 \neq \psi_2$. Either way, $\varphi \neq \psi$.

  The proof for the other connectives is similar. We will present the case of $\Box$.
- Let $p = \Box p_1$, $q = \Box q_2$, $\varphi \in p^e$ and $\psi \in q^e$. If $p \neq q$ then $p_1 \neq q_1$. Thus, we have for some $\varphi_1 \in p_1^e$, and $\psi_1 \in p_2^e$, (i) $\varphi = \alpha_a^1(\lceil \varphi_1 \rceil)$ or (ii) $\varphi = \alpha_a^2(\lceil T(X) \rceil, \lceil \varphi_1 \rceil)$ and (a) $\psi = \alpha_a^1(\lceil \psi_1 \rceil)$ or (b) $\psi = \alpha_a^2(\lceil T(X) \rceil, \lceil \psi_1 \rceil)$. It is clear that if (i) and (b) or (ii) and (a) then the claim is true. Suppose (a) and (i). Since $p_1 \neq q_1$, then by the induction hypothesis, $\varphi_1 \neq \psi_1$ and the claim is clear. Similarly when (ii) and (b). ∎

One simple consequence of the above lemma is:

Corollary 19 *If* $p^* \in q^e$*, then* $p = q$.

We are now ready to define an extension of RPQ within which the schemes characterizing knowledge and belief can safely be instantiated:

Definition 20 $E = \{\varphi \mid$ *for some* $p \in \mathcal{L}_\Box, \varphi \in p^e\}$.

We first note that E really extends RPQ. $RPQ \subseteq E$ since for any $p \in \mathcal{L}_\Box$, $p^* \in p^e$. But $RPQ \neq E$ since for any sentence $p \in \mathcal{L}_\Box$, $\alpha^2(\lceil T(X) \rceil, \lceil p^* \rceil) \in E$, but $\alpha^2(\lceil T(X) \rceil, \lceil p^* \rceil) \notin RPQ$. We shall now extend Theorem 9 to E, showing that the axiom schemes characterizing knowledge and belief can safely be instantiated with E sentences.

Theorem 21

(i) *For some consistent theory* Γ *of* $\mathcal{L}_\alpha$, $\{Q^{\mathbb{N}}\} \cup \mathcal{K}(E^s) \subseteq \Gamma$.
(ii) *For some theory* Γ *of* $\mathcal{L}_\alpha$ *and sentence* ψ *of* $\mathcal{L}_\alpha$, $\Gamma \not\vdash \alpha(\lceil Q^{\mathbb{N}} \rceil) \rightarrow \alpha(\lceil \psi \rceil)$ *and* $\{Q^{\mathbb{N}}\} \cup \mathcal{B}(E^s) \subseteq \Gamma$.

To state the main lemma that we use to prove Theorem 21, we must first revise Definition 15.

Definition 22 *[The variable assignments associated with h] Let h be a variable assignment of* $\mathcal{L}_\Box$ *into a definite* $\mathcal{L}_\Box$ *model* $\mathcal{M}$. *The set* A^e_h *of variable assignments associated with h is:*

$$\left\{ f \;\middle|\; \begin{array}{l} \textit{For any lower case variable } x: \;\; f(x) = h(x) \\ \textit{For any upper case variable } X: f(X) = g(\varphi), \textit{for some sentence } \varphi \in \mathcal{L}_\alpha \textit{ such that} \\ \hfill \textit{for some } p \in \mathcal{L}_\Box, \varphi \in p^e \textit{ and } \|p\|_{\mathcal{M}} = h(X) \end{array} \right\}$$

We next prove a lemma similar to Lemma 16.

Lemma 23 *For any definite* $\mathcal{L}_\Box$ *model* $\mathcal{M}(=<\mathcal{W}, \mathcal{R}, \Delta, \mathcal{I}>)$, *and for any world* $w \in \mathcal{W}$, *there is a model* $\mathcal{N}$ *of* $\mathcal{L}_\alpha$ *such that for any variable assignment h into* $\mathcal{M}$, *and for any* $p \in \mathcal{L}_\Box$, $\mathcal{M}, w \models p[h] \Longleftrightarrow \forall \varphi \in p^e, f \in A^e_h, \mathcal{N} \models \varphi[f]$.

Proof: The model $\mathcal{N}$ is similar to the one constructed in the proof of Lemma 16. The only difference is in the interpretation function I of $\mathcal{N}$, which is now defined as follows:

(i) Where σ is any predicate symbol other that $\mathcal{P}, T$ and α^i, or σ is an individual constant, $I(\sigma) = \mathcal{I}(w, \sigma)$, as in the proof of Lemma 16.
(ii) $I(T) = \{g(\varphi) \mid$ for some sentence $p \in \mathcal{L}_\Box$, $\varphi \in p^e$ and $\mathcal{M}, w \models p\}$.
(iii) $I(\mathcal{P}) = \{g(\varphi) \mid$ for some sentence $p \in \mathcal{L}_\Box$, $\varphi \in p^e\}$.
(iv) $I(\alpha^{k+l+1}) = \{(g(\varphi), \delta_1, ..., \delta_k, g(\psi_1), ..., g(\psi_l)) \mid$ for some $p, q_1, ..., q_l \in \mathcal{L}_\Box$, (letting $x_1, ..., x_k, X_1, ..., X_l$ be all the variables of p) $\varphi \in p^e$, $\psi_1 \in q^e_1$, ...,

$\psi_l \in q_l^e$ and $\mathcal{M}, w \models \Box p[(x_1/\delta_1), ..., (x_k/\delta_k), (X_1/\|q_1\|), ..., (X_l/\|q_l\|)]\}$.

The proof proceeds by induction on the length of p. We shall illustrate by considering the cases $p = X$, $p = \forall X q$ and $p = \Box q$.

$p = X$: $\varphi \in X^e$ only if $\varphi = T(X)$. We show that $\mathcal{M}, w \models X[h] \Longleftrightarrow \forall f \in A_h^e, \mathcal{N} \models T(X)[f]$.

$\Rightarrow$: Suppose, $\mathcal{M}, w \models X[h]$. Consider $f \in A_h^e$. By definition 22, $f(X) = g(\psi)$, for some $\psi \in \mathcal{L}_\alpha$ such that for some $q \in \mathcal{L}_\Box$, $\psi \in q^e$ and $\|q\|_\mathcal{M} = h(X)$. Since $\mathcal{M}, w \models X[h]$, $M, w \models q$. Thus by the definition of I, $\mathcal{N} \models T(X)[f]$.

$\Leftarrow$: Suppose $\forall f \in A_h^e, \mathcal{N} \models T(X)[f]$. Since $\mathcal{M}$ is a definite model, there is a sentence $q \in \mathcal{L}_\Box$ such that $\|q\|_\mathcal{M} = h(X)$. Since $q^* \in q^e$, it is possible to choose $f \in A_h^e$ such that $f(x) = g(q^*)$. Since $\mathcal{N} \models T(X)[f]$, $g(q^*) \in I(T)$. From the definition of I there is $r \in \mathcal{L}_\Box$ such that $q^* \in r^e$ and $\mathcal{M}, w \models r$. By Corollary 19, $q = r$. So, $\mathcal{M}, w \models q$ and we can conclude that $\mathcal{M}, w \models X[h]$.

$p = \forall X q$: We show that $\mathcal{M}, w \models \forall X q[h] \Longleftrightarrow \forall f \in A_h^e, \psi \in q^e, \mathcal{N} \models \forall X(\mathcal{P}(X) \to \psi)[f]$.

$\Rightarrow$: Suppose $\mathcal{M}, w \models \forall X q[h]$. Let $f \in A_h^e$, $\psi \in q^e$, $\delta \in \Delta$. Suppose $\mathcal{N} \models \mathcal{P}(X)[f(X/\delta)]$. Then, by the definition of I, there are sentences $r \in \mathcal{L}_\Box$, $\chi \in \mathcal{L}_\alpha$, such that $\delta = g(\chi)$, $\chi \in r^e$. By assumption, $\mathcal{M}, w \models q[h(X/\|r\|_\mathcal{M})]$. Also, $f(X/\delta) \in A^e_{h(X/\|r\|_\mathcal{M})}$, since $f \in A_h^e$. Thus, by the induction hypothesis, since $\psi \in q^e$, $\mathcal{N} \models \psi[f(X/\delta)]$. So, since δ is arbitrary within Δ, $\mathcal{N} \models \forall X(\mathcal{P}(X) \to \psi)[f]$. Since f and ψ are arbitrary within A_h^e and q^e, for each $f \in A_h^e$ and $\psi \in q^e$, $\mathcal{N} \models \forall X(\mathcal{P}(X) \to \psi)[f]$.

$\Leftarrow$: Suppose $\forall f \in A_h^e$, $\psi \in q^e$, $\mathcal{N} \models \forall X(\mathcal{P}(X) \to \psi)[f]$. Let $P \subseteq \mathcal{W}$. Let $f' \in A^e_{h(X/P)}$. Choose $r \in \mathcal{L}_\Box$ such that $\|r\|_\mathcal{M} = P$, $\chi \in \mathcal{L}_\alpha$ and $f \in A_h^e$ such that $f' = f(X/g(\chi))$. By the definition of $I(\mathcal{P})$, $\mathcal{N} \models \mathcal{P}(X)[f(X/g(\chi))]$. Choose $\psi \in q^e$. By the assumption, $\mathcal{N} \models \psi[f(X/g(\chi))]$, so $\forall \psi \in q^e$, $\mathcal{N} \models \psi[f']$. Thus, since f' was chosen arbitrarily within $A^e_{h(X/P)}$, $\forall f' \in A^e_{h(X/P)}$, $\psi \in q^e$, $\mathcal{N} \models \psi[f']$. By the induction hypothesis, $\mathcal{M}, w \models q[h(X/P)]$. Finally, since P was chosen arbitrarily, $\mathcal{M}, w \models \forall X q[h]$.

$p = \Box q$: There are two cases, (1) and (2), depending on whether p is a sentence.

(i) Assume that the free variables of q are $x_1, ..., x_k, X_1, ... X_l$. We show that $\mathcal{M}, w \models \Box q[h] \Longleftrightarrow \forall f \in A_h^e, \psi \in q^e \mathcal{N} \models \alpha^{k+l+1}(\lceil \psi \rceil, x_1, ..., x_k, X_1, ... X_j)[f]$.

$\Rightarrow$: Suppose $\mathcal{M}, w \models \Box q[h]$. Choose any $f \in A_h^e$. By definition 22, $f(x_i) = h(x_i)$, and we can choose $r_1, ..., r_l \in \mathcal{L}_\Box$ and $\chi_i \in r_i^e$ such that $f(X_i) = g(\chi_i)$ and $\|r_i\|_\mathcal{M} = h(X_i)$. Then,
$\mathcal{M}, w \models \Box q[(x_1/f(x_1)), ..., (x_k/f(x_k)), (X_1/\|r_1\|), ..., (X_l/\|r_l)\|]$. Choose $\psi \in q^e$. By the definition of I, $(g(\psi), f(x_1), ..., f(x_l), g(\chi_1), ..., g(\chi_l)) \in I(\alpha^{k+l+1})$, so $\mathcal{N} \models \alpha^{k+l+1}(\lceil \psi \rceil, x_1, ..., x_k, X_1, ... X_l)[f]$. Since ψ was chosen arbitrarily in q^e and f was chosen arbitrarily in A_h^e, $\forall f \in A_h^e, \psi \in q^e \mathcal{N} \models \alpha^{k+l+1}(\lceil \psi \rceil, x_1, ..., x_k, X_1, ... X_l)[f]$.

$\Leftarrow$: Suppose $\forall f \in A_h^e, \psi \in q^e, \mathcal{N} \models \alpha^{k+l+1}(\lceil \psi \rceil, x_1, ..., x_k, X_1, ... X_l)[f]$. Then, since $q^* \in q^e$, $\mathcal{N} \models \alpha^{k+l+1}(\lceil q^* \rceil, x_1, ..., x_k, X_1, ... X_l)[f]$. Choose

any $f \in A_h^e$. Then $(g(q^*), f(x_1), ..., f(x_k), f(X_1), ..., f(X_l)) \in I(\alpha^{k+l+1})$. By Definition 22, choose $r_1, ..., r_l \in \mathcal{L}_\Box$, $\chi_1, ..., \chi_l \in \mathcal{L}_\alpha$ such that $\chi_i \in r_i^e$, $\|r_i\|_\mathcal{M} = h(X_i)$ and $f(X_i) = g(\chi_i)$. By the definition of I (since formulas are recoverable from their translations),
$\mathcal{M}, w \models \Box q[(x_1/f(x_1)), ..., (x_k/f(x_k)), (X_1/\|r_1\|), ..., (X_l/\|r_l\|)]$. Thus $\mathcal{M}, w \models \Box q[h]$.

(ii) Assume q is a sentence. We first show that:

$$\text{For all } \psi \in \mathcal{L}_\alpha, \mathcal{N} \models \alpha^1(\lceil\psi\rceil) \Leftrightarrow \mathcal{N} \models \alpha^2(\lceil T(X)\rceil, \lceil\psi\rceil). \tag{1}$$

We have: $\mathcal{N} \models \alpha^1(\lceil\psi\rceil) \Leftrightarrow$ (by Definition of I) for some $q \in \mathcal{L}_\Box$ such that $\psi \in q^e$, $\mathcal{M}, w \models \Box q \Leftrightarrow$ for some $q \in \mathcal{L}_\Box$ such that $\psi \in q^e$, $\mathcal{M}, w \models \Box X[X/\|q\|_\mathcal{M}] \Leftrightarrow (g(T(X)), g(\psi)) \in I(\alpha^2) \Leftrightarrow \mathcal{N} \models \alpha^2(\lceil T(X)\rceil, \lceil\psi\rceil)$. Thus, to show that $\mathcal{M}, w \models \Box q \iff \forall\psi \in q^e\ \mathcal{N} \models \alpha^1(\lceil\psi\rceil)$ and $\mathcal{N} \models \alpha^2(\lceil T(X)\rceil, \lceil\psi\rceil)$, it is sufficient to show that $\mathcal{M}, w \models \Box q[h] \iff \mathcal{N} \models \alpha^1(\lceil\psi\rceil)$.

$\Rightarrow$: Suppose $\mathcal{M}, w \models \Box q$. Let $\psi \in q^e$. By definition of I, $g(\psi) \in I(\alpha^1)$, so $\mathcal{N} \models \alpha^1(\lceil\psi\rceil)$.

$\Leftarrow$: Suppose $\forall\psi \in q^e, \mathcal{N} \models \alpha^1(\lceil\psi\rceil)$. Let $\psi \in q^e$. In particular, since $q^* \in q^e$, $\mathcal{N} \models \alpha^1(\lceil q^*\rceil)$, so $g(q^*) \in I(\alpha^1)$. Then there is $r \in \mathcal{L}_\Box$ such that $q^* \in r^e$ and $\mathcal{M}, w \models \Box r$. By Corollary 19, $q = r$ and so $\mathcal{M}, w \models \Box q$. ∎

We are now in a position to prove Theorem 21.

Proof of Theorem 21. The proof is similar to that of Theorem 9. Here, $\mathcal{N}$ is the $\mathcal{L}_\alpha$ model gotten from $\mathcal{M}$ and w by the construction of Lemma 23. We show, for example, that $\forall\varphi \in E^s$, $\mathcal{N} \models \alpha^1(\lceil\varphi\rceil) \to \varphi$. Let $\varphi \in p^e$ for some $p \in \mathcal{L}_\Box$. Then, $\alpha^1(\lceil\varphi\rceil) \to \varphi \in (\Box p \to p)^e$. Since $\mathcal{M}, w \models \Box p \to p$, by Lemma 23, $\mathcal{N} \models \alpha^1(\lceil\varphi\rceil) \to \varphi$. The remaining cases follow analogously. ∎

6 Conclusions

Montague [12] and Thomason [16] signal a problem for the notion that the objects of implicit knowledge and belief are sentences. They show that paradoxes result when familiar axiom schemes characterizing these propositional attitudes are instantiated with certain self-referential sentences.

De Rivières and Levesque [6] characterize a class of sentences – the *regular sentences* – over which these axiom schemes can safely range. But restricting the schemes to regular sentences is a heavy-handed solution. For one thing, regular sentences do not quantify over propositional objects, so there are no regular representations of, for example, 'The system says only what it takes to be true' or 'There are among the user's beliefs some that are false'.

We have stated the problems discovered by Montague and Thomason. We have presented the solution of de Rivières and Levesque. We have argued that their solution is incomplete. Finally, we have extended it by characterizing a more inclusive class of sentences over which the axiom schemes characterizing knowledge and belief can safely range. This more inclusive class of sentences includes representations of our example sentences.

We expect that this result can be further extended, allowing the schemes to range over more sentences, though in the light of Theorems 1 and 2 there is a limit to what can be done. Maybe in a partial logical setting, within which self-referential sentences can be allowed to go without truth values, the axiom schemes can be allowed to range over the entire language. But this is not possible in the classical setting that has seemed attractive to researchers in artificial intelligence.

References

[1] W. Bibel. *Automated Theorem Proving.* Friedr. Vieweg and Sohn Verlagsgesellschaft GmbH, Braunschweig, 1982.

[2] W.W. Bledsoe and D.W. Loveland. *Automated theorem proving.* American Mathematical Society, Providence, 1984.

[3] A. Bundy. *The computer modelling of mathematical reasoning.* Academic Press, London, 1983.

[4] B. F. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, Cambridge, UK, 1980.

[5] E. Davis. *Representation of Commonsense Knowledge.* Morgan Kaufman Publishers, Inc, California, 1990.

[6] J. de Rivières and H. Levesque. The consistency of syntactical treatments of knowledge (how to compile quantificational modal logics into classical FOL. *Computational Intelligence*, 4:31–41, 1988.

[7] D. Gallin. *Intensional and higher-order modal logic : with applications to Montague semantics.* North Holland Pub. Co., Amsterdam, 1975.

[8] P. H. Grice. *Studies in the way of words.* Harvard Univ. Pr., Cambridge, 1989.

[9] A. Haass. The syntactic theory of belief and knowledge. *Artificial Intelligence*, 28(3):245–293, 1983.

[10] D. Lewis. *Counterfactuals.* Blackwell, Oxford, 1973.

[11] E. Mendelson. *Introduction to mathematical logic.* Wadsworth & Brooks/Cole, 1987. Third Edition.

[12] R Montague. Syntactical treatments of modality, with corollaries on reflection prinnciples and finite axiomatizability. In *Modal and Many-Valued Logics (Acta Philosophica Fennica, vol. 16).* Academic Bookstore, Helsinki, 1963. Reprinted in R. Montague (1974). *Formal Philosophy*, New Haven, pp. 286-302.

[13] L. Morgenstern. *Foundations of a Logic of Knowledge, Action, and Communication.* PhD thesis, New York University, 1988.

[14] D. Perlis. Language with self references I: Foundation. *Artificial Intelligence*, 25:301–322, 1985.

[15] A. Tarski. The concept of truth in formalized languages. In *Logic, Semantics, and Metamathematics: papers from 1923 to 1938.* Clarendon press, Oxford, 1956. translated by J.H. Woodger.

[16] R Thomason. A note on syntactical treatments of modality. *Synthese*, 44:391–395, 1980.