The Consistency of Syntactical Treatments of Knowledge

Jim des Rivières
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4

Hector J. Levesque
Department of Computer Science
University of Toronto and
The Canadian Institute for Advanced Research

Abstract

The relative expressive power of a sentential operator $\Box\alpha$ is compared to that of a syntactical predicate $L(\text{`}\alpha\text{'})$ in the setting of first-order logics. Despite results by Montague and by Thomason that claim otherwise, any of the so-called "modal" logics of knowledge and belief can be translated into classical first-order logics that have a corresponding predicate on sentences.

¡Modality, SI; Modal Logic, NO!

— *John McCarthy*

In most logics of knowledge and belief there is a symbol in the language that is intended to be interpreted as "knows" (or "believes"). In first-order logics there is some flexibility regarding the syntactic type of this symbol, namely the choice between a sentential operator symbol, say **K**, which would be used like the negation operator $\neg$ to prefix a *formula* (e.g., $\mathbf{K}(P \supset Q)$), and a predicate symbol Know, which would be used in the standard first-order manner to prefix a *term* that is the name of a formula (e.g., Know('$P \supset Q$')). We would like to report on our findings concerning the relative expressive power of sentential operators *versus* these so-called syntactical predicates and the apparent difficulties in consistently formalizing the predicate approach.

A glimpse at the literature on logics of knowledge and belief reveals a diversity of approaches, with the expected split between operators and predicates: Hintikka [2], Levesque [6,7], and Konolige [4] are among those who have put forward first-order logics extended with sentential operators for knowledge and belief; McCarthy [8], Creary [1], Moore [11], Konolige [3], and Perlis [14], are just some who have proposed first-order logics of knowledge and belief that employ one or more predicates on sentences (sometimes in the guise of *concepts* as pointed out by McCarthy in [8]). Although this is not the place to delve into the reasons for the split, we note that the issues touched on in this paper may have been contributing factors.[1]

When discussing the translation (or conversion, or reduction) of a sentential operator to a syntactical predicate, it is not immediately clear what properties must be preserved. In the context of the modality of necessity, where the symbol $\Box$ is the usual necessity operator, Quine has stated that

> there would be comfort in being able to regard '$\Box$' as mere shorthand for 'Nec' and a pair of quotation marks—thus '$\Box$(9 is odd)' for 'Nec '9 is odd' '. (Quine, [16, p. 268])

This is completely analogous to the way a sentence involving the connective $\equiv$ can be understood, *macro*-style, as shorthand for one expressed entirely in terms of the connective symbols $\supset$ and $\wedge$. Implicit in Quine's desideratum is the property that iterated necessity operators should be read in a like fashion—thus $\Box\Box$(9 is odd) would be short not for Nec('$\Box$(9 is odd)'), but rather Nec('Nec('9 is odd')').

Let us tentatively adopt just such a reading of $\Box$ and see how the modal laws should be reinterpreted. For simplicity, we will refer to the language with the $\Box$ symbol as the "modal" language, and the other one as the "classical" language. Further, we will use L as the name for the unary predicate symbol that appears in the classical language but not in the modal one, chiefly because what we have to say has nothing to do with necessity or any other specific modality.[2]

One of the axiom schemata of the standard epistemic and alethic modal logics is

[1] Moore certainly thought so: "The main reason that modal logics are generally favored over syntactic methods, however, is that there are severe difficulties in formalizing the syntactic approach." [11, p. 216]

[2] In fact, the main application in AI of our results will be for knowledge and belief.

$$\Box\alpha \supset \alpha,$$

where α is a schematic letter ranging over all sentences in the modal language. For example, if P and Q are nullary predicate letters in the modal language, the sentence

$$\Box(P \wedge \Box Q) \supset (P \wedge \Box Q)$$

is an instance of that axiom schema. When, following Quine, we reinterpret this as shorthand, we see it as an abbreviation for the sentence

$$L(\text{`}P \wedge L(\text{`}Q\text{'})\text{'}) \supset (P \wedge L(\text{`}Q\text{'})).$$

Using Quine's quasi-quotation to more correctly reflect the fact that α is being used as a schematic sentence letter, we can re-express the above schema as

$$L(\ulcorner\alpha\urcorner) \supset \alpha.$$

But the key question here is: Over which sentences should α range? The modal language? The entire classical language? Or perhaps just a subset thereof?

A sentence schema is best viewed as an abbreviation for a set of sentences in a language; convenience aside, it is the set that matters, nothing else. When trying to translate a schema in one language into one in another language, the schema should first be expanded into the set of sentences that it describes, then each of these sentences should be translated individually, and finally one can try to find a convenient schematic description of that resulting set. The other way of approaching it—simply translate the non-schematic portion of the schema and consider all schematic letters as ranging over sentences in the target language—need not always lead to the correct result.

If we carefully follow the steps mentioned above, the schema in the modal language expands to a particular (infinite) set of sentences in the modal language, and then each of these sentences is seen to be shorthand for some sentence in the classical language. The corresponding classical schema must cover all and only the sentences in this set. Note, however, there are some sentences in the classical language that have no shorthand equivalent in the modal language. For example, $\exists x\, L(x)$ has no short form, because of the occurrence of a variable in the argument position of an L predication.[3] In other words, our re-reading of sentences in the modal language only yields a subset of the classical language. Call a sentence *regular* if it belongs to this subset. Since the regular sentences are a proper subset of the classical language, and since every sentence described by a modal schema will correspond to a regular one, the correct re-interpretation of any modal schema will necessarily be a schema whose schematic letters range over just the regular sentences in the classical language. The correct re-reading of the schema:

$\Box\alpha \supset \alpha$ for all sentences α in the modal language,

[3]Recall that modal sentences may not use L (it's not a symbol in the modal language) and, as will be seen later, even formulas like $\exists x\, \Box\alpha$ will never give rise to formulas that have x in the initial argument position of L.

is, therefore, the schema:

$L(\ulcorner\alpha\urcorner) \supset \alpha$ for all *regular* sentences α in the classical language.

With this as our basis, we will show that there can be syntactical treatments of modal operators of the form Quine has wished for all along.

The remainder of this paper is organized as follows. After laying down our background assumptions in the next section, we review Montague's and Thomason's results. The matter of making precise our notational games is taken up in the section following that, along with proof that the simple way of deriving syntactical treatments of arbitrary sentential operators is consistent. We are then in a position to apply this technique to the standard modal logics to obtain classical first-order logics with all of the properties one might reasonably expect. Montague's and Thomason's negative results will be seen to apply only if the modal axiom schemata are inappropriately translated into ones in which the schematic sentence letters range over the *entire* classical language. We conclude with some suggested directions for future research.

Terminology and Notation

We assume throughout that the common base language will be a first-order logical language $\mathcal{L}$ with logical symbols $\neg$, $\supset$, and $\forall$, and replete with predicate, function and constant symbols.[4] The Greek letters α, β, and γ will be used for open formulas; ϕ, ψ, and σ will stand for sentences (closed formulas); S and T will stand for sets of sentences.

$\mathcal{L}(\Box)$ will be the "modal" language obtained by augmenting $\mathcal{L}$ with the unary sentential operator $\Box$, and $\mathcal{L}(L)$ will be the classical language obtained from $\mathcal{L}$ by adding the family of $(n+1)$-ary predicate symbols L_n, for each natural number n.[5] When we aren't concerned with quantifying into modal contexts, we'll often use the symbol L as a synonym for L_0. (In a later section we will suggest how to get by with a single 2-place predicate symbol.)

For the sake of brevity, we have focussed our attention on languages with one extra unary sentential operator. There is no difficulty in adapting our techniques to sets of operators of arbitrary arities.

Because we will be considering syntactical treatments of sentential operators, we need to be precise about what this involves. As we understand Quine and Montague, such a treatment necessarily involves a predicate of sentences as notational forms. These *syntactical* predicates would be prefixed to a term that serves to name some formula (or term) in some (perhaps different) language. The only syntactical predicates that we will encounter are the L_i, which are found only in the language $\mathcal{L}(L)$. We will require that our languages have a collection of closed terms that serve as names for each of the formulas and terms of $\mathcal{L}(L)$. These terms are called the *encoding terms* for $\mathcal{L}(L)$; the encoding term corresponding to the formula α will be written $\ulcorner\alpha\urcorner$. Distinct formulas will have distinct encoding terms. Although

[4]Other logical symbols like $\vee$ and $\exists$ can be introduced as abbreviations in the customary manner.

[5]Quine [16] calls L a *multigrade* predicate symbol.

we place no other constraints on the encoding scheme, we do wish to point out that the kind we have in mind would *not* employ an infinite number of distinct function and constant *symbols*. This is because we wish our systems to be able to express the standard notions of *elementary syntax* [10], including the functions for forming, dissecting, and categorizing the formulas of the language being encoded.[6] The encoding scheme plays only a background role, remaining invariant throughout the presentation.

Translation functions that map formulas of one language into formulas of another play a key role in this investigation. Embeddings are one kind of translation function that are of fundamental importance.

Definition Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be languages, both of which use at least the connectives of first-order logic. A translation function $^\diamond: \mathcal{L}_1 \to \mathcal{L}_2$ is an *embedding of* $\mathcal{L}_1$ *in* $\mathcal{L}_2$ iff

(i) $^\diamond$ maps atomic formulas to themselves; i.e., $\alpha^\diamond = \alpha$.

(ii) $^\diamond$ distributes over the connectives of first-order logic; i.e., $(\neg\alpha)^\diamond = \neg\alpha^\diamond$, $(\alpha \supset \beta)^\diamond = \alpha^\diamond \supset \beta^\diamond$, and $(\forall x\, \alpha)^\diamond = \forall x\, \alpha^\diamond$.

(iii) $\alpha^\diamond$ has the same free variables as α.

In other words, an embedding is a homomorphism from $\mathcal{L}_1$ to $\mathcal{L}_2$ except that it is almost entirely unrestricted in its treatment of non-standard symbols of $\mathcal{L}_1$, such as the operator $\Box$.

The generalization of theoremhood to include modal languages is defined in terms of embeddings. For the remainder of this section, unless otherwise indicated, the source language $\mathcal{L}_1$ can be either $\mathcal{L}(\Box)$ or $\mathcal{L}(L)$, with $\phi \in \mathcal{L}_1$ and $S \subseteq \mathcal{L}_1$; the target language $\mathcal{L}_2$ will always be $\mathcal{L}(L)$.

Definition ϕ is an *extended theorem* iff for every embedding $^\diamond$ of $\mathcal{L}_1$ in $\mathcal{L}(L)$, $\phi^\diamond$ is a theorem of first-order logic.

Clearly, if $\mathcal{L}_1$ does not contain any non-standard symbols, the only embedding possible is the identity function. The extended theorems of $\mathcal{L}(L)$ are, therefore, precisely the theorems of first-order logic.

On the other hand, ϕ is an extended theorem of the modal language $\mathcal{L}(\Box)$ iff ϕ with every subformula $\Box\beta$ replaced by any formula γ of $\mathcal{L}(L)$ with the same free variables is a theorem of first-order logic. Note, however, that not all substitution instances of theorems of first-order logic will be extended theorems. For example, while $(\forall x\, \alpha \supset \alpha(x/t))$ is a theorem of first-order logic for any α, a substitution instance of it,

$$\forall x\, \Box P(x) \supset \Box P(t),$$

[6] The reader will not be lead astray if he thinks of $\mathcal{L}$ as being the language of elementary number theory, $\mathcal{A}$, with non-logical symbols $=$, s (successor), $+$, $\cdot$, and 0. For a particular theory, take Robinson's finitely-axiomatized arithmetic theory **Q** [18], and understand the encoding terms to be a subset of the numerals 0, $s0$, $ss0$, etc., set in correspondence with the terms and formulas of $\mathcal{A}(L)$ via some sort of Gödel numbering scheme.

is not a theorem of quantified modal **S4**. Fortunately, it is not an extended theorem according to the above definition.

The notions of derivability and consistency can also be defined in a way that is compatible with first-order logic while, at the same time, being applicable to modal languages as well.

Definition ϕ is *derivable* from S (written $S \vdash \phi$) iff ϕ follows from S and the extended theorems of $\mathcal{L}_1$ by *modus ponens* alone.

Since the set of extended theorems is itself closed under *modus ponens*, $\vdash \phi$ iff ϕ is an extended theorem. Moreover, since *modus ponens* is the only rule of inference, the Deduction Theorem holds.

Definition S is *inconsistent* iff $S \vdash \phi$ for every ϕ in $\mathcal{L}_1$.

When $\mathcal{L}_1 = \mathcal{L}(L)$, these definitions coincide with the usual ones. For $\mathcal{L}(\Box)$, the above definitions automatically restrict our attention to modal systems that are "upward compatible" with first-order predicate calculus. While this immediately rules out some systems, such as intuitionistic and relevance logics, it sanctions the usual modal theories. For example, the set of theorems of quantified modal logic **S5** is consistent according to the above definition. In addition, we are allowing for *very* different sorts of modal theories. Although air-tight constraints are imposed on the interpretation of the connective symbols occurring outside the scope of a $\Box$ operator, the interpretation of the $\Box$ operator itself and, therefore, everything within its scope is almost totally unconstrained. The $\Box$ operator can be made as referentially opaque as desired. For example,

$$\{\forall x \Box P(x), \neg \forall y \Box P(y)\}$$

is consistent. In this case the '$\Box$' could be read as "uses the variable 'x' freely".

Review of Montague and Thomason

The first result, due to Montague [9], sets its sights on syntactical treatments of the standard modal logics that include the axiom of necessity, such as **T**, **S4**, and **S5**. In this theorem, and the one following it, the open formula α with free variable x should be viewed as a generalization of the atomic formula $L(x)$.

Theorem 1 (Montague, Theorem 1) Let $\mathcal{A}^+$ be a first-order language that contains at least the non-logical symbols of elementary number theory. Let T be a set of sentences of $\mathcal{A}^+$ and α a formula of $\mathcal{A}^+$ with one free variable. Suppose that the following conditions are met, for all sentences ϕ and ψ of $\mathcal{A}^+$.

(i) $T \vdash Q$, where Q is the single axiom for Robinson's arithmetic system **Q**,

(ii) $T \vdash \alpha(\ulcorner \phi \urcorner) \supset \phi$,

(iii) $T \vdash \alpha(\ulcorner \alpha(\ulcorner \phi \urcorner) \supset \phi \urcorner)$,

(iv) $T \vdash \alpha(\ulcorner \phi \supset \psi \urcorner) \supset (\alpha(\ulcorner \phi \urcorner) \supset \alpha(\ulcorner \psi \urcorner))$,

(v) $T \vdash \alpha(\ulcorner\phi\urcorner)$, if ϕ is a logical axiom.

Then T is inconsistent.

Proof: The proof involves a "self-referential" sentence of the form $\sigma \equiv \alpha(\ulcorner Q \supset \neg\sigma\urcorner)$, which is shown to be paradoxical. Refer to the original paper for details. ∎

From this result Montague reasons as follows:

> Now what general conclusions can be drawn from Theorems 1–4? In the first place, observe that the schemata in conditions (ii)–(v) of Theorem 1 are provable in the well-known systems of first-order modal logic with identity.... These schemata would, moreover, be provable in any reasonable extension of predicate logic of S1, the weakest of the Lewis modal calculi. Further, it is not unnatural to impose condition (i): modal logic, like ordinary logic, ought to be applicable to an arbitrary subject matter, including arithmetic....
>
> *Thus if necessity is to be treated syntactically, that is, as a predicate of sentences, as Carnap and Quine have urged, then virtually all of modal logic, even the weak system S1, must be sacrificed.* (Montague, [9, p. 294]; Italics added)

Theorem 1 obviously applies to standard approaches to idealized knowledge (e.g., Hintikka [2]), since they share the same basic axiom schemata with the modal logics. But logics of idealized belief, for which the axiom schema corresponding to condition (ii) is inappropriate, are not covered. With the close ties between knowledge and truth, and given Tarski's famous theorem [17] on the non-definability of a truth predicate,[7] one might always hope that belief predicates would not be susceptible to paradoxes of self-reference. However, as the next theorem due to Thomason [19] shows, logics of idealized belief—notably weak **S5**—are also in jeopardy.

Theorem 2 (Thomason, Theorem 2)[8] Let $\mathcal{A}^+$ be a first-order language that contains at least the non-logical symbols of elementary number theory. Let T be a set of sentences of $\mathcal{A}^+$ and α a formula of $\mathcal{A}^+$ with one free variable. Suppose that the following conditions are met, for all sentences ϕ and ψ of $\mathcal{A}^+$.

(i) $T \vdash \alpha(\ulcorner\phi\urcorner) \supset \alpha(\ulcorner\alpha(\ulcorner\phi\urcorner)\urcorner)$,

(ii) $T \vdash \alpha(\ulcorner\alpha(\ulcorner\phi\urcorner) \supset \phi\urcorner)$,

(iii) $T \vdash \alpha(\ulcorner\phi\urcorner)$, if ϕ is a logical axiom,

(iv) $T \vdash \alpha(\ulcorner\phi \supset \psi\urcorner) \supset (\alpha(\ulcorner\phi\urcorner) \supset \alpha(\ulcorner\psi\urcorner))$.

Then for all sentences $\psi \in \mathcal{A}^+$,

(v) $T \vdash \alpha(\ulcorner Q\urcorner) \supset \alpha(\ulcorner\psi\urcorner)$

where Q is the single axiom for **Q**.

[7] In fact, Montague considered his result to be a refinement of Tarski's.

[8] With some minor changes: no relativization, sentences in place of arbitrary formulas, and a bug fix.

Proof: Similar to the proof of Montague's theorem, this proof involves a paradoxical sentence of the form $\sigma \equiv (Q \supset \alpha(\ulcorner\neg\sigma\urcorner))$. Refer to the original paper for details. ∎

Thomason concludes:

> Though this theorem does not show T to be inconsistent, it does establish that T would become inconsistent upon the addition of $[\alpha(\ulcorner Q\urcorner)]$ and $\neg\alpha(\ulcorner\psi\urcorner)$, for any formula ψ. This seems to show a coherent theory of idealized belief as a syntactical predicate to be problematic. (Thomason, [19, p. 393])

Thus the syntactical predicate approach to knowledge and belief appears to be seriously flawed — and in a way that the operator approach is not. In what follows, we will show that these appearances are indeed deceiving and are based on a misconception of what should be required of a syntactical treatment of modality.

A Translation-Based Syntactical Treatment

In this section we will define a particular embedding * from formulas of $\mathcal{L}(\Box)$ to formulas of $\mathcal{L}(L)$ and show that this mapping preserves the important property of derivability. That is, for all $T \subseteq \mathcal{L}(\Box)$ and $\sigma \in \mathcal{L}(\Box)$, $T \vdash \sigma$ iff $T^* \vdash \sigma^*$. We will see that * maps $\mathcal{L}(\Box)$ to a subset of $\mathcal{L}(L)$. Consequently, sets of sentences arising from the proper translation of sets of modal sentences will only dictate how the predicate L should treat a subset of $\mathcal{L}(L)$, while remaining completely neutral on the rest.

The Translation Function *

We introduce a translation function from the modal language to the classical one in order to make precise the idea of reading $\Box$ as shorthand for L and a pair of quotation marks.

Definition The translation function * is defined to be the embedding of $\mathcal{L}(\Box)$ in $\mathcal{L}(L)$ with the property

$$(\Box\alpha)^* = L_n(\ulcorner\alpha^*\urcorner, x_1, x_2, \ldots, x_n)$$

where $x_1, x_2, \ldots x_n$ are the free variables of the formula α, listed in some predetermined, fixed order.

* is clearly 1–1, since $L_n \notin \mathcal{L}(\Box)$ and distinct formulas have distinct encoding terms.

Definition $\alpha \in \mathcal{L}(L)$ is *regular* iff $\alpha = \gamma^*$ for some $\gamma \in \mathcal{L}(\Box)$. In other words, the set of regular formulas is just $\mathcal{L}(\Box)^*$ (see Figure 1).

Reductions

For a mapping such as * to be considered to be a reduction of the modal *operator* to a syntactical predicate, it must, at the very least, preserve derivability.

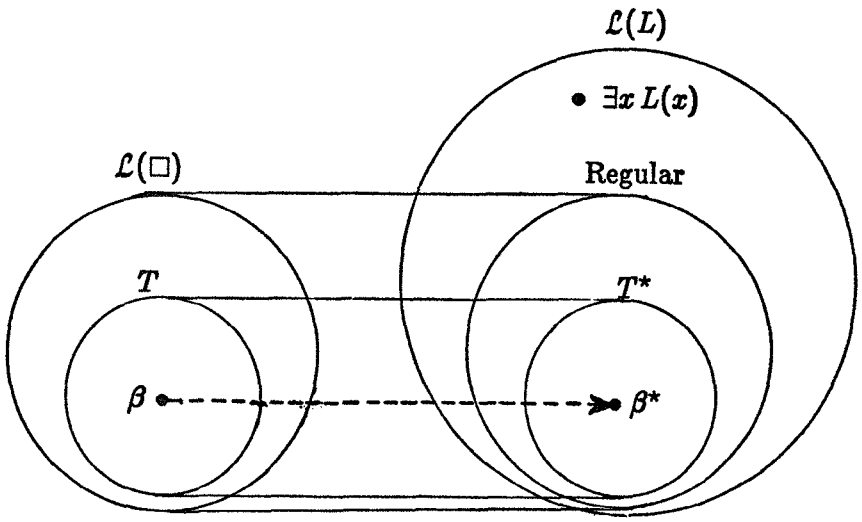

Figure 1: The Regular Formulas of $\mathcal{L}(L)$.

Definition A translation function $^\diamond$: $\mathcal{L}_1 \to \mathcal{L}_2$ is said to be a *reduction* of $S \subseteq \mathcal{L}_1$ to $T \subseteq \mathcal{L}_2$ if, for all $\phi \in \mathcal{L}_1$,

$$S \vdash \phi \iff T \vdash \phi^\diamond.$$

Morgan's syntactic approach [12] provides a good example of how to reduce arbitrary logics to classical first-order ones. Note, however, that a reduction need not preserve any aspect of a sentence's structure. On the other hand, we are interested in the *embedding* *, which observes all of the normal connectives, mapping $\neg$ to $\neg$, $\supset$ to $\supset$, etc., and also preserves all atomic formulas. This is because, following Quine, we wish to re-interpret the $\Box$ operator somehow as a predicate, but leave the rest of the modal sentence unchanged.

More generally, we are also interested in translation functions that can be used to reduce not just a single set of sentences but all such sets.

Definition A translation function $^\diamond$: $\mathcal{L}_1 \to \mathcal{L}_2$ is said to be a *general reduction* of $\mathcal{L}_1$ to $\mathcal{L}_2$ if for every $T \subseteq \mathcal{L}_1$, $^\diamond$ is a reduction of T to $T^\diamond = \{\phi^\diamond \mid \phi \in T\}$.

Lemma 3 General reductions preserve consistency. That is, if $^\diamond$ is a general reduction of $\mathcal{L}_1$ to $\mathcal{L}_2$, then for all $T \subseteq \mathcal{L}_1$, T is consistent iff $T^\diamond$ is consistent.

Proof: $T \subseteq \mathcal{L}_1$ inconsistent $\iff$ $U \vdash \phi$ for all $\phi \in \mathcal{L}_1$ $\iff$ $T^\diamond \vdash \phi^\diamond$ $\iff$ $T^\diamond \vdash (\sigma \wedge \neg\sigma)^\diamond$ for all σ $\iff$ $T^\diamond \vdash \sigma^\diamond \wedge \neg\sigma^\diamond$ $\iff$ $T^\diamond$ inconsistent. ∎

Properties of *

Lemma 4 For any $\phi \in \mathcal{L}(\Box)$, if $\phi^\diamond$ is satisfiable for some embedding $^\diamond$ of $\mathcal{L}(\Box)$ in $\mathcal{L}(L)$, then ϕ^* is also satisfiable.

Proof: If $\phi^\diamond$ is satisfiable, then it is satisfiable in a term model $M = (\mathcal{T}, I)$ whose domain $\mathcal{T}$ includes all of the closed terms of $\mathcal{L}(L)$. Define $M^* = (\mathcal{T}, I^*)$ be the first-order model structure for $\mathcal{L}(L)$ with domain of interpretation $\mathcal{T}$ and agreeing with M on the interpretation of all non-logical symbols except $\{L_0, L_1, \ldots\}$; i.e, let $I^*(P) = I(P)$ for each predicate symbol P of $\mathcal{L}$. For each $n \geq 0$, define

$$I^*(L_n) = \{(\ulcorner \alpha^* \urcorner, t_1, \ldots, t_n) \in \mathcal{T}^{n+1} \mid \alpha \in \mathcal{L}(\Box) \text{ with free variables } x_1, \ldots, x_n \text{ and } M, \nu^{x_1,\ldots,x_n}_{t_1,\ldots,t_n} \models (\Box\alpha)^\diamond\}.$$

We show by induction on the structure of any formula $\alpha \in \mathcal{L}(\Box)$ that, for all variable assignments ν,

$$M, \nu \models \alpha^\diamond \iff M^*, \nu \models \alpha^*.$$

There are five cases:

(1) α is atomic. It follows immediately from the fact that $\alpha^\diamond = \alpha^*$ and that M and M^* agree on the interpretation of all predicate symbols and terms of $\mathcal{L}$ (i.e., exclusive of the L_i predicates).

(2) $\alpha = \Box\beta$. Let $x_1, \ldots, x_n$ be the free variables of β.

$$\begin{aligned} M, \nu \models (\Box\beta)^\diamond &\iff (\ulcorner \beta^* \urcorner, \nu(x_1), \ldots, \nu(x_n)) \in I^*(L_n) \quad (\text{since } \gamma^* = \beta^* \text{ iff } \gamma = \beta) \\ &\iff M^*, \nu \models L_n(\ulcorner \beta^* \urcorner, x_1, \ldots, x_n) \\ &\iff M^*, \nu \models (\Box\beta)^*. \end{aligned}$$

(3) $\alpha = \neg\beta$.

$$\begin{aligned} M, \nu \models (\neg\beta)^\diamond &\iff M, \nu \models \neg\beta^\diamond \\ &\iff M, \nu \not\models \beta^\diamond \\ &\iff M^*, \nu \not\models \beta^* \qquad (\text{by the induction hypothesis}) \\ &\iff M^*, \nu \models \neg\beta^* \ (= (\neg\beta)^*). \end{aligned}$$

(4) $\alpha = \beta \supset \gamma$. Similar to the preceding case.

(5) $\alpha = \forall x\, \beta$.

$$\begin{aligned} M, \nu \models (\forall x\, \beta)^\diamond &\iff M, \nu \models \forall x\, \beta^\diamond \\ &\iff M, \nu^x_t \models \beta^\diamond \text{ for all } t \in \mathcal{T} \\ &\iff M^*, \nu^x_t \models \beta^* \text{ for all } t \in \mathcal{T} \qquad (\text{by the induction hypothesis}) \\ &\iff M^*, \nu \models \forall x\, \beta^* \ (= (\forall x\, \beta)^*). \end{aligned}$$

Thus, for all $\alpha \in \mathcal{L}(\Box)$ and all assignments ν,

$$M, \nu \models \alpha^\diamond \iff M^*, \nu \models \alpha^*.$$

∎

Lemma 5 For all $\phi \in \mathcal{L}(\Box)$, $\vdash \phi \iff \vdash \phi^*$.

Proof: ($\Rightarrow$) If $\vdash \phi$ then ϕ is an extended theorem of $\mathcal{L}(\Box)$. Therefore, by the definition of extended theorem, $\vdash \phi^\diamond$ for every embedding $^\diamond$ of $\mathcal{L}(\Box)$ in $\mathcal{L}(L)$, of which * is included.
($\Leftarrow$) If $\nvdash \phi$ then ϕ is not an extended theorem. So, $\nvdash \phi^\diamond$ for some embedding $^\diamond$. By first-order completeness, $\neg\phi^\diamond$ is satisfiable. By Lemma 4, $\neg\phi^*$ must also be satisfiable. Therefore $\nvdash \phi^*$ by first-order soundness. ∎

Theorem 6 * is a general embedding reduction of $\mathcal{L}(\Box)$ to $\mathcal{L}(L)$.

Proof: Take any $T \subseteq \mathcal{L}(\Box)$ and any $\phi \in \mathcal{L}(\Box)$. Then

$$\begin{aligned} T \vdash \phi &\iff \vdash \sigma_1 \supset \cdots \supset \sigma_k \supset \phi \text{ for some } \sigma_1, \ldots, \sigma_k \in T \\ &\iff \vdash (\sigma_1 \supset \sigma_2 \supset \cdots \supset \sigma_k \supset \phi)^* && \text{(by Lemma 5)} \\ &\iff \vdash \sigma_1^* \supset \sigma_2^* \supset \cdots \supset \sigma_k^* \supset \phi^* \\ &\iff T^* \vdash \phi^*. \end{aligned}$$

Hence * is a general reduction of $\mathcal{L}(\Box)$ to $\mathcal{L}(L)$. ∎

Montague and Thomason Revisited

By Theorem 6, given any consistent set of sentences over the modal language $\mathcal{L}(\Box)$ one can always reduce it using * to a consistent set of sentences of $\mathcal{L}(L)$ in which the role of the extra sentential operator is played instead by a syntactical predicate. This treatment is entirely consonant with the systematic reading of the operator $\Box$ as shorthand for the multigrade predicate L and a pair of quotation marks that Quine has countenanced. Applied to a genuine modal theory such as **S5**, one obtains a classical first-order system that rightfully deserves to be considered as *a syntactical treatment of modality.*

But how, then, are we to reconcile this with Montague's result, which seemed to show that syntactical treatments of modality are not possible? The answer is quite simple. Notice that in conditions (ii)–(v) of Theorem 1, which are supposed to be the predicate counterparts of the modal theorem schemata $(\Box\phi \supset \phi)$, $\Box(\Box\phi \supset \phi)$, etc., respectively, are schemata with ϕ and ψ ranging over the *entire* classical language. These conditions are overly stringent. As mentioned earlier, the correct predicate counterparts of the modal schemata are the classical schemata (ii)–(v) with ϕ and ψ ranging over *just the regular subset* of the classical language. As the following result illustrates, by restricting ϕ and ψ in conditions (ii)–(v) to be regular, the claim of inconsistency no longer follows. This means that the irregular sentences are indeed the source of any inconsistencies (a stronger claim than the simple observation that Montague's *proof* employs irregular sentences).

Non-Theorem 7 (cf. Theorem 1) Let $\mathcal{A}^+$ be a first-order language that contains at least the non-logical symbols of elementary number theory. Let T be a set of sentences of $\mathcal{A}^+$ and α a formula of $\mathcal{A}^+$ with one free variable. Suppose that the following conditions are met, for all *regular* sentences ϕ and ψ of $\mathcal{A}^+$.

(i) $T \vdash Q$, where Q is the single axiom for Robinson's arithmetic system **Q**,

(ii) $T \vdash \alpha(\ulcorner\phi\urcorner) \supset \phi$,

(iii) $T \vdash \alpha(\ulcorner\alpha(\ulcorner\phi\urcorner) \supset \phi\urcorner)$,

(iv) $T \vdash \alpha(\ulcorner\phi \supset \psi\urcorner) \supset (\alpha(\ulcorner\phi\urcorner) \supset \alpha(\ulcorner\psi\urcorner))$,

(v) $T \vdash \alpha(\ulcorner\phi\urcorner)$, if ϕ is a logical axiom.

Then T is inconsistent.

Counterexample: Let $S \subseteq \mathcal{A}(\Box)$ consist of the theorems of the modal system **S5** together with the single axiom Q of Robinson's arithmetic. S is consistent since it is satisfiable in a suitable Kripke model with arithmetic. For $\mathcal{A}^+$ use $\mathcal{A}(L)$, for α use the open formula $L_0(x)$, and let $T = S^\star$. Condition (i) holds since $Q = Q^\star$ and $Q \in S$. $(\Box\sigma \supset \sigma) \in S$ for every $\sigma \in \mathcal{A}(\Box)$, because this is a theorem schema for **S5**. Hence $S^\star \vdash L_0(\ulcorner\sigma^{\star}\urcorner) \supset \sigma^\star$ for every $\sigma \in \mathcal{A}(\Box)$; i.e., $S^\star \vdash \alpha(\ulcorner\phi\urcorner) \supset \phi$ for every *regular* $\phi \in \mathcal{A}(L)$. Similarly, conditions (iii)–(v) follows in virtue of the corresponding properties of **S5**. By Theorem 6, * is a general reduction, which are always consistency preserving by Lemma 3. So $S^\star$ is consistent, providing a counterexample to this slightly modified statement of Montague's theorem. ∎

Similar considerations apply equally well to Thomason's result.

Non-Theorem 8 (cf. Theorem 2) Let $\mathcal{A}^+$ be a first-order language that contains at least the non-logical symbols of elementary number theory. Let T be a set of sentences of $\mathcal{A}^+$ and α a formula of $\mathcal{A}^+$ with one free variable. Suppose that the following conditions are met, for all *regular* sentences ϕ and ψ of $\mathcal{A}^+$.

(i) $T \vdash \alpha(\ulcorner\phi\urcorner) \supset \alpha(\ulcorner\alpha(\ulcorner\phi\urcorner)\urcorner)$,

(ii) $T \vdash \alpha(\ulcorner\alpha(\ulcorner\phi\urcorner) \supset \phi\urcorner)$,

(iii) $T \vdash \alpha(\ulcorner\phi\urcorner)$, if ϕ is a logical axiom,

(iv) $T \vdash \alpha(\ulcorner\phi \supset \psi\urcorner) \supset (\alpha(\ulcorner\phi\urcorner) \supset \alpha(\ulcorner\psi\urcorner))$.

Then for all sentences $\psi \in \mathcal{A}^+$,

(v) $T \vdash \alpha(\ulcorner Q\urcorner) \supset \alpha(\ulcorner\psi\urcorner)$

where Q is the single axiom for **Q**.

Counterexample: Use the same counterexample as in the preceding non-theorem. ∎

Why it Works

Non-Theorems 7 and 8 indicate quite clearly that the restriction to regular sentences makes a significant difference. What is happening is this. Both Montague's and Thomason's proofs rely on the existence of fixed points for certain formulas in the language. In particular, Montague's proof requires that there be a sentence ϕ satisfying $T \vdash \phi \equiv \alpha(\ulcorner Q \supset \neg\phi\urcorner)$. The existence of such a *fixed-point* sentence ϕ follows from the Fixed-Point Theorem (a proof appears in [10]). The particular fixed point ϕ whose construction is given in his proof has the form

$$\forall x\,(x = \ulcorner \forall y\,(\delta(x,y) \supset \alpha(y))\urcorner \supset \forall y\,(\delta(x,y) \supset \alpha(y)))$$

where δ is a (regular) formula with two free variables. The important thing to note is that this ϕ is *not* a regular sentence because it contains the subformula $L(y)$ $(= \alpha(y))$. Similarly, Thomason's Theorem 2 relies on a fixed point ϕ such that $T \vdash \phi \equiv (Q \supset \alpha(\ulcorner\neg\phi\urcorner))$, and, again, the ϕ constructed would not be regular. Furthermore, *no regular sentence could be used in place of these ϕ sentences.* By restricting our attention to the regular sentences we are failing to provide axioms that apply to those fixed point sentences upon which their results hinge.

Other Considerations

Compatibility with First-Order Theories with Equality

Note that if T is a modal theory with a built-in equality predicate, it need not be the case that T^* be compatible with first-order predicate calculus with equality. In particular, T might confuse the encoding terms; e.g., $T \vdash \ulcorner\sigma^*\urcorner = \ulcorner\neg\sigma^*\urcorner$ for some σ. Then, even though $\ulcorner\sigma^*\urcorner$ and $\ulcorner\neg\sigma^*\urcorner$ are distinct encoding terms and the sentence $(\Box\sigma \wedge \neg\Box\neg\sigma)$ is consistent, T^* would be inconsistent with the full set of theorems of $=$ applied to L because $T^* \vdash L(\ulcorner\sigma^*\urcorner) \wedge \neg L(\ulcorner\neg\sigma^*\urcorner)$. To guarantee that such difficulties do not arise, a restriction must be placed on T to ensure that distinct encoding terms cannot be conflated.

Assume that the notion of extended theorem is redefined in terms of embeddings into first-order predicate calculus *with equality*, and that derivability, consistency, etc. are reinterpreted accordingly.

Definition T is *adequate* iff $T \nvdash s = t$ for all distinct encoding terms s and t.

Robinson's arithmetic system **Q** is adequate for encoding expressions using a subset of the numerals, since $Q \vdash s \neq t$ for all distinct closed terms s and t.

Theorem 9 When restricted to adequate sets of sentences, * is a general embedding reduction of $\mathcal{L}(\Box)$ to $\mathcal{L}(L)$.

Proof: A variation of Lemma 4 must be shown to hold. The restiction that ϕ be adequate is needed to ensure the existence of a model for $\phi^\diamond$ that interprets distinct encoding terms as distinct domain elements (being equivalence classes defined over the terms). Then Lemma 5 and Theorem 6 go through, *mutatis mutandis*. We will not repeat the proofs here; instead, we leave them as an exercise for the obsessed reader. ∎

Comparison to Perlis' Truth Schema

This work has a definite tie-in to the recent work on limited truth predicates. We will discuss this briefly and mention how that has a bearing on the finite axiomatizability of syntactical treatments of knowledge and belief.

Perlis has recently argued [13] that it is overly restrictive to work within a representational framework that precludes any chance of even *expressing* a statement that is potentially paradoxical. After all, we routinely make statements about our own statments, reason about the our own reasoning and our relationship with the external world, and so forth. On occasion we may stumble into a paradox, but we usually react to it more with amusement than anything else.

Perlis has advocated working in classical first-order logic containing a syntactical predicate for truth with the ability to explicitly refer to any formula in the language. Beginning with a theory T over a language $\mathcal{L}$, he shows how to extend it to another classical first-order theory $P(T)$ over the extended language $\mathcal{L}(\mathtt{True})$, True being the suitably-limited truth predicate for the full language $\mathcal{L}(\mathtt{True})$. The axiom schema that is used to extend T to $P(T)$ is

$$\mathtt{True}(\ulcorner\phi\urcorner) \equiv \phi^\circ \quad \text{for all sentences } \phi \in \mathcal{L}(\mathtt{True}), \tag{1}$$

where ϕ° is a variant of ϕ obtained by performing certain, straightforward transformations on some of its subformulas (we need not be concerned with the details here).

Our work relates to his in the following manner. Extend Perlis' base theory T over $\mathcal{L}$ to $R(T)$ over the language $\mathcal{L}(\Box)$ by adjoining all instances of the axiom schema

$$\Box\phi \equiv \phi \quad \text{for all sentences } \phi \in \mathcal{L}(\Box).$$

This axiom schema is obviously consistent; all it says is that $\Box$ is a logical *no-op*. We can now translate $\mathcal{L}(\Box)$ to $\mathcal{L}(\mathtt{True})$ with the embedding that satisfies $(\Box\phi)^\bullet$ to $\mathtt{True}(\ulcorner\phi^\bullet\urcorner)$; i.e., $^\bullet$ is a minor variation on our * that assumes that there is no quantification into modal contexts. It can be shown that $R(T)^\bullet$ is a proper subtheory of $P(T)$. That is, Perlis' limited truth schema (1) includes all of the sentences that a translation-based schema would contain (i.e., all of the regular ones). Indeed, any useful truth schema should handle the regular sentences at the very least. In a sense, $R(T)^\bullet$ is a *minimal* theory, a standard by which the adequacy of other proposals such as Perlis' can be gauged. This is just as one might have expected: whereas our concern was to provide a syntactical treatment of modality without falling prey to the paradoxes of self-reference, Perlis was going after a limited form of self-reference.

Finite Axiomatizability

As Quine has pointed out in [16], a family of predicate symbols like L_i can always be reduced to a single 2-place predicate symbol L_ϵ that takes a finite sequence of variable-value pairings (i.e., an environment/a-list) as its second argument. We would revamp * so that

$$(\Box\alpha)^* = L_\epsilon(\ulcorner\alpha^*\urcorner, \langle\langle\ulcorner x_1\urcorner, x_1\rangle, \langle\ulcorner x_2\urcorner, x_2\rangle, \ldots, \langle\ulcorner x_n\urcorner, x_n\rangle\rangle),$$

where $x_1, x_2, \ldots x_n$ are the free variables of α.

Having a finite language is a necessary start if you are to construct a finite axiomatization, but it is clearly not sufficient. For people interested in translating standard modal theories, one must also attend to problems concerning axiom schemata. Under what circumstances can an axiom schema be replaced by a finite set of axioms? We do not have the full answer

yet. However, it does appear that many of the common axiom schemata can be collapsed into single axioms. For example, the schema $\Box(\phi \supset \phi)$, when translated, could be rendered

$$\forall x \forall e\, \mathtt{regular}(x) \supset L_e(\mathtt{implies}(x,x),e),$$

where `regular` and `implies` are predicates and functions defined over sentence encodings (and easily finitely axiomatized).[9] Unfortunately, a schema like $(\Box\phi \supset \phi)$ cannot be handled exactly as above because a schematic variable appears outside the scope of any $\Box$. It would appear, however, that such schemata can be handled indirectly via a limited truth predicate, like the ones discussed above, in addition to the syntactical predicate being axiomatized. Truth predicates, as it happens, *do* have finite axiomatizations. We hope to be able to report on this at a later date.

Concluding Remarks

The spectre of inconsistency has always loomed over those research programmes which (for good reasons and bad) have attempted to formalize modalities in first-order terms. McCarthy expressed the fear as follows:

> We have not yet investigated the matter, but plausible axioms for necessity or knowledge expressed in terms of concepts may lead to the paradoxes discussed in Kaplan and Montague (1960) and Montague (1963). Our intention is that the paradoxes can be avoided by restricting the axioms concerning knowledge, and necessity of statements about necessity. The restrictions will be somewhat unintuitive as are the restrictions necessary to avoid the paradoxes of naive set theory. (McCarthy, [8, p. 146])

This research has attempted to allay these fears. It shows that any intensional operator governed by a reasonable modal theory, that is, a theory containing all the extended theorems and closed under *modus ponens*, can be treated syntactically in a simple and intuitive way. This certainly applies to (more or less) conventional logics of knowledge such as [7], but also to non-standard ones such as [4] and [5]. There is no danger of introducing inconsistency as long as the treatment does not explicitly insist on more than its modal logic counterpart. It is only those that go beyond this, for example to deal with self-reference directly, that are at risk. This, we feel, vindicates Quine and should help to dispel the erroneous impression suggested by the results of Montague and Thomason to the effect that classical first-order languages were unable to serve as the basis for logics of the modalities and the propositional attitudes, in effect *forcing* one to employ intensional logics. As our results show, predicate approaches are the more expressive of the two (or, at any rate, *not* the *less* expressive). Clearly more care is required with them in order to avoid inconsistencies, but perhaps this is a direct consequence of their greater expressive power.

[9] However, such sentences may not themselves be regular, and, for that reason, need not be among the sentences that are explicitly believed (or known, or whatever other propositional attitude the logic might be formalizing). An interesting open problem is to identify whether or not these irregular sentences can also be made objects of belief without causing problems.

Acknowledgements

We would like especially to thank Calvin Ostrum for bringing several relevant papers to our attention, and for making helpful suggestions on the overall structure of the paper, many of which have been adopted. Wilf LaLonde and Dave Etherington also gave us invaluable feedback on an earlier draft. This research has been supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

[1] Creary, Lewis G., "Propositional Attitudes: Fregean Representations and Simulative Reasoning," *Proceedings of IJCAI-6*, Volume 1, Tokyo, 1979, 176–181.

[2] Hintikka, Jaakko, *Knowledge and Belief*, Ithaca: Cornell University Press, 1962.

[3] Konolige, Kurt, "A First-Order Formalisation of Knowledge and Action for a Multi-Agent Planning System," J.E. Hayes, D. Michie, and Y.-H. Pao (Eds.) *Machine Intelligence* **10**, Chichester: Ellis Horwood and New York: Halstead Press, 1982, 41–72.

[4] Konolige, Kurt, *A Deduction Model of Belief and its Logics*, Stanford Computer Science Report STAN-CS-84-1022, Stanford, June 1984.

[5] Lakemeyer, Gerhard, "Steps Towards a First-Order Logic of Explicit Knowledge and Belief," *these proceedings.*

[6] Levesque, Hector, "A Formal Treatment of Incomplete Knowledge Bases," FLAIR Technical Report No. 3, Fairchild, Palo Alto, 1982.

[7] Levesque, Hector, "Foundations of a Functional Approach to Knowledge Representation," *Artificial Intelligence* **23**, 1984, 155–215.

[8] McCarthy, John, "First-Order Theories of Individual Concepts and Propositions," in J.E. Hayes, D. Michie, and L.I. Mikulick (eds.), *Machine Intelligence* **9**, Chichester: Ellis Horwood and New York: Halstead Press, 1979, 129–147.

[9] Montague, Richard, "Syntactical Treatment of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability," *Acta Philosophica Fennica* **16**, 1963, 153–167. Reprinted in R. Montague, *Formal Philosophy*, New Haven: Yale University Press, 1974, 286–302.

[10] Montague, Richard, "Theories Incomparable with Respect to Relative Interpretability," *Journal of Symbolic Logic* **27**, 1962, 195–211.

[11] Moore, Robert C., "Reasoning about Knowledge and Action," SRI Technical Note 191, Menlo Park, October 1980.

[12] Morgan, Charles G., "Methods for Automated Theorem Proving in Nonclassical Logics," IEEE *Transactions on Computers* **C-25**, 8, August 1976, 852–862.

[13] Perlis, Donald, "Languages with Self-Reference I: Foundations," *Artificial Intelligence* **25**, 1985, 301–322.

[14] Perlis, Donald, *Language, Computation, and Reality*, Ph.D. dissertation, Department of Computer Science, University of Rochester, Rochester, NY, 1981. (Manuscript received from author.)

[15] Quine, Willard V.O., "Three Grades of Modal Involvement," *Proceedings of the XI^{th} International Congress of Philosophy*, Volume 14, North-Holland, Amsterdam, 1953. Reprinted in W.V.O. Quine, *The Ways of Paradox and Other Essays, Revised and Enlarged Edition*, 158–176. *[Not referenced.]*

[16] Quine, Willard V.O., "Intensions Revisited", in P.A. French, T.E. Uehling Jr., H.K. Wettstein (eds.) *Contemporary Perspectives in the Philosophy of Language*, Minneapolis: University of Minnesota Press, 1979, 268–274.

[17] Tarski, Alfred, "Der Wahrheitsbegriff in den formalisierten Sprachen," *Studia Philosophia* **1**, 1936, 261–405. English translation: "The Concept of Truth in Formalized Languages" appears in A. Tarski, *Logic, Semantics, and Metamathematics*, Oxford, 1956.

[18] Tarski, Alfred, A. Mostowski, and R. Robinson, *Undecidable Theories*, Amsterdam, 1953.

[19] Thomason, Richmond H., "A Note on Syntactical Treatments of Modality," *Synthese* **44**, 1980, 391–395.