Notre Dame Journal of Formal Logic
Volume 23, Number 4, October 1982

# Commutativity and Self-Reference

C. SMORYŃSKI

A common occurrence in fixed-point theory is the sharing of some fixed point by commuting operators[1]: Quite often, if $F$, $G$ are such that, for all $x$, $FG(x) = GF(x)$, there is some element $x$ such that $x = F(x) = G(x)$. How common is this occurrence in the arithmetical fixed-point theory, i.e., in the theory of arithmetical self-reference? The ongoing modal analysis of this self-reference offers a tool to study this question. Put differently, this question offers a test for the coherence of the modal theory.

In the present paper I will offer some simple observations in this direction. Though the results are not difficult, I think they do testify to the coherence of the modal theory. Moreover, through the exhibition of nontrivial differences between varying types of self-reference, they offer some indication of the possibility of a rich theory of arithmetical self-reference.

In its weakest and most popular form, the Diagonalisation Theorem asserts there to be, for any formula $\psi v$ with only the variable $v$ free, a sentence $\phi$ such that

$$PA \vdash \phi \longleftrightarrow \psi(\ulcorner\phi\urcorner).$$

Well, $\phi$ is not a fixed point in the strictest sense: On the left it is a sentence and on the right a numerical code for the sentence. Composing $\psi$ with the code-assigning function will give an operator on sentences,

$$\Psi(\phi) = \psi(\ulcorner\phi\urcorner),$$

for which we still do not have a fixed-point equation,

$$\Psi(\phi) = \phi, \qquad (*)$$

but merely an equivalence. Now, if $\Psi$ preserves provable equivalence, we can identify equivalent sentences and so obtain a genuine fixed-point equation (*). I shall call such operators *extensional* and shall deal solely with extensional

*Received March 30, 1981*

operators and their fixed points here—though I shall stick to linguistic rather than functorial notation.

Is the notion of extensionality anything more than a red herring? I really do not know. It would seem to be relevant on the superficial ground that intuition from fixed point theory should be more reliable when one is dealing with genuine fixed points than when one is not. At one point below I will use extensionality; but, since the particular application will depend on somewhat more than mere extensionality, no inference about the value of extensionality in itself can be drawn.

Let us get down to business:

Example: Let $\psi v$ be any formula with only $v$ free such that, for all sentences $\phi$,

$$PA \vdash \psi(\ulcorner \neg \phi \urcorner) \longleftrightarrow \neg \psi(\ulcorner \phi \urcorner).$$

Then $\psi$ and $\neg \psi$ commute, but share no fixed point.

The proof is immediate. Less immediate are good instances of this example. I cite two:

i. Let $\psi v$ assert $v$ to be the code of a formula containing an odd number (or: an even number) of symbols;
ii. Let $\psi v$ define truth for a model of $PA$.

The first of these is simple, but not very exciting. The second is much more interesting, but requires a bit of effort to construct: One needs the Orey Compactness Theorem (which can be found in [2]; but cf. [3] and [5] for related formulae and their fixed points). I note that formula (ii) is extensional and seems to be more interesting for this reason: $\psi(\ulcorner \phi \urcorner)$ depends on what $\phi$ says rather than on how $\phi$ says it.

Just as it is not the case that every function has a fixed point or that every pair of commuting functions possessing fixed points will share a fixed point, it is not, as we have just seen, the case that commuting formulae will share a fixed point. Fixed-point theory depends on special properties of the functions (e.g., continuity) or of the underlying space on which the functions are defined (e.g., well-orderedness). By analogy, the theory of self-reference will have to invoke special properties to establish particular results. In the present context, the best understood relevant special property is the modal one—that $\psi v$ be modal in character.

There are now several systems of modal logic dedicated to the study of various types of self-reference. The first, simplest, most successful, and most widely known of these is the system $L$, the language of which is given by:

Propositional variables: $p, q, r, \ldots$
Truth values: $\top, \bot$
Propositional connectives: $\neg, \wedge, \vee, \rightarrow$
Modal operator: $\Box$.

Arithmetical interpretations $*$ of the modal language are determined by first assigning arithmetical sentences $\phi_p$ to variables $p$ and then defining:

i. $p^*$: $\phi_p$
ii-iii. $\top^*$: $0 = 0$; $\bot^*$: $0 = 1$

iv-vi. $(A \circ B)^*$: $A^* \circ B^*$, for $\circ \in \{\wedge, \vee, \rightarrow\}$
vii. $(\neg A)^*$: $\neg A^*$
viii. $(\Box A)^*$: $Pr_{PA}(\ulcorner A^* \urcorner)$,

where $Pr_{PA}(v)$ is the canonical proof predicate for $PA$. If $A(p_1, \ldots, p_n)$ has only the propositional variables $p_i$ and $\phi_i = p_i^*$, we will write $A(\phi_1, \ldots, \phi_n)$ for $A(p_1, \ldots, p_n)^*$.

The most popular result about $L$ and its interpretations is Solovay's Completeness Theorem ([8]):

**Solovay's Completeness Theorem** *For any modal sentence $A$, $L \vdash A$ iff $\forall^*(PA \vdash A^*)$.*

Solovay's Completeness Theorem is very interesting, but not in itself very useful. What is useful is a technical lemma upon which this result and a generalization rested. But that lies ahead of us.

The second most popular result is the de Jongh-Sambin Fixed Point Theorem (cf. [4] and [7]). Notice that, if each occurrence of $p$ in $A(p)$ lies within the scope of a $\Box$, the corresponding occurrences of $\phi$ in $A(\phi)$ will not be occurrences of $\phi$ as a subformula of $A(\phi)$, but merely references to $\phi$ via $\ulcorner \phi \urcorner$. Thus, if each occurrence of $p$ is so restricted, there will be a sentence $\phi$ such that $PA \vdash \phi \leftrightarrow A(\phi)$. Because of this, the restriction on $p$ that each of its occurrences in $A(p)$ lie within the scope of a $\Box$ is called the *diagonalisation restriction* (DR).

**De Jongh-Sambin Fixed Point Theorem** *Let $p$ obey the DR in $A(p)$. There is a sentence $B$ constructed from the variables of $A(p)$ other than $p$ and such that*

i. $L \vdash [p \leftrightarrow A(p)] \wedge \Box[p \leftrightarrow A(p)] \rightarrow .p \leftrightarrow B$
ii. $L \vdash B \leftrightarrow A(B)$.

As one of the two who originally conjectured this result, I rather like it. I must nevertheless admit that, as was the case with Solovay's Completeness Theorem, the result is more interesting than useful. What is generally useful is the weaker uniqueness result of Bernardi and de Jongh (cf. [1] and [6]):

**Uniqueness Theorem** *Let $p$ obey the DR in $A(p)$. Then $L \vdash \Box[p \leftrightarrow A(p)] \wedge \Box[q \leftrightarrow A(q)] \rightarrow \Box(p \leftrightarrow q)$.*

The arithmetical interpretation of this is: If $p$ obeys the DR in $A(p)$, then, for any interpretation of the variables of $A(p)$ other than $p$, if

$$PA \vdash \phi_0 \leftrightarrow A(\phi_0) \text{ and } PA \vdash \phi_1 \leftrightarrow A(\phi_1),$$

then

$$PA \vdash \phi_0 \leftrightarrow \phi_1.$$

In other words, if $\psi v$ arises from a modal formula $A(p)$ in which $p$ obeys the DR, then $\psi v$ has (up to provable equivalence) a unique fixed point. The significance of this for fixed points of commuting operators is immediate:

**First Common Fixed Point Theorem** *Suppose $\psi_1 v$, $\psi_2 v$ arise from modal formulae $A(p)$ and $B(p)$ in which p obeys the DR. If, for all sentences $\phi$,*

$$PA \vdash \psi_1(\ulcorner\psi_2(\ulcorner\phi\urcorner)\urcorner) \longleftrightarrow \psi_2(\ulcorner\psi_1(\ulcorner\phi\urcorner)\urcorner),$$

*there is a single sentence $\phi$ such that*

$$PA \vdash \phi \longleftrightarrow \psi_1(\ulcorner\phi\urcorner) \text{ and } PA \vdash \phi \longleftrightarrow \psi_2(\ulcorner\phi\urcorner).$$

*Proof:* Diagonalise on $\psi_1$ to produce a sentence $\phi$ such that $PA \vdash \phi \longleftrightarrow \psi_1(\ulcorner\phi\urcorner)$. By extensionality, $PA \vdash \psi_2(\ulcorner\phi\urcorner) \longleftrightarrow \psi_2(\ulcorner\psi_1(\ulcorner\phi\urcorner)\urcorner)$. By commutativity, $PA \vdash \psi_2(\ulcorner\phi\urcorner) \longleftrightarrow \psi_1(\ulcorner\psi_2(\ulcorner\phi\urcorner)\urcorner)$ and $\psi_2(\ulcorner\phi\urcorner)$ is a fixed point of $\psi_1$. By the uniqueness of the fixed point of $\psi_1$, this yields $PA \vdash \phi \longleftrightarrow \psi_2(\ulcorner\phi\urcorner)$. QED

Remark: The proof of the First Common Fixed Point Theorem establishes a bit more than stated: If $\psi_1$ arises from $A(p)$ with $p$ obeying the DR, and $\psi_2$ is extensional, then the commutativity of $\psi_1$ and $\psi_2$ entails the existence of a common fixed point.

The remark is a bit nicer than the theorem, although both results are slightly too trivial to be very pleasing. Another disappointing feature of the theorem is that such instances of commutativity are not all that common. In ordinary algebra, for example, the formulae $A(p) = p \wedge C$ and $B(p) = p \wedge D$ (with $p$ not occurring in $C$ or $D$) obviously commute: $L \vdash AB(p) \longleftrightarrow BA(p)$. The diagonalisation restriction largely kills this:

Example: Let $A(p) = \Box p \wedge C$, $B(p) = \Box p \wedge D$, with $p$ absent from $C, D$. Then the following are equivalent:

i. $L \vdash AB(p) \longleftrightarrow BA(p)$
ii. $L \vdash C \wedge \Box C \longleftrightarrow D \wedge \Box D$
iii. $L \vdash AA(p) \longleftrightarrow BB(p)$.

Moreover, under any of these assumptions, $L \vdash A^n(p) \longleftrightarrow B^n(p)$ for all $n > 1$.

*Proof:* (Those unfamiliar with $L$ should first consult, e.g., [8].) Note that

$$L \vdash AB(p) \longleftrightarrow . \Box^2 p \wedge \Box D \wedge C$$
$$L \vdash BA(p) \longleftrightarrow . \Box^2 p \wedge \Box C \wedge D.$$

i ⇒ ii. Letting $p = \top$,

$$\begin{aligned} L \vdash AB(\top) \longleftrightarrow BA(\top) &\Rightarrow L \vdash \Box D \wedge C \longleftrightarrow \Box C \wedge D \\ &\Rightarrow L \vdash C \wedge \Box C \rightarrow . \Box D \rightarrow D \\ &\Rightarrow L \vdash \Box C \rightarrow \Box(\Box D \rightarrow D) \\ &\Rightarrow L \vdash \Box C \rightarrow \Box D \\ &\Rightarrow L \vdash C \wedge \Box C \rightarrow \Box D \wedge D. \end{aligned}$$

The converse implication is similar.

ii ⇒ iii. Note that

$$L \vdash A^2(p) \longleftrightarrow . \Box^2 p \wedge \Box C \wedge C$$
$$L \vdash B^2(p) \longleftrightarrow . \Box^2 p \wedge \Box D \wedge D.$$

The conclusion is immediate.

iii ⇒ i. Letting $p = \top$,

$$\begin{aligned} L \vdash A^2(\top) \longleftrightarrow B^2(\top) &\Rightarrow L \vdash C \wedge \Box C \longleftrightarrow D \wedge \Box D \\ &\Rightarrow L \vdash D \wedge C \wedge \Box C \longleftrightarrow D \wedge \Box D \\ &\Rightarrow L \vdash \Box D \wedge \Box C \longleftrightarrow \Box D \\ &\Rightarrow L \vdash \Box D \rightarrow \Box C \\ &\Rightarrow L \vdash C \wedge \Box D \rightarrow C \wedge \Box C \\ &\Rightarrow L \vdash C \wedge \Box D \rightarrow D \wedge \Box D \\ &\Rightarrow L \vdash C \wedge \Box D \rightarrow D \wedge \Box C. \end{aligned}$$

The converse implication being similarly established, we have

$$\begin{aligned} L \vdash A^2(p) \longleftrightarrow B^2(p) &\Rightarrow L \vdash \Box D \wedge C \longleftrightarrow D \wedge \Box C \\ &\Rightarrow L \vdash \Box^2 p \wedge \Box D \wedge C \longleftrightarrow \Box^2 p \wedge \Box C \wedge D \\ &\Rightarrow L \vdash AB(p) \longleftrightarrow BA(p). \end{aligned}$$

The final assertion is trivial. QED

That $A(p) \longleftrightarrow B(p)$ does not follow from commutativity in this case is shown by the following counterexample:

Counterexample: Let $A(p) = \Box p \wedge q$, $B(p) = \Box p \wedge (\Box q \rightarrow q)$. Then:

i. $L \vdash AB(p) \longleftrightarrow BA(p)$
ii. $L \nvdash A(p) \longleftrightarrow B(p)$.

I leave the proof to the reader.

This example and its accompanying counterexample illustrate the potential for interesting, or at least amusing, results. From the panoply of possibilities, permit me to propose a pair of open problems:

Problem 1: Find a number of nontrivial pairs $A(p)$ and $B(p)$ in which $p$ obeys the DR such that $L \vdash AB(p) \longleftrightarrow BA(p)$.

Problem 2: Does there exist a formula $A(p)$ in which $p$ obeys the DR such that, for all $B(p)$ in which $p$ obeys the DR, it follows from $L \vdash AB(p) \longleftrightarrow BA(p)$ that, for some $n \geqslant 1$, either $L \vdash B(p) \longleftrightarrow A^n(p)$ or $L \vdash A(p) \longleftrightarrow B^n(p)$?

Should these prove too easy, I also note there is the general question of providing a simple computable criterion for the commutativity of such formulas $A(p)$ and $B(p)$. Since I have settled for citing such simple-seeming questions, it should come as no surprise to the reader that I have, at this point in time, nothing more to report on the commutativity of such operations. Where do I intend to go from here?

Let us back up a bit. Recall the formulae

$$A(p) = p \wedge C, \qquad B(p) = p \wedge D,$$

whose commutativity is evident (provided, of course, $p$ is absent from $C$ and $D$). $A(p)$ has the obvious fixed point $C$ and $B(p)$ the fixed point $D$. Moreover, they further share the fixed point $C \wedge D$. Surely the modal theory should be able to account for this, i.e., this sharing of a fixed point should be explainable as being a more-or-less trivial instance of some general result. This is my next goal.

First, I should isolate the “cause” for the existence of fixed points to formulae like $A(p)$ and $B(p)$ as just described. To do this, I must first define what "formulae like $A(p)$ and $B(p)$" means:

**Definition** The notion *$p$ is weakly positive in $A(p)$* is inductively defined by:

i. $p$ is weakly positive in $\top$, $\bot$ and any variable $q$
ii. $p$ is weakly positive in $\Box B$ for any formula $B$ (whether or not $p$ occurs in $B$)
iii. if $p$ is weakly positive in $B$ and $C$, then $p$ is weakly positive in $B \wedge C$ and $B \vee C$.

The adverb "weakly" is explained by clause ii: $p$ can do all sorts of negative things, provided it limits such behaviour to occurring within the scopes of boxes.

The self-referential interest in formulae in which $p$ is weakly positive stems from their possession of fixed points: Write $A(p) = A(p, q_1, \ldots, q_n)$, with all variables exhibited (this notation will be freely invoked whenever convenient), and suppose $p$ is weakly positive in $A$. Then, for any choice $\theta_1, \ldots, \theta_n$ of arithmetical sentences, there is another arithmetical sentence $\phi$ such that

$$PA \vdash \phi \longleftrightarrow A(\phi, \theta_1, \ldots, \theta_n).$$

To see this, simply write $A(p)$ in disjunctive normal form (maximal subformulas of the form $\Box C$ being considered strange new variables):

$$L \vdash A(p) \longleftrightarrow . \; B_1(p) \vee (B_2(p) \wedge p)$$

where $p$ obeys the DR in $B_1(p)$ and $B_2(p)$. Note that the quantifier complexities of the formulae $B_1(\phi, \theta_1, \ldots, \theta_n)$ and $B_2(\phi, \theta_1, \ldots, \theta_n)$ are governed by the complexities of the formulae $\theta_1, \ldots, \theta_n$ and not by that of $\phi$, which is only referred to via $\ulcorner\phi\urcorner$ in these sentences. Thus, if this complexity is $\Sigma_{n+1}$ and if $\phi \in \Sigma_{n+1}$, $A(\phi)$ will also be $\Sigma_{n+1}$, and we can temporarily modify $A$ to solve the equivalence

$$PA \vdash \phi \longleftrightarrow . \; B_1(\phi) \vee [B_2(\phi) \wedge Tr_{\Sigma_{n+1}}(\ulcorner\phi\urcorner)],$$

where $Tr_{\Sigma_{n+1}}(v)$ is a $\Sigma_{n+1}$ truth definition for $\Sigma_{n+1}$ formulae. Since $\phi \in \Sigma_{n+1}$, we can erase $Tr_{\Sigma_{n+1}}(\cdot)$ from this last displayed formula and conclude

$$PA \vdash \phi \longleftrightarrow A(\phi).$$

Let me digress briefly to give a quick application:

Example: Let $A(p) = p \wedge q$, $B(p) = \Box p \wedge (\Box q \longleftrightarrow q)$. Then:

i. $L \vdash AB(p) \longleftrightarrow BA(p)$
ii. For no $n, m \geqslant 1$ does $L \vdash A^n(p) \longleftrightarrow B^m(p)$
iii. For any $\theta$, there is a sentence $\phi$ such that

$$PA \vdash \phi \longleftrightarrow A(\phi,\theta) \text{ and } PA \vdash \phi \longleftrightarrow B(\phi,\theta).$$

*Proof:* i. Note

$$L \vdash AB(p) \longleftrightarrow [\Box p \wedge (\Box q \longleftrightarrow q)] \wedge q$$
$$L \vdash BA(p) \longleftrightarrow \Box p \wedge \Box q \wedge (\Box q \longleftrightarrow q).$$

ii. For $m, n \geqslant 1$,

$$\begin{aligned} L &\vdash A^n(p) \longleftrightarrow p \wedge q \\ L &\vdash B^m(p) \longleftrightarrow \Box^m p \wedge \Box[\Box q \longleftrightarrow q] \wedge (\Box q \longleftrightarrow q) \\ &\qquad \longleftrightarrow \Box^m p \wedge \Box q \wedge (\Box q \longleftrightarrow q). \end{aligned}$$

Letting $q = \top$, $p = \bot$ quickly reveals

$$L \nvdash A^n(p) \longleftrightarrow B^m(p).$$

iii. Since $p$ obeys the DR in $B(p)$ and $A(p)$ gives rise to an extensional operator, our earlier Remark applies. QED

With this example, I have finally given a nontrivial pair of commuting formulae sharing a fixed point. Of course, there is also the pair $A(p) = p \wedge C$, $B(p) = p \wedge D$, with $p$ absent from $C$, $D$. How does one explain examples such as this modally? The answer, of course, is: with Kripke models.

**Definition** A *Kripke model* is a triple $\underline{K} = (K, <, \Vdash)$, where $(K, <)$ is a finite partially ordered set with least element $\alpha_0$ and $\Vdash$ is a "forcing relation" or system of truth valuations indexed by $(K, <)$ and satisfying: For all $\alpha \in K$,

ii-iii. $\alpha \Vdash \top$, $\alpha \nVdash \bot$
iv-vi. $\alpha \Vdash A \circ B$ iff $(\alpha \Vdash A) \circ (\alpha \Vdash B)$, for $\circ \in \{\wedge, \vee, \rightarrow\}$
vii. $\alpha \Vdash \neg A$ iff $\alpha \nVdash A$
viii. $\alpha \Vdash \Box A$ iff $\forall \beta > \alpha(\beta \Vdash A)$.

The relevance of Kripke models to $L$ is quickly explained by the following oft-proven theorem:

**Completeness Theorem** *For any modal formula $A$, $L \vdash A$ iff $A$ is valid in all Kripke models, i.e. for all $\underline{K}$ and all $\alpha \in K$, $\alpha \Vdash A$.*

The relevance of Kripke models to the problem at hand is also quickly explained:

**Second Common Fixed Point Theorem** *Suppose $p$ is weakly positive in both $A(p)$ and $B(p)$, and that $L \vdash AB(p) \longleftrightarrow BA(p)$. Then: For any Kripke model $K = (K, <, \Vdash)$, the forcing relation $\Vdash$ can be extended to encompass a new variable $p_0$ in such a way as to make valid both $p_0 \longleftrightarrow A(p_0)$ and $p_0 \longleftrightarrow B(p_0)$. In other words, in any Kripke model, $A(p)$ and $B(p)$ share a fixed point.*

*Proof:* We mimic de Jongh's proof (cf. [6]) that fixed points of formulae in which $p$ obeys the DR are implicitly defined in all Kripke models; i.e., we show by induction from the top down how to extend $\Vdash$ to include $p_0$ in such a way as to always guarantee $\alpha \Vdash p_0 \longleftrightarrow A(p_0)$ and $\alpha \Vdash p_0 \longleftrightarrow B(p_0)$.

For notational convenience, we let $\Vdash_0$ be the extension of $\Vdash$ which we are constructing.

Consider a node $\alpha \in K$. Our basic hypothesis is that $\Vdash$ is defined on all of $(K, <)$; and our induction hypothesis is that $\Vdash_0$ has been defined for all $\beta > \alpha$ in such a way that, for $\beta > \alpha$, $\beta \Vdash_0 p_0 \longleftrightarrow A(p_0)$ and $\beta \Vdash_0 p_0 \longleftrightarrow B(p_0)$. Write $A(p_0)$ and $B(p_0)$ in disjunctive normal form:

$$\begin{aligned} L &\vdash A(p_0) \longleftrightarrow C_1(p_0) \vee [C_2(p_0) \wedge p_0] \\ L &\vdash B(p_0) \longleftrightarrow D_1(p_0) \vee [D_2(p_0) \wedge p_0], \end{aligned}$$

where $p_0$ obeys the DR in each $C_i$ and each $D_h$. Because of the DR, even though the truth value of $p_0$ at $\alpha$ has not yet been determined, the truth values of each $C_i$ and each $D_h$ at $\alpha$ have been determined.

*Case 1.* $\alpha \Vdash_0 A(p_0)$ regardless of how we decide whether or not $\alpha \Vdash_0 p_0$. Let $\alpha \Vdash_0 p_0$, so that $\alpha \Vdash_0 p_0 \longleftrightarrow A(p_0)$. But then $\alpha \Vdash_0 [p_0 \longleftrightarrow A(p_0)] \wedge \Box[p_0 \longleftrightarrow A(p_0)]$ and, since

$$L \vdash [E \longleftrightarrow F] \wedge \Box[E \longleftrightarrow F] \rightarrow\!\!. \ G(E) \longleftrightarrow G(F),$$

we see that

$$\alpha \Vdash_0 B(p_0) \longleftrightarrow B[A(p_0)].$$

By the commutativity of $A$, $B$,

$$\begin{aligned} \alpha \Vdash_0 B(p_0) &\longleftrightarrow A[B(p_0)] \\ \Vdash_0 B(p_0) &\longleftrightarrow \top, \end{aligned}$$

this last equivalence following from the case assumption that $\alpha \Vdash_0 A(p_0)$ regardless of our decision on $\alpha \Vdash_0 p_0$. Since we chose $\alpha \Vdash_0 p_0$, it follows that $\alpha \Vdash_0 B(p_0) \longleftrightarrow p_0$.

*Case 2.* $\alpha \nVdash_0 A(p_0)$ regardless of how we decide whether or not $\alpha \Vdash_0 p_0$. It follows similarly that, if we define $\alpha \nVdash_0 p_0$, we will have $\alpha \Vdash_0 p_0 \longleftrightarrow A(p_0)$ and $\alpha \Vdash_0 p_0 \longleftrightarrow B(p_0)$.

*Case 3.* The truth value of $A(p_0)$ is undetermined. Without loss of generality, we can assume the same to hold of $B(p_0)$. Replacing each $C_i$, and each $D_h$ by their truth values, we see that $A(p_0)$ and $B(p_0)$ become positive in $p_0$. Thus, indeterminacy requires

$$\begin{aligned} \alpha \nVdash_0 p_0 &\Rightarrow \alpha \nVdash_0 A(p_0), B(p_0) \\ \alpha \Vdash_0 p_0 &\Rightarrow \alpha \Vdash_0 A(p_0), B(p_0). \end{aligned}$$

Hence either decision about $\alpha \Vdash_0 p_0$ will yield

$$\alpha \Vdash_0 p_0 \longleftrightarrow A(p_0) \text{ and } \alpha \Vdash_0 p_0 \longleftrightarrow B(p_0). \qquad \text{QED}$$

As the observant reader might have noticed, we have strayed from our original path. The First Common Fixed Point Theorem dealt with self-referential sentences; while the Second Common Fixed Point Theorem concerns itself only with modal formulae. Does the Second Common Fixed Point Theorem have anything to say about arithmetical self-reference, and, if so, what? This question is where the depth lies and my shallow answer will be incomplete and rather unsatisfying.

**Corollary 1** *Let $A(p, p_1, \ldots, p_n)$ and $B(p, p_1, \ldots, p_n)$ have only the variables shown, with p weakly positive in both formulae. Suppose for all arithmetical sentences $\phi, \theta_1, \ldots, \theta_n$,*

$$PA \vdash A(B(\phi, \theta_1, \ldots, \theta_n), \theta_1, \ldots, \theta_n) \longleftrightarrow B(A(\phi, \theta_1, \ldots, \theta_n), \theta_1, \ldots, \theta_n).$$

*Then there are arithmetical sentences $\phi, \theta_1, \ldots, \theta_n$ such that*

$$PA \vdash \phi \longleftrightarrow A(\phi, \theta_1, \ldots, \theta_n) \text{ and } PA \vdash \phi \longleftrightarrow B(\phi, \theta_1, \ldots, \theta_n).$$

*Proof:* By Solovay's Completeness Theorem, from the commutativity of all instances of $A(\phi, \theta_1, \ldots, \theta_n)$ and $B(\phi, \theta_1, \ldots, \theta_n)$, it follows that

$$L \vdash AB(p) \longleftrightarrow BA(p).$$

By the Second Common Fixed Point Theorem, for any Kripke model $\underline{K} = (K, <, \Vdash)$ we can assume some $p$ to be forced in such a way as to validate the equivalences $p \longleftrightarrow A(p)$ and $p \longleftrightarrow B(p)$. Now, we wish to reinterpret this arithmetically.

Solovay has shown the following: Let $\underline{K} = (K, <, \Vdash)$ with minimum node $\alpha_0$ and suppose

i. $\alpha_0 \Vdash E$
ii. $\alpha_0 \Vdash \Box F \rightarrow F$, for all $\Box F$ occurring as subformulae of $E$. Then

$E^*$ is true in the standard model for some interpretation $*$.

We shall apply Solovay's result to

$$E = \Box[p \longleftrightarrow A(p)] \wedge \Box[p \longleftrightarrow B(p)].$$

Let $\underline{K} = (K, <, \Vdash)$, with minimum node $\alpha_0$, be any model in which $p \longleftrightarrow A(p)$ and $p \longleftrightarrow B(p)$ are valid. Define a sequence of models $\underline{K}_0 = \underline{K}, \underline{K}_1, \underline{K}_2, \ldots$ as follows:

$K_n = K \cup \{\alpha_1, \ldots, \alpha_n\}$
$\alpha_0 > \alpha_1 > \ldots > \alpha_n$
$\alpha_i \Vdash q$ iff $\alpha_0 \Vdash q,$

for all propositional variables $q$ other than $p$. Define $\alpha_i \Vdash p$ in accordance with the procedure of the proof of the Second Common Fixed Point Theorem.

Note that, for any subformula $\Box F$ of $E$, if $\alpha_n \nVdash \Box F$, then $\alpha_{n+1} \nVdash \Box F$– hence the truth of $\Box F$ is constant from some $\alpha_n$ on. Choose $n$ large enough for these values to be constant and consider $\underline{K}_n$ (and also $\underline{K}_{n+1}$). Now, for any subformula $\Box F$ of $E$,

$$\begin{aligned} \alpha_n \Vdash \Box F &\Rightarrow \alpha_{n+1} \Vdash \Box F \\ &\Rightarrow \alpha_n \Vdash F, \end{aligned}$$

whence

$$\alpha_n \Vdash \Box F \rightarrow F;$$

i.e., Solovay's conditions hold and we can apply his lemma.

By Solovay's lemma, there is some interpretation $p^* = \phi$, $p_i^* = \theta_i$ making $E^*$ true, i.e.

$$\mathbf{N} \models Pr_{PA}(\ulcorner \phi \longleftrightarrow A(\phi, \theta_1, \ldots, \theta_n)\urcorner) \wedge Pr_{PA}(\ulcorner \phi \longleftrightarrow B(\phi, \theta_1, \ldots, \theta_n)\urcorner),$$

i.e.,

$$PA \vdash \phi \longleftrightarrow A(\phi, \theta_1, \ldots, \theta_n) \text{ and } PA \vdash \phi \longleftrightarrow B(\phi, \theta_1, \ldots, \theta_n). \qquad \text{QED}$$

The disappointing thing about Corollary 1 is that the parametric hypothesis is too strong and the parametric conclusion too weak. When there are no parameters, the relative strengths of hypothesis and conclusion match and we get the pleasing result:

**Corollary 2** *Suppose $p$ is weakly positive in $A(p)$ and $B(p)$ and that $p$ is the only variable occurring in these formulae. If, for all sentences $\phi$,*

$$PA \vdash AB(\phi) \longleftrightarrow BA(\phi),$$

*there is a sentence $\phi$ such that*

$$PA \vdash \phi \longleftrightarrow A(\phi) \textit{ and } PA \vdash \phi \longleftrightarrow B(\phi).$$

What about the case with parameters?

Conjecture: Let $A(p, p_1, \ldots, p_n)$ and $B(p, p_1, \ldots, p_n)$ have only the variables shown, with $p$ weakly positive in both formulae. Let $\theta_1, \ldots, \theta_n$ be arithmetic sentences such that, for all sentences $\phi$,

$$PA \vdash A(B(\phi, \theta_1, \ldots, \theta_n), \theta_1, \ldots, \theta_n) \longleftrightarrow B(A(\phi, \theta_1, \ldots, \theta_n), \theta_1, \ldots, \theta_n).$$

Then there is a sentence $\phi$ such that

$$PA \vdash \phi \longleftrightarrow A(\phi, \theta_1, \ldots, \theta_n) \text{ and } PA \vdash \phi \longleftrightarrow B(\phi, \theta_1, \ldots, \theta_n).$$

Problem 3: Establish the Conjecture by modal considerations similar to those used in proving Corollary 1.

## NOTE

[1] This fact was brought to my attention during a lecture by Dana Scott, who was bemoaning the failure of models of the λ-calculus to account for this aspect of Kleene's recursion theorems.

## REFERENCES

[1] Bernardi, C., "The fixed-point theorem for diagonalizable algebras," *Studia Logica,* vol. 34 (1975), pp. 239-251.

[2] Feferman, S., "Arithmetization of metamathematics in a general setting," *Fundamenta Mathematicae*, vol. 49 (1960), pp. 35-92.

[3] Manevitz, L. and J. Stavi, "$\Delta_2^0$ operators and alternating sentences in arithmetic," *The Journal of Symbolic Logic,* vol. 45 (1980), pp. 144-154.

[4] Sambin, G., "An effective fixed-point theorem in intuitionistic diagonalizable algebras," *Studia Logica,* vol. 35 (1976), pp. 345-361.

[5] Smoryński, C., "The incompleteness theorems," in *Handbook of Mathematical Logic,* ed., J. Barwise, North-Holland, Amsterdam, 1977.

[6] Smoryński, C., "Beth's theorem and self-referential sentences," in *Logic Colloquium '77,* ed., A. Macintyre, L. Pacholski, and J. Paris, North-Holland, Amsterdam.

[7] Smoryński, C., "Calculating self-referential statements I: explicit calculations," *Studia Logica,* vol. 38 (1979), pp. 17-36.

[8] Solovay, R., "Provability interpretations of modal logic," *Israel Journal of Mathematics,* vol. 25 (1976), pp. 287-304.

*Department of Mathematics*
*Ohio State University*