356
Notre Dame Journal of Formal Logic
Volume 28, Number 3, July 1987

Quantified Modal Logic and Self-Reference

C. SMORYŃSKI

The propositional modal logic of provability, which I denote *PrL* for Provability Logic, has proven to be a useful tool in studying self-reference in Peano Arithmetic, *PA*. The three chief results about *PrL* are: (i) Solovay's Completeness Theorems, (ii) the de Jongh-Sambin Theorem, and (iii) the Uniqueness Theorem. Solovay's First Completeness Theorem asserts that *PrL* is the modal logic of provability in *PA*, i.e., it axiomatizes those schemata in the language of $\Box$ which are provable in *PA*. Solovay's Second Completeness Theorem tells what schemata are always true assertions when interpreted in the language of arithmetic. These results are not only aesthetically pleasing, but they, particularly the Second Theorem, are extremely useful in establishing incompleteness results—the prototypical applications of self-reference. The de Jongh-Sambin Theorem and the Uniqueness Theorem, although they do not help in establishing the results obtained by self-reference, do tell us something more about self-reference itself: According to the de Jongh-Sambin Theorem, for all reasonable modal formulas $A(p)$, there is a modal formula D not containing p and such that $PrL \vdash D \leftrightarrow A(D)$; the Uniqueness Theorem, due independently to Bernardi, de Jongh, and Sambin, asserts that the fixed point to $A(p)$ is unique.

Whenever one decides to expand the context of the discussion, the first question one asks is whether or not these three results carry over. Thus, for example, in their analysis of Rosser sentences, Guaspari and Solovay expanded the language of *PrL* to accommodate the Rosser trick of comparing witnesses to provability assertions, added a few axioms about these comparisons, proved the completeness of their system with respect to arithmetic interpretations, and then applied this result to show the failure of the Uniqueness Theorem and to make a few observations on definability. In studying multimodal logics and their interpretations, one finds the same results: Suppose, e.g., one has two modal operators $\Box$ and Δ to be interpreted as provability in *PA* and *ZF*, respectively. Carlson has proven three completeness theorems for axiomatizing the schemata

Received May 28, 1985

provable in *PA*, the schemata provable in *ZF*, and the true schemata, respectively, and the author has generalized the de Jongh-Sambin Theorem and the Uniqueness Theorem to this context. Finally, I cite Montagna's work on the modal logic *QGL* obtained by adding the predicate calculus to *PrL*, i.e., expanding *PrL* from propositional modal logic to quantified modal logic. Montagna proved the incompleteness of *QGL* with respect to arithmetic interpretations and the failure of the analogue to the de Jongh-Sambin Theorem; but he did obtain one positive result: The Uniqueness Theorem holds.

The reasons for testing the analogues to the three fundamental results may or may not be good. A completeness theorem shows that one has made a correct analysis of the concept involved. It is not a priori necessary to have a complete analysis of the concept to obtain positive results. Thus, for example, the author's generalization of the uniqueness and explicit definability of fixed points in $\Box$ and Δ, cited above, is actually given in a weak multimodal theory and applies to a variety of settings, including some to be discussed in the present paper. On the other hand, negative results can require a completeness theorem: Guaspari and Solovay proved their arithmetic completeness theorem to show that the nonuniqueness result in their modal system had arithmetic significance. Similarly, one of the goals of the present paper is to prove that Montagna's nondefinability result in *QGL* yields a nondefinability result in *PA*. This is accomplished by mimicking the completeness proof of Solovay.

The status of the uniqueness and explicit definability theorems is different: These are not normative results, but were surprises to the experts. The explicit definability and uniqueness results are very special and can be viewed as manifestations of the limited expressibility of the propositional modal language, and their failure in a given context can thus, perhaps, be interpreted as a sign of some minimal expressive power. Counterexamples are to be expected—perhaps even sought. One of the goals of the present paper is to clarify, however slightly, the extent of these two theorems. We shall see that an axiomatic analysis of their proofs yields a very general result and that more general results usually are not possible. (Actually, half of this has already been done by the author and will merely be cited.)

The plan of this paper is as follows: In Section 1 I will describe, without proof, the result of my multimodal analysis of the de Jongh-Sambin Theorem. This will give us some idea of the extent of the de Jongh-Sambin Theorem and the Uniqueness Theorem, leaving us with the task of showing the conditions cited to be best possible—a task not completely performed. The result will also be applied directly in the sequel. Section 2 reviews *QGL* and Montagna's undefinability result. The perspective obtained from the result of the previous section suggests that Montagna's formula, albeit quite simple, is logically too complex and a simplification is obtained, still using Montagna's delightful proof. These counterexamples cease to be counterexamples if the Barcan Schema is added to *QGL*, and I thus give an example of the undefinability of a fixed point when *BS* is added to *QGL*. The Barcan Schema fails generally under the arithmetic interpretation and one might question the arithmetic significance of *QGL* + *BS*. Although I cannot really answer such a question satisfactorily, in the next section I adapt a proof of Solovay's First Completeness Theorem to this counterexample and thereby obtain an undefinability result in *PA*.

1 The multi-modal analysis of self-reference The proof of the de Jongh-Sambin Theorem has two parts. The first, crucial, part consists of deriving $\Box C(t) \leftrightarrow \Box C[\Box C(t)]$ in *PrL*. Once one has this, a fairly algebraic argument using only the Substitution Lemma yields the general result. We get some idea of the general scope of validity of the Theorem by introducing a new operator $\nabla p = \Box C(p)$ and asking what properties of ∇ are needed in the derivation of the equivalence cited. As I pointed out already in [6], we use the completeness schema,

$$\nabla A \rightarrow \Box \nabla A.$$

The Substitution Lemma in *PrL* also yields the schema,

$$\Box (A \leftrightarrow B) \rightarrow . \nabla A \leftrightarrow \nabla B.$$

It turns out that this is enough—in a sense I will shortly make precise.

Definition SR^- is the system of bimodal logic with language, axioms, and rules of inference as follows:

Language:
Propositional variables: $p, q, r, \ldots$
Truth values: t, f
Propositional connectives: $\sim, \wedge, \vee, \rightarrow$
Modal operators: $\Box, \nabla$.

Axioms:
A1 All (Boolean) tautologies
A2 $\Box A \wedge \Box (A \rightarrow B) \rightarrow \Box B$
A3$_\Box$ $\Box A \rightarrow \Box \Box A$
A4 $\Box (\Box A \rightarrow A) \rightarrow \Box A$
A5 $\Box (A \leftrightarrow B) \rightarrow . \nabla A \leftrightarrow \nabla B.$

Rules:
R1 $A, A \rightarrow B / B$
R2 $A / \Box A.$

If we ignore ∇, Axioms A1–A4 and Rules R1–R2 are just the axioms and rules of *PrL*. Axiom schema A5 is a substitutivity schema. It is the minimal axiom needed for the general Substitution Lemma.

Substitution lemma *Let* $A(p)$ *be given.*
(i) $SR^- \vdash \boxed{\Box} (B \leftrightarrow C) \rightarrow . A(B) \leftrightarrow A(C)$
(ii) $SR^- \vdash \Box (B \leftrightarrow C) \rightarrow \Box [A(B) \leftrightarrow A(C)]$,

where $\boxed{\Box}$ *is the so-called* strong box, *defined by* $\boxed{\Box} D = D \wedge \Box D.$

The Substitution Lemma, the Formalized Löb's Theorem for $\Box$ (A4), and the following notion are all that are necessary to prove the Uniqueness Theorem.

Definition Let $A(p)$ be a formula of the language of SR^-. We say p is *modalized* in $A(p)$ if every occurrence of p in $A(p)$ is inside the scope of one of the modal operators.

Uniqueness theorem *Let p be modalized in $A(p)$.*
(i) $SR^- \vdash \boxed{S}[p \leftrightarrow A(p)] \wedge \boxed{S}[q \leftrightarrow A(q)] \rightarrow .p \leftrightarrow q$
(ii) $SR^- \vdash \Box[p \leftrightarrow A(p)] \wedge \Box[q \leftrightarrow A(q)] \rightarrow \Box(p \leftrightarrow q)$.

As I noted earlier, the generalization of the de Jongh-Sambin Theorem needs an extra assumption on $\triangledown$.

Definition SR is the modal system extending SR^- by the addition of the axiom schema, $\mathbf{A3}_{\triangledown}$: $\triangledown A \rightarrow \Box\triangledown A$.

Explicit definability theorem *Let p be modalized in $A(p)$. There is a sentence D of the bimodal language possessing only those atoms of $A(p)$ other than p and such that*

$$SR \vdash D \leftrightarrow A(D).$$

I do not propose to prove these results here. The proofs can be found in Chapter 4 of [7]. I note merely that the proof of the Uniqueness Theorem in SR^- adapts Bernardi's proof from [1] for *PrL* and the proof of the Explicit Definability Theorem requires an adaptation of Sambin's Extension of Löb's Theorem from [5].

As noted in the introduction, one of the goals of this paper is to see to what extent the above results are best possible. Before attempting this, let me quickly note an easy improvement: Instead of adding a single new operator $\triangledown$ to *PrL* and appropriate axioms to obtain SR^- and SR, add n new operators $\triangledown_1, \ldots, \triangledown_n$ and axiom schemata

$\mathbf{A3}_{\triangledown_i}$ $\triangledown_i A \rightarrow \Box\triangledown_i A$, $i = 1, \ldots, n$
$\mathbf{A5}_{\triangledown_i}$ $\Box(A \leftrightarrow B) \rightarrow .\triangledown_i A \leftrightarrow \triangledown_i B$, $i = 1, \ldots, n$,

as appropriate to obtain SR_n^- and SR_n. The Explicit Definability and Uniqueness Theorems carry over to these extended theories.

To see to what extent SR^- and SR (or, SR_n^- and SR_n) are arithmetically best possible, we must consider arithmetic interpretations of the bimodal (or, multimodal) language. An arithmetic interpretation $*$ is inductively defined by first choosing arithmetic sentences p_i^* for propositional atoms p_i and then extending the map by: t^* is $\bar{0} = \bar{0}$, f^* is $\bar{0} = \bar{1}$, $(\sim A)^*$ is $\sim(A^*)$, $(\Box A)^*$ is $Pr_{PA}(\ulcorner A^* \urcorner)$, $(A \circ B)^*$ is $A^* \circ B^*$ for $\circ \in \{\wedge, \vee, \rightarrow\}$, and $(\triangledown A)^*$ will be $\rho(\ulcorner A^* \urcorner)$ for an appropriate formula ρ. What makes ρ appropriate? Well, according to application, the axioms of SR^- and/or SR ought to be provable schemata in PA.

Definition A formula $\rho(v)$ of the language of arithmetic, with only the free variable indicated, is *substitutable* if, for all sentences ϕ, ψ,

$$PA \vdash Pr_{PA}(\ulcorner \phi \leftrightarrow \psi \urcorner) \rightarrow .\rho(\ulcorner \phi \urcorner) \leftrightarrow \rho(\ulcorner \psi \urcorner).$$

Noting that the class of substitutable formulas is closed under the usual logical operations and application of Pr_{PA} and ρ, the full content of the Uniqueness Theorem is the uniqueness of the single fixed point $p \leftrightarrow \triangledown p$. Arithmetically, this yields:

Arithmetic uniqueness theorem *Let $\rho(v)$ be substitutable. For any sentences ϕ, ψ,*

$$PA \vdash \phi \leftrightarrow \rho(\ulcorner\phi\urcorner) \;\&\; PA \vdash \psi \leftrightarrow \rho(\ulcorner\psi\urcorner) \Rightarrow PA \vdash \phi \leftrightarrow \psi.$$

Proof: Observe,

(*) $SR^- \vdash \boxed{\mathbb{S}}[p \leftrightarrow \nabla p] \wedge \boxed{\mathbb{S}}[q \leftrightarrow \nabla q] \rightarrow .p \leftrightarrow q.$

Let ϕ, ψ be fixed points of ρ:

$$PA \vdash \phi \leftrightarrow \rho(\ulcorner\phi\urcorner), \qquad PA \vdash \psi \leftrightarrow \rho(\ulcorner\psi\urcorner).$$

Then

$$PA \vdash Pr_{PA}(\ulcorner\phi \leftrightarrow \rho(\ulcorner\phi\urcorner)\urcorner), \qquad PA \vdash Pr_{PA}(\ulcorner\psi \leftrightarrow \rho(\ulcorner\psi\urcorner)\urcorner),$$

and, letting $p^* = \phi$, $q^* = \psi$, we have proven the premise of the interpretation of (*) in PA. Thus, PA proves the conclusion: $(p \leftrightarrow q)^*$; i.e., $PA \vdash \phi \leftrightarrow \psi$.

The question of the generality of the derivation of the Uniqueness Theorem in SR^- can now be viewed as asking if the Arithmetic Uniqueness Theorem is the best possible; i.e., can the condition of the substitutability of ρ be weakened? Modally, substitutability seems natural enough. Arithmetically, it appears more a proof-generated concept. More natural would be extensionality:

Definition A formula $\rho(v)$ of the language of arithmetic, with only the free variable indicated, is *extensional* if, for all sentences ϕ, ψ,

$$PA \vdash \phi \leftrightarrow \psi \Rightarrow PA \vdash \rho(\ulcorner\phi\urcorner) \leftrightarrow \rho(\ulcorner\psi\urcorner).$$

Midway between extensionality and substitutability is provable extensionality:

Definition A formula $\rho(v)$ of the language of arithmetic, with only the free variable indicated, is *provably extensional* if, for all sentences ϕ, ψ,

$$PA \vdash Pr(\ulcorner\phi \leftrightarrow \psi\urcorner) \rightarrow Pr(\ulcorner\rho(\ulcorner\phi\urcorner) \leftrightarrow \rho(\ulcorner\psi\urcorner)\urcorner).$$

Does the Arithmetic Uniqueness Theorem hold for arbitrary extensional or provably extensional formulas? The answer is *no*.

Counterexample By the Orey Compactness Theorem, there is a formula $Tr_M(v)$ that offers within PA the truth definition for a model of PA. Now, it happens that

$$PA \nvdash \forall v(Pr_{PA}(v) \rightarrow Tr_M(v)),$$

but (as pointed out by the referee and not immediately obvious),

$$PA \vdash \forall v(Pr_{PA}(v) \rightarrow Pr_{PA}(\ulcorner Tr_M(v)\urcorner).$$

From this, provable extensionality follows easily: For any ϕ, ψ,

$$\begin{aligned} PA &\vdash Pr_{PA}(\ulcorner\phi \leftrightarrow \psi\urcorner) \rightarrow Pr_{PA}(\ulcorner Tr_M(\ulcorner\phi \leftrightarrow \psi\urcorner)\urcorner) \\ &\vdash Pr_{PA}(\ulcorner\phi \leftrightarrow \psi\urcorner) \rightarrow Pr_{PA}(\ulcorner Tr_M(\ulcorner\phi\urcorner) \leftrightarrow Tr_M(\ulcorner\psi\urcorner)\urcorner), \end{aligned}$$

by the fact that Tr_M is a truth definition. But now let ϕ be any theorem of PA and observe

$$PA \vdash \phi \Rightarrow PA \vdash Tr_M(\ulcorner \phi \urcorner) \wedge \phi$$
$$\Rightarrow PA \vdash \phi \leftrightarrow Tr_M(\ulcorner \phi \urcorner], \quad (1)$$

$$\Rightarrow PA \vdash \sim\phi \leftrightarrow \sim Tr_M(\ulcorner \phi \urcorner)$$
$$\Rightarrow PA \vdash \sim\phi \leftrightarrow Tr_M(\ulcorner \sim \phi \urcorner), \quad (2)$$

again since Tr_M is a truth definition for some model. By (1) and (2), ϕ and $\sim\phi$ are fixed points to $Tr_M(v)$, and they clearly cannot be proven equivalent.

And what about explicit definability? First, what is the arithmetic meaning of my Explicit Definability Theorem for *SR*? Again, Axiom schema A5 requires $\rho(v)$ to be substitutable. Schema A3 requires

$$PA \vdash \rho(\ulcorner \phi \urcorner) \rightarrow Pr_{PA}(\ulcorner \rho(\ulcorner \phi \urcorner) \urcorner),$$

for all sentences ϕ. Now, the most general condition on ρ guaranteeing this derivability assertion is that ρ be Σ_1.

Arithmetic explicit definability theorem Let ρ be a substitutable Σ_1-formula with only v free and specify that ∇ is to be interpreted by ρ under arithmetic interpretations $*$. Let p be modalized in $A(p)$. There is a sentence D in the bimodal language such that, for all $*$, $PA \vdash D^* \leftrightarrow A(D)^*$.

Proof: For some D, $SR \vdash D \leftrightarrow A(D)$. By choice of ρ, one has

$$SR \vdash B \Rightarrow PA \vdash B^*$$

for all $*$. In particular, $PA \vdash D^* \leftrightarrow A(D)^*$.

We can now ask ourselves if this can be improved. In purely modal terms, this means: Can we weaken *SR* and still derive the modal form of the Explicit Definability Theorem? In arithmetic terms, this means: Can we weaken the assumption that ρ be a substitutable Σ_1-formula and still derive the Arithmetic Definability Theorem? An affirmative answer to the former question yields an affirmative answer to the later; a negative answer to the latter yields a negative answer to the former.

Of course, if ρ is a substitutable Π_1-formula, $\sim\rho$ is a substitutable Σ_1-formula and the Arithmetic Explicit Definability Theorem will hold for ρ.

The question arises: Does the Arithmetic Explicit Definability Theorem hold for arbitrary substitutable ρ? The answer is no, as we shall see in the sequel. It follows that we cannot drop axiom $A3_\nabla$ and derive the Explicit Definability Theorem in SR^-.

And what about substitutability? Can it be dropped as an assumption? The only negative result that comes to mind concerns the Rosser sentence. In [3], Guaspari and Solovay extended the language of *PrL* by a witness comparison, $\preccurlyeq$. The formula

$$\nabla p = \Box p \preccurlyeq \Box \sim p$$

has the intended interpretation

$$\rho(v): \quad \exists v_0[Prov(v_0, v) \wedge \forall v_1 < v_0 \sim Prov(v_1, neg(v))],$$

where $\mathrm{neg}(\ulcorner\phi\urcorner) = \ulcorner\sim\phi\urcorner$. In their modal logic, they show

$$D \leftrightarrow \nabla\sim D$$

cannot be derived for any sentence D not mentioning p. Their arithmetic completeness result, by which we expect a counterexample to the Arithmetic Explicit Definability Theorem for this ρ, doesn't quite yield what we want. They replace *Prov* by *Der*, where

$$PA \vdash \forall v[\exists v_0 Prov(v_0, v) \leftrightarrow \exists v_0 Der(v_0, v)];$$

i.e., their arithmetic interpretations of ∇ vary ρ as well as the variables. This is enough to show that, if we drop the substitutability condition on ρ, we do not get the uncited *uniformity* of the Arithmetic Explicit Definability Theorem, whence the modal Explicit Definability Theorem cannot be derived if we drop axiom schema A5 from *SR*.

How far does this example take us? Another thing Guaspari and Solovay's analysis shows is that ∇ is not extensional: One can fail to derive even $\nabla p \leftrightarrow \nabla(p \wedge p)$, even though $p \leftrightarrow p \wedge p$ is derivable. Hence, many of the interpretations ρ must fail to be extensional. Thus, we can ask:

Question Does the Arithmetic Explicit Definability Theorem hold for all Σ_1-formulas $\rho(v)$ with only v free?

We have already seen the answer is no if we want some modal uniformity to the definition. If we ask about weakening, instead of dropping, substitutability, the question becomes:

Question Does the Arithmetic Explicit Definability Theorem hold for extensional, Σ_1-formulas $\rho(v)$? provably extensional, Σ_1-formulas $\rho(v)$?

For uniform positive results, the obvious approach is to work modally:

Question Can one prove the modal Explicit Definability Theorem if one replaces A5 in *SR* by

A5′ $\Box(A \leftrightarrow B) \rightarrow \Box(\nabla A \leftrightarrow \nabla B)$,

or

R3 $A \leftrightarrow B / \nabla A \leftrightarrow \nabla B$?

These three questions have all been solved negatively by Albert Visser since I posed them as open problems in the original draft of this paper. Here I give the weaker counterexample: the Arithmetic Explicit Definability Theorem fails for logically complex $\rho(v)$. This is done in Section 3, below. Like the example of imitation Rosser sentences, it rests on a modal undefinability result, to which we turn in Section 2.

2 Self-reference in quantified modal logic We begin with a definition:

Definition The system *QGL* is the system with language, axioms, and rules of inference as follows:

Language:
Individual variables: $v_0, v_1, \ldots$
Propositional variables: $p, q, r, \ldots$
Truth values: t, f
n-ary predicate symbols: $R_0^n, R_1^n, \ldots$
Propositional connectives: $\sim, \wedge, \vee, \rightarrow$
Quantifiers: $\forall, \exists$
Modal operator: $\Box$.

Axioms:
A1 All instances of axioms of the predicate calculus in this language
A2 $\Box A \wedge \Box(A \rightarrow B) \rightarrow \Box B$
A3 $\Box A \rightarrow \Box\Box A$
A4 $\Box(\Box A \rightarrow A) \rightarrow \Box A$.

Rules:
R1 $A, A \rightarrow B/B$
R2 $A/\Box A$.

Under A1 is assumed some axiomatization of the predicate calculus requiring only R1.

The reason for studying *QGL* is obscure to me, perhaps merely generalization for generalization's sake. Be that as it may, Montagna's paper has shown *QGL* to be quite interesting. My favorite result of his paper, to which the present paper is a response, is the following:

Counterexample 1 Let $A(p) = \forall u \exists v \Box(p \rightarrow Ruv)$. There is no sentence D in the language of *QGL* for which $QGL \vdash D \leftrightarrow A(D)$.

Except for the usual problem involving the clash of free and bound variables, which problem vanishes if we adopt the simple device of using separate free and bound variables, the Substitution Lemma holds in *QGL* and we may deem $A(p)$ (better: any arithmetic interpretation of $A(p)$) substitutable. My intuition (or: confusion—not only are we not yet talking about arithmetic interpretations, but *QGL* was demonstrated by Montagna to be incomplete with respect to arithmetic interpretations) is that the block to the definability of any fixed point to $A(p)$ is the quantifier alternation $\forall u \exists v$. Now, $\Box$ implicitly has an existential quantifier, whence $\exists v$ ought not to be really necessary:

Counterexample 2 Let $A(p) = \forall u \Box(p \rightarrow Pu)$. There is no sentence D in the language of *QGL* for which $QGL \vdash D \leftrightarrow A(D)$.

Proof: The proof follows Montagna's proof for his counterexample. It is done by giving a Kripke model $\mathbf{K} = (K, R, D, \Vdash)$, where

i. K is a nonempty set of nodes α
ii. $R \subseteq K \times K$ is a partial ordering
iii. D is a map associating a nonempty set D_α with each $\alpha \in K$; it is assumed that $\alpha R \beta \Rightarrow D_\alpha \subseteq D_\beta$
iv. $\Vdash$ is the forcing relation.

Montagna's paper presents a completeness proof for *QGL* with respect to those models it is valid in. For us, the crucial thing is soundness: *QGL* is valid in any Kripke model in which the converse to R is well-founded.

We prove the result by presenting one model **K** in which each equivalence $D \leftrightarrow A(D)$ fails. For the domain K of the model, we take the set ω of natural numbers. R is the converse to the natural ordering of these numbers,

$$xRy \text{ iff } y < x.$$

For the domains D_x, we simply truncate the integers:

$$D_x = \{x, x+1, \dots\} = \omega - x.$$

Finally, we define $\Vdash$ by

$$x \Vdash P\bar{y} \text{ iff } x = y \text{ or } y > x+1.$$

(All other predicates can be ignored or trivially interpreted.) Pictorially, we have:

$$\begin{array}{lll} 0 & \{0, 1, 2, \dots\} & \sim P1 \\ & \quad | & \\ 1 & \quad [1, 2, \dots\} & \sim P2 \\ & \quad\ | & \\ 2 & \quad\ \{2, \dots\} & \sim P3 \\ & \quad\ | & \\ & \quad\ . & \\ & \quad\ . & \\ & \quad\ . \quad . & \end{array}$$

The key to the undefinability result is the utter simplicity of the structure. Its components, the set K, the relation R, the relation $x \in D_y$, and the atomic forcing relation $x \Vdash P\bar{y}$, are all definable in the structure $(\omega, =, <, S, 0)$. The definability of forcing extends to all formulas of *QGL* via the clauses in the definition of forcing:

$x \Vdash \sim B$: $\sim(x \Vdash B)$
$x \Vdash B \circ C$: $(x \Vdash B) \circ (x \Vdash C)$ for $\circ \in \{\wedge, \vee, \rightarrow\}$
$x \Vdash \Box B$: $\forall y > x (y \Vdash B)$
$x \Vdash \forall u B u$: $\forall u \geq x (x \Vdash Bu)$
$x \Vdash \exists u B u$: $\exists u \geq x (x \Vdash Bu)$.

The quantifier elimination for the theory of $(\omega, =, <, S, 0)$ yields

Claim 1 *For any sentence B in the language of QGL, $\{x: x \Vdash B\}$ is either finite or cofinite.*

The undefinability result follows from this and one further

Claim 2 *Assume $p \leftrightarrow A(p)$ is valid in the model* **K**. *Then: For all $x \in \omega$, $x \Vdash p$ iff x is even.*

Proof: By induction on x.

Basis: $x = 0$ or $x = 1$. $0 \Vdash p$ since 0 forces any boxed statement.
For $x = 1$, observe that $1 \in D_1$ and
$1R0 \Vdash p \wedge \sim P\bar{1}$,

whence $1 \nVdash \Box(p \to P\bar{1})$
$\nVdash \forall u \Box(p \to Pu)$.

Induction step: First observe that
$2x \Vdash \Box P\bar{y}$ for all $y \geq 2x + 1$
and $2x \Vdash \Box(p \to P\overline{2x})$
since $2x - 1 \nVdash p$
and $2x - 1$ is the only node above $2x$ failing to force $P\overline{2x}$. Thus,
$\forall y \geq 2x (2x \Vdash \Box(p \to P\bar{y}))$;
i.e., $2x \Vdash \forall u \Box(p \to Pu)$;
i.e., $2x \Vdash p$.

For the odd case, observe

$$2x \Vdash p \;\&\; 2x \nVdash P(\overline{2x+1}) \Rightarrow 2x+1 \nVdash \Box(p \to P(\overline{2x+1})) \Rightarrow 2x+1 \nVdash \forall u \Box(p \to Pu) \Rightarrow 2x+1 \nVdash p.$$

This completes the proof of the Claim and therewith that of the undefinability result.

Given Montagna's published proof, I need not have gone into as much detail as I did, but I really am taken with the proof. I shall show more restraint with my next counterexample.

Why do I need another counterexample? Well, as much as I like these counterexamples, they are counterexamples in modal logic, not in arithmetic. Since *QGL* is not complete with respect to arithmetic interpretations, this counterexample merely makes it plausible that the Arithmetic Definability Theorem fails for complex substitutable formulas $\rho(v)$. The varying domains of **K** would complicate the arithmetic simulation of the counterexample. Thus, I want a counterexample to explicit definability in *QGL* using a model with constant domains. This is readily obtained from the model already at hand. Let $A(p) = \forall u[Qu \to \Box(P \to Pu)]$. The model will have ω as its constant domain and Q will single out the old domains. Before making a formal statement of the counterexample, I digress to discuss a couple of related matters.

In quantified modal logic, the syntactic counterpart of models with constant domains is the Barcan schema,

BS $\forall u \Box Au \to \Box \forall u Au$,

and its converse,

$\Box \forall u Au \to \forall u \Box Au$.

The converse is provable in *QGL*:

$QGL \vdash \Box(\forall u Au \to Au)$	by A1, R2
$\vdash \forall u \Box(\forall u Au \to Au)$	
$\vdash \forall u(\Box \forall u Au \to \Box Au)$	by A2
$\vdash \Box \forall u Au \to \forall u \Box Au$	by predicate logic.

In a similar manner,

$$QGL \vdash \exists u \Box Au \to \Box \exists u Au.$$

Thus, if $A = Q_1 u_1 \ldots Q_k u_k \Box B$, with any string $Q_1, \ldots, Q_k$ of quantifiers,

$$QGL + BS \vdash A \to \Box A.$$

Because of this, the counterexamples to explicit definability fail to be counterexamples when *BS* is added to *QGL* (and *R*2 retained).

Counter-counterexample Let $A(p) = \forall u_1 \exists u_2 \ldots Q_k u_k \Box (p \to Pu_1 \ldots u_k)$, where $k \geq 1$ and the quantifiers alternate. Let $D = A(t) = \forall u_1 \exists u_2 \ldots Q_k u_k \Box$-$Pu_1 \ldots u_k$. Then:

$$QGL + BS \vdash D \leftrightarrow A(D).$$

As we have just seen,

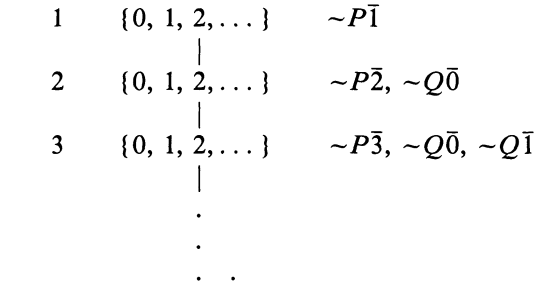
$$QGL + BS \vdash A(p) \to \Box A(p).$$

Thus, we can interpret *SR* in *QGL* + *BS* by making $\nabla p = A(p)$ and appeal to the Explicit Definability Theorem to conclude the existence of *D*. The exact form, $D = \nabla t$, is verified by looking at the actual proof of the Explicit Definability Theorem (for which cf. Chapter 4 of [7]).

This counter-counterexample generalizes the two counterexamples given for *QGL*. Interpreting SR_n in *QGL* + *BS*, we see that no propositional combination of formulas of the form $Q_1 u_1 \ldots Q_k u_k \Box B$ will give us counterexamples in *QGL* + *BS*. We need a counterexample in which the quantifiers cannot be pushed against the box. This happens with the counterexample promised:

Counterexample 3 Let $A(p) = \forall u [Qu \to \Box (p \to Pu)]$. There is no modal sentence *D* in the language of *QGL* such that $QGL + BS \vdash D \leftrightarrow A(D)$.

This doesn't quite reduce to Counterexample 2 because *Q* was previously ignored (or, trivially interpreted) and it now singles out the domains. Nonetheless, the proof is virtually identical, because the relation $x \Vdash Q\bar{y}$ is also definable in $(\omega, =, <, S, 0)$. Let me simply describe the model **K**:



I have relabeled the nodes of *K* in preparation for the next section. Forcing of the predicates *P*, *Q* is still definable:

$$x \Vdash P\bar{y} \text{ iff } y \neq x$$
$$x \Vdash Q\bar{y} \text{ iff } x < y + \bar{2}.$$

This model is not quite as nice as it could be. The fact that 0 is treated differently from the other elements of the domain is inelegant. If we decide not

to force $P\bar{0}$ at the top and slide all the other failures down, we get the more elegant model:

$$\begin{array}{l} 1 \cdot \sim P\bar{0} \\ 2 \cdot \sim P\bar{1},\ \sim Q\bar{0} \\ 3 \cdot \sim P\bar{2},\ \sim Q\bar{0},\ \sim Q\bar{1} \\ \cdot \\ \cdot \\ \cdot\ \cdot \end{array}$$

In this model, Qu is equivalent to $\Box Pu$. Thus, we can eliminate Q:

Counterexample 4 Let $A(p) = \forall u[\Box\Box Pu \rightarrow \Box(p \rightarrow Pu)]$. There is no modal sentence D in the language of QGL such that $QGL + BS \vdash D \leftrightarrow A(D)$.

Once again, for any sentence B in the language of QGL, the set $\{x: x \Vdash B\}$ is definable in the language of $(\omega, =, <, S, 0)$, whence finite or cofinite. And, once again, if we assume $p \leftrightarrow A(p)$ throughout K, p is forced at alternate nodes—this time the odd-numbered ones. I leave the details to the reader.

Our next goal is to transform this latest counterexample from a modal to an arithmetic one.

3 An arithmetic counterexample The goal of the present section is to prove the following:

Theorem *Let* $A(p) = \forall u[\Box\Box Pu \rightarrow \Box(p \rightarrow Pu)]$. *There is no sentence* D *in the language of QGL such that, for all arithmetic interpretations* $*$, $PA \vdash D^* \leftrightarrow A(D)^*$.

An arithmetic interpretation $*$ of the language of QGL is defined by mapping propositional atoms p to sentences p^* and predicates $Pu_1 \ldots u_k$ to formulas $P^*u_1 \ldots u_k = \phi u_1 \ldots u_k$. Given this beginning, the rest of the interpretation is automatic: t^* is $\bar{0} = \bar{0}$, f^* is $\bar{0} = \bar{1}$, $(\sim A)^* = \sim A^*$, $(A \circ B)^* = A^* \circ B^*$ for $\circ \in \{\wedge, \vee, \rightarrow\}$, $(QuAu)^* = Qu(A^*u)$ for $Q \in \{\forall, \exists\}$ $(\Box Au_1 \ldots u_k)^* = Pr_{PR}(\ulcorner A^*\dot{u}_1 \ldots \dot{u}_k \urcorner)$, where we may relabel bound variables in A^*u if necessary and where $\ulcorner \phi\dot{u}_1 \ldots \dot{u}_k \urcorner$ denotes $\text{sub}(\ulcorner \phi u_1 \ldots u_k \urcorner, num(u_1), \ldots, num(u_k))$, the code of the result of substituting the u_i-th numeral for the variable u_i in ϕ, e.g., $sub(\ulcorner \phi uv \urcorner, \ulcorner \bar{3} \urcorner, \ulcorner \bar{4} \urcorner) = \ulcorner \phi(\bar{3}, \bar{4}) \urcorner$.

To prove the theorem, we will construct a single interpretation $*$ by simulating the Kripke model $\mathbf{K} = (\{1,2,\ldots\}, R, D, \Vdash)$ of Counterexample 4 of the preceding section. This simulation is a straightforward application of the method used by Solovay in proving the Arithmetic Completeness Theorem for PrL.

We begin by setting $0Rx$ for all $x \in K$. Then we define a function F via the Recursion Theorem by

$$F0 = 0$$

$$F(x+1) = \begin{cases} y,\ Prov_{PA}(x+1, \ulcorner L \neq \bar{y} \urcorner)\ \&\ FxRy \\ Fx, \text{ otherwise (i.e., no such } y \text{ exists}), \end{cases}$$

where $L = \lim_{x \to \infty} F(x)$: The formula $L = v$ is $\exists v_0 \forall v_1 > v_0 (Fv_1 = v)$.

The only difference between this function F and that used by Solovay is that this F is trying (not) to grow through an *infinite* frame $(K \cup \{0\}, R)$. This frame is so close to being finite that all of Solovay's results and proofs carry over and I need only list the most relevant ones:

Lemma 1 *$PA \vdash L$ exists; i.e., $PA \vdash \exists v \exists v_0 \forall v_1 > v_0 (Fv_1 = v)$.*

Lemma 2

(i) $PA \vdash L = u > \bar{0} \wedge uRv \rightarrow \sim Pr_{PA}(\ulcorner L \neq \dot{v} \urcorner)$
(ii) $PA \vdash L = u > \bar{0} \wedge u \neq v \wedge \sim uRv \rightarrow Pr_{PA}(\ulcorner L \neq \dot{v} \urcorner)$
(iii) $PA \vdash L = u > \bar{0} \rightarrow Pr_{PA}(\ulcorner L \neq \dot{u} \urcorner)$
(iv) $PA \vdash L = \bar{0} \wedge v > \bar{0} \rightarrow \sim Pr_{PA}(\ulcorner L \neq \dot{v} \urcorner)$.

Restated, Lemmas 2(i) and 2(iv) assert

(i) $PA \vdash L = u > \bar{0} \wedge uRv \rightarrow Con(PA + L = \dot{v})$
(iv) $PA \vdash L = \bar{0} \wedge v > \bar{0} \rightarrow Con(PA + L = \dot{v})$.

In particular, for any $x > 0$,

$$PA \vdash L = \bar{0} \rightarrow Con(PA + L = \bar{x}).$$

Lemma 3 $PA \vdash L = u > \bar{0} \rightarrow Pr_{PA}(\ulcorner \exists v (\dot{u}Rv \wedge L = v) \urcorner)$.

Lemma 4

(i) *$L = \bar{0}$ is true*
(ii) *For any $x > 0$, $PA + L = \bar{x}$ is consistent.*

These lemmas yield all the information we will need about F. The interested reader can consult Solovay's original paper [8] for the proofs.

We wish to use the limit L to simulate the model $\mathbf{K}$ within PA. To this end, we first note that $\Vdash$ is definable in the sublanguage of that of PA given by the language of the structure $(\omega, =, <, S, 0)$:

"$u \Vdash Pv$": $u \neq \bar{S}v$
"$u \Vdash \sim B$": $\sim$"$u \Vdash B$"
etc.

To simulate this, we define $*$ by

$$P^*v\text{: } L \neq \bar{S}v$$

(and any other predicate is trivially interpreted).

Lemma 5 *For any formula B of the language of QGL,*

(i) $PA \vdash$ "$u \Vdash B$" $\wedge\, u > \bar{0} \rightarrow (L = u \rightarrow B^*)$
(ii) $PA \vdash$ "$u \nVdash B$" $\wedge\, u > 0 \rightarrow (L = u \rightarrow \sim B^*)$.

Proof: By induction on the complexity of B.
Basis: For $B = t, f$ (or having any atomic predicate other than P) the result is trivial. For $B = Pv$, the implications (i) and (ii) are,

$$u \neq \bar{S}v \wedge u > \bar{0} \rightarrow (L = u \rightarrow L \neq \bar{S}v)$$
$$u = \bar{S}v \wedge u > 0 \rightarrow (L = u \rightarrow L = \bar{S}v),$$

respectively. But these are clearly derivable in PA.

Induction step: The propositional cases, $B = \sim C$, $C \wedge D$, $C \vee D$, or $C \to D$ are trivial.

Let $B = \Box C$: This is not quite the same as in Solovay's proof. Observe,

$$PA \vdash \text{“}u \Vdash \Box C\text{”} \wedge u > \bar{0} \to \forall v(uRv \to \text{“}v \Vdash C\text{”}) \\ \to \forall v(uRv \to (L = v \to C^*)),$$

by induction hypothesis. Thus,

$$PA \vdash Pr_{PA}(\ulcorner \text{“}\dot{u} \Vdash \Box C\text{”} \wedge \dot{u} > \bar{0} \to \forall v(\dot{u}Rv \to (L = v \to C^*))\urcorner) \\ \vdash Pr_{PA}(\ulcorner \text{“}\dot{u} \Vdash \Box C\text{”} \wedge \dot{u} > \bar{0}\urcorner) \to Pr_{PA}(\ulcorner \forall v(\dot{u}Rv \to (L = v \to C^*))\urcorner). \quad (1)$$

Once again we use the quantifier-elimination for the language of $(\omega, =, <, S, 0)$: “$u \Vdash \Box C$” is quantifier-free, whence

$$PA \vdash \text{“}u \Vdash \Box C\text{”} \to Pr_{PA}(\ulcorner \text{“}\dot{u} \Vdash \Box C\text{”}\urcorner).$$

Similarly,

$$PA \vdash u > \bar{0} \to Pr_{PA}(\ulcorner \dot{u} > \bar{0}\urcorner),$$

whence (1) yields

$$PA \vdash \text{“}u \Vdash \Box C\text{”} \wedge u > \bar{0} \to Pr_{PA}(\ulcorner \forall v(\dot{u}Rv \to (L = v \to C^*))\urcorner). \quad (2)$$

Now, Lemma 3 yields

$$PA \vdash L = u \wedge u > \bar{0} \to Pr_{PA}(\ulcorner \exists v(\dot{u}Rv \wedge L = v)\urcorner).$$

With (2) this yields

$$PA \vdash \text{“}u \Vdash \Box C\text{”} \wedge L = u \wedge u > \bar{0} \to Pr_{PA}(\ulcorner C^*\urcorner) \\ \vdash \text{“}u \Vdash \Box C\text{”} \wedge u > \bar{0} \to (L = u \to (\Box C)^*).$$

Next, observe

$$PA \vdash \text{“}u \nVdash \Box C\text{”} \wedge u > \bar{0} \to \exists v(uRv \wedge \text{“}v \nVdash C\text{”} \wedge v > \bar{0}) \\ \to \exists v(uRv \wedge Pr_{PA}(\ulcorner \text{“}\dot{v} \nVdash C\text{”} \wedge \dot{v} > \bar{0}\urcorner),$$

again using the fact that “$v \nVdash C$” is quantifier-free. Thus,

$$PA \vdash \text{“}u \nVdash \Box C\text{”} \wedge u > \bar{0} \to \exists v(uRv \wedge Pr_{PA}(\ulcorner L = \dot{v} \to \sim C^*\urcorner),$$

by induction hypothesis, whence

$$PA \vdash \text{“}u \nVdash \Box C\text{”} \wedge u > \bar{0} \to \exists v(uRv \wedge Pr_{PA}(\ulcorner C^* \to L \neq \dot{v}\urcorner)) \\ \to \exists v[uRv \wedge (Pr_{PA}(\ulcorner C^*\urcorner) \to Pr_{PA}(\ulcorner L \neq \dot{v}\urcorner))]. \quad (*)$$

But now recall Lemma 2(i):

$$PA \vdash L = u \wedge u > \bar{0} \wedge uRv \to \sim Pr_{PA}(\ulcorner L \neq \dot{v}\urcorner).$$

Together with (*) this yields

$$PA \vdash \text{“}u \nVdash \Box C\text{”} \wedge u > \bar{0} \wedge L = u \to \sim Pr_{PA}(\ulcorner C^*\urcorner);$$

i.e.,

$$PA \vdash \text{“}u \nVdash \Box C\text{”} \wedge u > \bar{0} \to (L = u \to \sim(\Box C)^*).$$

The remaining cases $B = \forall vCv$ and $B = \exists vCv$ are complementary and we need only show:

$$PA \vdash \text{“}u \Vdash \forall vCv\text{”} \wedge u > \bar{0} \rightarrow (L = u \rightarrow \forall vC^*v)$$
$$PA \vdash \text{“}u \Vdash \exists vCv\text{”} \wedge u > \bar{0} \rightarrow (L = u \rightarrow \exists vC^*v).$$

For $Q \in \{\forall, \exists\}$, observe

$$\begin{aligned} PA \vdash \text{“}u \Vdash QvCv\text{”} \wedge u > \bar{0} &\rightarrow Qv(\text{“}u \Vdash Cv\text{”} \wedge u > \bar{0}) \\ &\rightarrow Qv(L = u \rightarrow C^*v) \\ &\rightarrow (L = u \rightarrow QvC^*v). \end{aligned}$$

We are now in position to prove the Theorem: Let D be in the language of QGL. For some $x > 0$,

$$x \Vdash D \leftrightarrow A(D).$$

Thus “$\bar{x} \Vdash D \leftrightarrow A(D)$” is true and, since it is quantifier-free,

$$PA \vdash \text{“}\bar{x} \Vdash D \leftrightarrow A(D)\text{”}.$$

By Lemma 5,

$$PA \vdash L = \bar{x} \rightarrow \sim(D \leftrightarrow A(D))^*.$$

By Lemma 4(ii), $PA + L = \bar{x}$ is consistent. Thus, $PA + \sim(D \leftrightarrow A(D))^*$ is consistent; i.e.,

$$PA \nVdash D^* \leftrightarrow A(D)^*.$$

REFERENCES

[1] Bernardi, Claudio, “The uniqueness of the fixed point theorem in every diagonalisable algebra,” *Studia Logica*, vol. 35 (1976), pp. 335–343.

[2] Carlson, Timothy, “Modal logics with several operators and provability interpretations,” *Israel Journal of Mathematics*, vol. 54 (1986), pp. 14–24.

[3] Guaspari, David, and Robert M. Solovay, “Rosser sentences,” *Annals of Mathematical Logic*, vol. 16 (1979), pp. 81–99.

[4] Montagna, Franco, “The predicate modal logic of provability,” *Notre Dame Journal of Formal Logic*, vol. 25 (1984), pp. 179–189.

[5] Sambin, Giovanni, “Un estensione del theorema di Löb,” Rendiconti del Seminario Mathematico dell’ Universita Padova, vol. 52 (1974), pp. 193–199.

[6] Smoryński, C., “Calculating self-referential statements, I; explicit calculations,” *Studia Logica*, vol. 38 (1979), pp. 17–36.

[7] Smoryński, C., *Self-Reference and Modal Logic*, Springer-Verlag, New York, 1986.

[8] Solovay, Robert M., “Provability interpretations of modal logic,” *Israel Journal of Mathematics*, vol. 25 (1976), pp. 287–304.

Department of Mathematics and Computer Science
San José State University
San José, California 95192