INTERPRETABILITY IN TERMS OF MODELS [1])

BY

RICHARD MONTAGUE

(Communicated by Prof. HEYTING at the meeting of February 27, 1964)

Between two sets of sentences, Φ and Ψ, several relations of relative strength may hold, for instance, those expressed as follows:

(1) all members of Ψ are derivable from Φ (by the axioms and rules of first-order logic with identity);
(2) the theory axiomatized by Ψ is interpretable in the theory axiomatized by Φ;
(3) the theory axiomatized by Ψ is relatively interpretable in the theory axiomatized by Φ.

The well-known, so-called *syntactical*, definitions of these relations have an accidental character,[2]) and it seems desirable to find alternative characterizations with greater philosophic interest. In particular, we might seek interesting relations between the *models* of Φ and Ψ which correspond to (1), (2), (3).

For the relation (1) the problem was solved by Gödel's completeness theorem. The present paper gives analogous equivalences involving the relations (2) and (3), at least for those cases in which the set Ψ is finite. Syntactical formulations of the results are also obtained: a finitely axiomatizable theory is interpretable (or relatively interpretable) in a theory Φ if and only if it is interpretable (or relatively interpretable) in every complete extension of Φ.

§ 1. *Preliminaries.*

Concerning the framework within which (first-order) metamathematics is to be conducted, the literature is in some details not uniform and in

---

[1]) This paper, which reports a talk given before the Euratom workgroup in Amsterdam in the fall of 1962, was prepared partly at the University of Amsterdam under Euratom Contract No. 010–60–12 and partly under United States National Science Foundation Grant No. NSF–GP 1603 (Montague). I am grateful also to Dr. K. L. de Bouvère and Professors Andrzej Ehrenfeucht, Solomon Feferman, and Alfred Tarski for helpful discussion; Professor Tarski made suggestions leading to the improvement of several formulations.

[2]) The relation expressed by (1) is the basic notion of first-order logic. The relations expressed by (2) and (3) were first explicitly defined in TARSKI, MOSTOWSKI, and ROBINSON [1].

others not entirely definite. It is therefore perhaps not wholly gratuitous to sketch here a suitable framework.[3])

Metamathematics is seen as an extension of set theory. As to what is meant by 'set theory' we may maintain a certain degree of indeterminacy. All the objects to which the present paper will refer are to be *sets*; we leave open the question whether there also exist *individuals* and *proper classes*. For set-theoretic axioms we may thus choose either the system of Zermelo-Fraenkel or that of Bernays-Morse, and in either case we may either allow or not for the existence of individuals.[4])

The extension consists in adding to set theory as primitive symbols the 0-place operation symbols '$\neg$', '$\rightarrow$', '$\wedge$', '$\vee$', '$\leftrightarrow$', '[', ']', '$\bigwedge$', '$\bigvee$', '$=$' (respectively read 'the negation symbol', 'the implication symbol', 'the conjunction symbol', 'the disjunction symbol', 'the biconditional symbol', 'the left bracket', 'the right bracket', 'the universal quantifier', 'the existential quantifier', and 'the identity symbol'), the 1-place operation symbol 'v' (read, in the context 'v_n', 'the n^{th} variable'), and the 2-place operation symbols 'p' and 'f' (read, in the contexts '$p_{\alpha,n}$' and '$f_{\alpha,n}$', 'the α^{th} n-place predicate' and 'the α^{th} n-place operation symbol (or functor)' respectively), and adding the following axiom: if α, β are distinct ordinals and n, k are distinct natural numbers, then the set $\{\neg, \rightarrow, \wedge, \vee, \leftrightarrow, [,], \bigwedge, \bigvee, =, v_n, v_k, p_{\alpha,n}, p_{\beta,n}, p_{\alpha,k}, f_{\alpha,n}, f_{\beta,n}, f_{\alpha,k}\}$ contains exactly 18 members, each of which is a 1-place sequence.

Such are the axiomatic foundations of our metatheory. We now introduce by definition a number of metamathematical notions. The *logical constants* are $\neg$, $\rightarrow$, $\wedge$, $\vee$, $\leftrightarrow$, [,], $\bigwedge$, $\bigvee$, and $=$. A *variable* is a sequence v_n where n is a natural number, an *n-place predicate* is a sequence $p_{\alpha,n}$ where α is an ordinal, and an *n-place operation symbol* is a sequence $f_{\alpha,n}$ where α is an ordinal. A *predicate* or an *operation symbol* is a sequence which, for some natural number n, is an n-place predicate or an n-place operation symbol. An *atomic expression* is either a logical constant, a variable, a predicate, or an operation symbol. An *expression* is a finite sequence each of whose 1-place subsequences is an atomic expression. Concatenation of sequences is indicated by juxtaposition.

A *language* is a set of predicates and operation symbols. The set of *terms of* a language Γ is the smallest set Θ containing all variables and such that $F\zeta_0 \ldots \zeta_{n-1} \in \Theta$ whenever F is an n-place operation symbol in Γ and $\zeta_0, \ldots, \zeta_{n-1} \in \Theta$. An *atomic formula of* Γ is an expression having the form $\zeta = \eta$, where ζ, η are terms of Γ, or else the form $P\zeta_0 \ldots \zeta_{n-1}$, where P is an n-place predicate in Γ and $\zeta_0, \ldots, \zeta_{n-1}$ are terms of Γ. The set of *formulas of* Γ is the smallest set Φ containing all atomic formulas of Γ and such that (1) $\neg \phi \in \Phi$ whenever $\phi \in \Phi$, (2) $[\phi \rightarrow \psi]$, $[\phi \wedge \psi]$, $[\phi \vee \psi]$,

---

[3]) A closely related, though not identical, approach is given in Tarski [1].

[4]) For formulations of the various systems of set theory mentioned see Montague, Scott, and Tarski [1].

$[\phi \leftrightarrow \psi] \in \Phi$ whenever $\phi, \psi \in \Phi$, and (3) $\wedge x\phi$, $\vee x\phi \in \Phi$ whenever x is a variable and $\phi \in \Phi$.

A *standard atomic formula of* Γ is an expression of the form $Pv_0 \ldots v_{n-1}$, where P is an n-place predicate in Γ, or $Fv_0 \ldots v_{n-1} = v_n$, where F is an n-place operation symbol in Γ. (It will be recalled that $v_0, \ldots, v_n$ are the first $n+1$ variables.) The set of *standard formulas of* Γ is the smallest set containing all standard atomic formulas of Γ, together with all formulas $x=y$, where x, y are variables, and satisfying conditions (1)–(3) above.

The notion of a *free variable* is understood in the usual way, and a *sentence* (or *standard sentence*) *of* Γ is a formula (or standard formula) of Γ without free variables. By a *term*, *atomic formula*, *formula*, *sentence*, *standard atomic formula*, *standard formula*, or *standard sentence* is understood an expression which is respectively a term, atomic formula, formula, sentence, standard atomic formula, standard formula, or standard sentence of some language.

By a *derivation from* a set Φ of sentences is understood a finite sequence constructed in the usual way on the basis of members of Φ and the axioms and rules of first-order logic with identity. Several exact characterizations are possible; for an example see MONTAGUE and HENKIN [1]. A formula ψ is said to be *derivable* from a set Φ of sentences, in symbols $\Phi \vdash \psi$, if there is a derivation from Φ which contains ψ as a constituent, and two sentences are *logically equivalent* if each is derivable from the unit set of the other. It is easily seen that every sentence of a language Γ is logically equivalent to a standard sentence of Γ. A *theory* is a set Φ of sentences such that, for some language Γ, all members of Φ are sentences of Γ, and $\psi \in \Phi$ whenever ψ is a sentence of Γ and derivable from Φ. The language Γ is uniquely determined by the theory Φ and is called the *language of* Φ. If Φ is a theory and Ψ a set of sentences, then Φ is said to be *axiomatized by* Ψ if Ψ is a subset of Φ and all members of Φ are derivable from Ψ; and a theory is called *finitely axiomatizable* if it is axiomatized by some finite set of sentences.

We turn now to the syntactical characterization of interpretability and relative interpretability. Suppose that f is a function whose domain is a set of standard atomic formulas and whose range is a set of formulas. For any standard formula ϕ, we can define the *substitution in* ϕ *based on* f, or ϕ_f, by the following recursion: if ϕ is a standard atomic formula, then ϕ_f is $f(\phi)$ if ϕ is in the domain of f, and ϕ otherwise; if x, y are variables, then $(x=y)_f$ is $x=y$; if ϕ, ψ are standard formulas and x is a variable, then $(\neg\, \phi)_f$ is $\neg(\phi_f)$, $[\phi \rightarrow \psi]_f$ is $[\phi_f \rightarrow \psi_f]$, $(\wedge x\phi)_f$ is $\wedge x\phi_f$; and the other logical constants behave analogously. If Φ is a theory, Φ_f is to be the smallest theory containing ϕ_f whenever ϕ is a standard sentence and a member of Φ.

If Φ, Ψ are theories, then Φ is said to be *interpretable* in Ψ if there exists a function f such that (1) the domain of f is a set of standard atomic

formulas, (2) the range of f consists of formulas of the language of Ψ, (3) whenever ϕ is in the domain of f, $f(\phi)$ is a formula whose free variables are among those of ϕ, and (4) $\Phi_f \subseteq \Psi$.

Let P be a 1-place predicate. The *relativization of a formula* ϕ to P, or $\phi^{(P)}$, is defined by the following recursion: if ϕ is an atomic formula, then $\phi^{(P)}$ is ϕ; if ϕ and ψ are formulas, then $(\neg\, \phi)^{(P)}$ is $\neg\, \phi^{(P)}$, $[\phi \to \psi]^{(P)}$ is $[\phi^{(P)} \to \psi^{(P)}]$, $[\phi \wedge \psi]^{(P)}$ is $[\phi^{(P)} \wedge \psi^{(P)}]$, $[\phi \vee \psi]^{(P)}$ is $[\phi^{(P)} \vee \psi^{(P)}]$, and $[\phi \leftrightarrow \psi]^{(P)}$ is $[\phi^{(P)} \leftrightarrow \psi^{(P)}]$; if in addition x is the variable v_0, then $(\wedge x\phi)^{(P)}$ is $\wedge\, x[Px \to \phi^{(P)}]$ and $(\vee\, x\phi)^{(P)}$ is $\vee\, x[Px \wedge \phi^{(P)}]$; and if x is a variable other than v_0, then $(\wedge x\phi)^{(P)}$ is $\wedge x[\vee v_0[Pv_0 \wedge x = v_0] \to \phi^{(P)}]$ and $(\vee x\phi)^{(P)}$ is $\vee\, x[\vee\, v_0[Pv_0 \wedge x = v_0] \wedge \phi^{(P)}]$. It is obvious that if ϕ is a formula and x any variable, then $(\wedge\, x\phi)^{(P)}$ is logically equivalent to $\wedge\, x[Px \to \phi^{(P)}]$ and $(\vee\, x\phi)^{(P)}$ to $\vee\, x[Px \wedge \phi^{(P)}]$. The reason for giving less natural conditions in the recursive definition is to insure that if ϕ is a standard formula, $\phi^{(P)}$ will be one also.

The *relativization of a theory* Φ to P, or $\Phi^{(P)}$, is the smallest theory containing $\phi^{(P)}$ whenever $\phi \in \Phi$. A theory Φ is said to be *relatively interpretable* in a theory Ψ if $\Phi^{(P)}$ is interpretable in Ψ, for some 1-place predicate P not in the language of Φ.

The notions of interpretability and relative interpretability are due to Tarski and are defined in Tarski, Mostowski, and Robinson [1]. The definition of interpretability given in the first edition of that monograph, though equivalent to the present definition in the case of theories with finite languages, turned out to have undesirable consequences in the general case. A revised definition, completely equivalent to the one given here, is to appear in the forthcoming second edition of Tarski, Mostowski, and Robinson [1]. The present rather simple syntactical characterization of interpretability was worked out in collaboration with Feferman.

Lemma 1. Suppose that Φ is a theory, Ψ is a set of standard sentences of the language of Φ, Φ is axiomatized by Ψ, and f is a function satisfying conditions (1) and (3) of the definition of interpretability. Let Γ be the set of sentences ψ_f for which ψ is in Ψ; and let Δ be the set of sentences

$$\wedge\, v_0 \ldots \wedge\, v_{n-1} \vee\, v_{n+1} \wedge\, v_n[v_n = v_{n+1} \leftrightarrow (Av_n \ldots v_{n-1} = v_n)_f],$$

where n is a natural number and A is an n-place operation symbol in the language of Φ. Then Φ_f is axiomatized by $\Gamma \cup \Delta$.

Lemma 2. If Φ, Ψ are theories such that Φ is interpretable in Ψ, and P is a 1-place predicate not in the language of Φ or the language of Ψ, then $\Phi^{(P)}$ is interpretable in $\Psi^{(P)}$.

Proof. Let f be a function satisfying conditions (1)–(4) of the definition of interpretability (of Φ in Ψ). Let g be a function whose domain is the set of standard atomic formulas of the language of Φ and which satisfies the following conditions. If F is an n-place predicate in the language of Φ,

then $g(Fv_0 \ldots v_{n-1})$ is $(Fv_0 \ldots v_{n-1})_f{}^{(P)}$. If A is a 0-place operation symbol in the language of Φ, then $g(A=v_0)$ is $[(A=v_0)_f{}^{(P)} \wedge Pv_0]$. If $n>0$ and A is an n-place operation symbol in the language of Φ, then $g(Av_0 \ldots v_{n-1}=v_n)$ is the formula

$$[[Pv_0 \wedge \ldots \wedge Pv_n \wedge (Av_0 \ldots v_{n-1}=v_n)_f{}^{(P)}] \vee$$
$$[\neg [Pv_0 \wedge \ldots \wedge Pv_{n-1}] \wedge v_n=v_0]].$$

Now it is clear that $\Phi^{(P)}$ is axiomatized by the set of sentences $\phi^{(P)}$ for which ϕ is a standard sentence in Φ. Therefore, by Lemma 1,

(1) $\qquad \Phi^{(P)}{}_g$ is axiomatized by $\Gamma \cup \Delta$,

where Γ is the set of sentences $\phi^{(P)}{}_g$ for which ϕ is a standard sentence in Φ, and Δ is the set of sentences $\wedge v_0 \ldots \wedge v_{n-1} \vee v_{n+1} \wedge v_n[v_n = =v_{n+1} \leftrightarrow (Av_0 \ldots v_{n-1}=v_n)_g]$ such that n is a natural number and A is an n-place operation symbol in the language of Φ.

We show that

(2) $\qquad \Gamma \subseteq \Psi^{(P)}$.

Assume that ϕ is a standard sentence in Φ. Then ϕ_f is in Ψ. Therefore $\phi_f{}^{(P)}$ is in $\Psi^{(P)}$. But $\phi^{(P)}{}_g$ is logically equivalent with $\phi_g{}^{(P)}$, and hence with $\phi_f{}^{(P)}$. Therefore $\phi^{(P)}{}_g$ is in $\Psi^{(P)}$, and (2) is established.

It follows easily from the definition of g that $\Delta \subseteq \Psi^{(P)}$ and that the language of $\Phi^{(P)}{}_g$ is included in that of $\Psi^{(P)}$. Therefore, by (1) and (2), $\Phi^{(P)}{}_g$ is included in $\Psi^{(P)}$, and $\Phi^{(P)}$ is consequently interpretable in $\Psi^{(P)}$.

It has been mentioned in Tarski, Mostowski, Robinson [1] that the relation of interpretability is transitive. The following simple lemma makes the same assertion concerning relative interpretability.

Lemma 3. If Φ, Ψ, Γ are theories such that Φ is relatively interpretable in Ψ and Ψ in Γ, then Φ is relatively interpretable in Γ.

Proof. From the hypothesis it follows that $\Phi^{(P)}$ is interpretable in Ψ and $\Psi^{(Q)}$ is interpretable in Γ, for some distinct 1-place predicates P and Q not in the language of Φ or Ψ. Hence, by Lemma 2, $\Phi^{(P)(Q)}$ is interpretable in $\Psi^{(Q)}$. Let R be a new 1-place predicate. By considering the function f whose domain is $\{Rv_0\}$ and which is such that $f(Rv_0)$ is $[Pv_0 \wedge Qv_0]$, we see that $\Phi^{(R)}$ is interpretable in $\Phi^{(P)(Q)}$. Thus, by the transitivity of interpretability, $\Phi^{(R)}$ is interpretable in Γ, and Φ is relatively interpretable in Γ.

Two theories are said to be *mutually interpretable* if each is interpretable in the other. The following simple lemma will be useful.

Lemma 4. If Φ is any finitely axiomatizable theory, then there exists a finitely axiomatizable theory which is mutually interpretable with Φ and whose language contains no operation symbols.

Proof. Assume that at least one operation symbol is in the language of Φ, for otherwise the lemma is trivial. Let A be such an operation symbol, and let A have the further property that no operation symbol with a smaller number of places than A is in the language of Φ. Since Φ is finitely axiomatizable, there is a standard sentence ϕ in Φ such that all members of Φ are derivable from $\{\phi\}$; we may choose ϕ in such a way that A occurs in ϕ. Let f be a biunique function whose domain is the set of standard atomic formulas beginning with an operation symbol and occurring in ϕ, and which is such that whenever n is a natural number and C is an n-place operation symbol occurring in ϕ, $f(Cv_0 \ldots v_{n-1}=v_n)$ is $Pv_0 \ldots v_n$, for some $(n+1)$-place predicate P not in the language of Φ. Let Ψ be the intersection of all theories whose languages contain the predicates in the language of Φ, and which contain ϕ_f as well as all sentences

$$\wedge v_0 \ldots \wedge v_{n-1} \vee v_{n+1} \wedge v_n[v_n = v_{n+1} \leftrightarrow f(Cv_0 \ldots v_{n-1}=v_n)],$$

where n is a natural number and C is an n-place operation symbol occurring in ϕ. It is clear that Ψ is a finitely axiomatizable theory. To see that Φ is interpretable in Ψ, we need only consider the function $f \cup g$, where g is that function whose domain is the set of standard atomic formulas of the language of Φ beginning with an operation symbol but not occurring in ϕ, and which is such that whenever $Cv_0 \ldots v_{n-1}=v_n$ is such a formula, $g(Cv_0 \ldots v_{n-1}=v_n)$ is $f(Av_0 \ldots v_{k-1}=v_n)$, where k is the number of places of A. On the other hand, Ψ is obviously interpretable in Φ, using the function $\breve{f}$.

Let us now turn to model-theoretic notions. A *model* is an ordered pair $\langle A, f\rangle$, where A is a non-empty set and f is a function whose domain is a language and which assigns appropriate meanings to the predicates and operation symbols in that language; that is, $f(P)$ is a set of ordered n-tuples of elements of A whenever P is an n-place predicate in the domain of f, and whenever F is an n-place operation symbol in the domain of f, $f(F)$ is a set of ordered $(n+1)$-tuples of elements of A satisfying the condition that there is exactly one x for which $\langle a_0, \ldots, a_{n-1}, x\rangle \in f(F)$ whenever $a_0, \ldots, a_{n-1} \in A$. By the *universe* and the *language* of a model $\langle A, f\rangle$ are respectively meant A and the domain of f. We shall say that a finite sequence $\langle a_0, \ldots, a_{n-1}\rangle$ *satisfies* a formula ϕ in a model $\langle A, f\rangle$ under the usual conditions. These amount roughly to saying that the free variables of ϕ are among $v_0, \ldots, v_{n-1}$, ϕ is a formula of the language of the model, and ϕ holds when each of its free variables v_i is interpreted as denoting a_i, each of its predicates or operation symbols P as denoting $f(P)$, and its bound variables as ranging over A; logical constants, including the identity symbol, are to receive their usual meaning.[5]) A sentence ϕ is said to be *true* in a model $\langle A, f\rangle$ if ϕ is satisfied in $\langle A, f\rangle$ by every finite sequence of elements of A. A model $\mathfrak{A}$ is said to be a *model of* a set Φ of sentences if every member of Φ is true in $\mathfrak{A}$.

[5]) For an exact definition of satisfaction in a model see TARSKI and VAUGHT [1].

It is now possible to formulate the Completeness Theorem, which gives the model-theoretic characterization of the syntactical relation (1) considered at the beginning of this paper: if Φ and Ψ are sets of sentences, then all members of Ψ are derivable from Φ if and only if, for each model $\mathfrak{A}$ of Φ, if all members of Ψ are sentences of the language of $\mathfrak{A}$, then $\mathfrak{A}$ is a model of Ψ.[6])

§ 2. *Interpretability.*

For the analogous characterization of the relation of interpretability we must introduce another familiar model-theoretic notion, *definability*. If $\mathfrak{A}$ is a model and R an n-place relation over $\mathfrak{A}$ (that is, a set of ordered n-tuples of elements of the universe of $\mathfrak{A}$), then R is said to be *definable* in $\mathfrak{A}$ if there is a formula ϕ of the language of $\mathfrak{A}$ such that the free variables of ϕ are among $v_0, \ldots, v_{n-1}$ and, for all $a_0, \ldots, a_{n-1}$ in the universe of $\mathfrak{A}$, $\langle a_0, \ldots, a_{n-1}\rangle \in R$ if and only if $\langle a_0, \ldots, a_{n-1}\rangle$ satisfies ϕ in $\mathfrak{A}$. A *model* $\langle B, g\rangle$ is said to be *definable* in a model $\langle A, f\rangle$ if $B = A$ and every relation in the range of g is definable in $\langle A, f\rangle$. For $\langle B, g\rangle$ to be *relatively definable* in $\langle A, f\rangle$ we require again that every relation in the range of g be definable in $\langle A, f\rangle$, but replace the requirement that $B = A$ by the weaker condition that B (or, more exactly, the 1-place relation corresponding to the set B) be definable in $\langle A, f\rangle$.

The model-theoretic characterization of interpretability can now be given.

Theorem 1. If Φ is a theory and Ψ is a finitely axiomatizable theory, then Ψ is interpretable in Φ if and only if, for each model $\mathfrak{A}$ of Φ, there is a model of Ψ which is definable in $\mathfrak{A}$.

Proof. Assume the hypothesis. The implication from left to right in the conclusion is obvious. In proving the converse implication, we shall first consider the case in which the language of Ψ contains no operation symbols. Let us make this assumption, and assume in addition that

(1) for every model $\mathfrak{A}$ of Φ, there is a model of Ψ which is definable in $\mathfrak{A}$.

Let F be the set of functions f such that the domain of f is the set of standard atomic formulas of the language of Ψ and, for each such formula ϕ, $f(\phi)$ is a formula of the language of Φ whose free variables are among those of ϕ; and let ψ be a member of Ψ which is a standard sentence and such that all members of Ψ are derivable from $\{\psi\}$.

I shall show that

(2) there is a finite subset D of F such that, for every model $\mathfrak{A}$ of Φ, there exists f in D such that ψ_f is true in $\mathfrak{A}$.

---

[6]) See GÖDEL [1], where only those sets of sentences are considered which are at most denumerable. It was Henkin's observation, in HENKIN [1], that the Completeness Theorem could be extended so as to apply to arbitrary sets of sentences.

Assume that (2) does not hold. It follows that for every finite subset D of F, there is a model of $\Phi \cup \{\neg\, \psi_f : f \in D\}$. Hence, by the Compactness Theorem,[7]) there is a model $\mathfrak{A}$ of $\Phi \cup \{\neg\, \psi_f : f \in F\}$. But it follows from (1) that there exists f in F such that ψ_f is true in $\mathfrak{A}$. We have thus arrived at contradiction, and (2) is proved.

Now let $\{f_0, \ldots, f_n\}$ be a finite subset of F satisfying (2). (Since F is not empty, we may clearly assume, as is implicit in this representation, that the subset contains at least one member.) Then if $\mathfrak{A}$ is any model of Φ, the disjunction $[\psi_{f_0} \vee \ldots \vee \psi_{f_n}]$ is true in $\mathfrak{A}$. Hence, by the Completeness Theorem,

$$\Phi \vdash [\psi_{f_0} \vee \ldots \vee \psi_{f_n}]. \tag{3}$$

Let f be that function whose domain is the set of standard atomic formulas of the language of Ψ and which is such that, for every such formula ϕ, $f(\phi)$ is the formula

$$[[\psi_{f_0} \wedge f_0(\phi)] \vee [\neg\, \psi_{f_0} \wedge \psi_{f_1} \wedge f_1(\phi)] \vee \ldots \vee [\neg\, \psi_{f_0} \wedge \ldots \wedge \neg\, \psi_{f_{n-1}} \wedge \psi_{f_n} \wedge f_n(\phi)]].$$

For every ϕ in the domain of f, we have:

$$\{\psi_{f_0}\} \vdash [f(\phi) \leftrightarrow f_0(\phi)],$$
$$\{\neg\, \psi_{f_0}, \psi_{f_1}\} \vdash [f(\phi) \leftrightarrow f_1(\phi)],$$
$$\vdots$$
$$\{\neg\, \psi_{f_0}, \ldots, \neg\, \psi_{f_{n-1}}, \psi_{f_n}\} \vdash [f(\phi) \leftrightarrow f_n(\phi)].$$

Hence

$$\{\psi_{f_0}\} \vdash [\psi_f \leftrightarrow \psi_{f_0}],$$
$$\{\neg\, \psi_{f_0}, \psi_{f_1}\} \vdash [\psi_f \leftrightarrow \psi_{f_1}],$$
$$\vdots$$
$$\{\neg\, \psi_{f_0}, \ldots, \neg\, \psi_{f_{n-1}}, \psi_{f_n}\} \vdash [\psi_f \leftrightarrow \psi_{f_n}].$$

Therefore, by (3) and sentential logic,

$$\Phi \vdash \psi_f;$$

and under our special assumption that the language of Ψ contains no operation symbols, this is sufficient to show that Ψ is interpretable in Φ.

Let us now turn to the general situation, in which the language of Ψ may contain operation symbols, and again make the assumption (1). By Lemma 4 there exists a finitely axiomatizable theory Ψ' which is mutually interpretable with Ψ and whose language contains no operation symbols. From the fact that Ψ' is interpretable in Ψ, together with our assumption (1) concerning Ψ, it is seen that (1) holds for Ψ'. Hence, by the case considered above, Ψ' is interpretable in Φ; and therefore so is Ψ, by the fact Ψ is interpretable in Ψ'.

---

[7]) The Compactness Theorem, which is an immediate consequence of the Completeness Theorem as stated above, asserts that if every finite subset of a given set of sentences has a model, then the set itself has a model.

By a result in DE BOUVÈRE [1], based on work in KEISLER [1], the assertion that a relation R is definable in a model $\mathfrak{A}$ is equivalent to a purely mathematical condition on R and $\mathfrak{A}$, that is, a mathematically natural condition which does not involve, as does the notion of definability, any reference to metamathematical objects. (The notion of a purely mathematical condition can of course not be defined in mathematical terms, but has instead an esthetic character.) It follows by Theorem 1 that under the hypothesis of that theorem the assertion that Ψ is interpretable in Φ is equivalent to a purely mathematical condition on the class of models of Φ and the class of models of Ψ.

Ehrenfeucht has shown by means of an unpublished example that Theorem 1 becomes false if the hypothesis that Ψ be finitely axiomatizable is omitted. On the other hand, Orey and Feferman have shown that if Φ, Ψ are *recursively enumerable* theories with finite languages and Φ is *reflexive*, then Ψ is interpretable in Φ if and only if every finitely axiomatizable theory included in Ψ is interpretable in Φ.[8]) Feferman has pointed out that consequently, on the basis of Theorem 1, the hypothesis in Theorem 1 that Ψ be finitely axiomatizable can be replaced by the condition that Φ and Ψ both have finite languages, that they both be recursively enumerable, and that Φ be reflexive.

Two open problems suggest themselves. One might try to find a simple model-theoretic condition equivalent to interpretability for arbitrary theories. On the other hand, one might consider the model-theoretic relation in Theorem 1, that is, the condition that for each model $\mathfrak{A}$ of Φ, there is a model of Ψ which is definable in $\mathfrak{A}$, and try to find a simple syntactical condition equivalent to that condition for arbitrary theories Φ and Ψ.

We easily derive from Theorem 1 a corresponding model-theoretic characterization of relative interpretability, again under an assumption of finite axiomatizability.

Theorem 2. If Φ is a theory and Ψ is a finitely axiomatizable theory, then Ψ is relatively interpretable in Φ if and only if, for each model $\mathfrak{A}$ of Φ, there is a model of Ψ which is relatively definable in $\mathfrak{A}$.

Proof. Assume the hypothesis. As with Theorem 1, the implication in the conclusion from left to right is obvious. Assume, then, that

(1) for every model $\mathfrak{A}$ of Φ, there is a model of Ψ which is relatively definable in $\mathfrak{A}$.

---

[8]) See FEFERMAN [1], Theorem 8. 10. A theory Φ is *reflexive* if, loosely speaking, one can prove in Φ the consistency of each finitely axiomatizable theory included in Φ. For the exact notion of a reflexive theory see MONTAGUE [1], where it was shown that a number of familiar theories, for instance, all extensions of (first-order) Peano's arithmetic or of Zermelo-Fraenkel set theory are reflexive. The characterization of recursive enumerability as applied to theories with finite languages presents no problem.

By Lemma 4, there is a finitely axiomatizable theory Ψ' which is mutually interpretable with Ψ and whose language contains no operation symbols. From the fact that Ψ' is interpretable in Ψ we conclude that (1) holds for Ψ' as well as for Ψ. Let P be a new 1-place predicate. It follows that for every model $\mathfrak{A}$ of Φ, there is a model of $\Psi'^{(P)}$ which is definable in $\mathfrak{A}$; also, because Ψ' is finitely axiomatizable and has a language containing no operation symbols, $\Psi'^{(P)}$ is finitely axiomatizable. Thus we may apply Theorem 1 and conclude that $\Psi'^{(P)}$ is interpretable in Φ; therefore Ψ' is relatively interpretable in Φ. But Ψ is relatively interpretable, because interpretable, in Ψ'. Hence, by Lemma 3, Ψ is relatively interpretable in Φ.

As immediate corollaries of Theorems 1 and 2 we have the following results concerning the syntactical notions of interpretability and relative interpretability. A theory Ψ is called an *extension* of a theory Φ if Ψ includes Φ and has the same language as Φ; and a *complete* theory is a theory Φ such that for every sentence ϕ of the language of Φ, either ϕ or $\neg\phi$ is in Φ.

Theorem 3. If Φ is a theory and Ψ a finitely axiomatizable theory, then Ψ is interpretable in Φ if and only if Ψ is interpretable in every complete extension of Φ.

Theorem 4. If Φ is a theory and Ψ a finitely axiomatizable theory, then Ψ is relatively interpretable in Φ if and only if Ψ is relatively interpretable in every complete extension of Φ.

University of California,
Los Angeles

REFERENCES

BOUVÈRE, K. L. DE [1]. A mathematical characterization of explicit definability. Indagationes Mathematicae, **25**, 264–274 (1963).

FEFERMAN, S. [1]. Arithmetization of metamathematics in a general setting. Fundamenta Mathematicae, **49**, 35–92 (1960).

GÖDEL, K. [1]. Die Vollständigkeit der Axiome des logischen Funktionenkalküls. Monatshefte für Mathematik und Physik, **37**, 349–360 (1931).

HENKIN, L. [1]. The completeness of the first-order functional calculus. The Journal of Symbolic Logic, **14**, 159–166 (1949).

KEISLER, H. J. [1]. Ultraproducts and elementary classes. Indagationes Mathematicae, **23**, 477–495 (1961).

MONTAGUE, R. [1]. Contributions to the axiomatic foundations of set theory. Dissertation, Berkeley (1957).

————, and L. HENKIN [1]. On the definition of 'formal deduction'. The Journal of Symbolic Logic, **21**, 129–136 (1956).

————, D. S. SCOTT and A. TARSKI [1]. An axiomatic approach to set theory. Amsterdam, forthcoming.

TARSKI, A. [1]. A simplified formalization of predicate logic with identity. Archiv für mathematische Logik und Grundlagenforschung, forthcoming.

————, A. MOSTOWSKI and R. M. ROBINSON [1]. Undecidable theories. Amsterdam (1953).

————, and R. L. VAUGHT [1]. Arithmetical extensions of relational systems. Compositio Mathematica, **13**, 81–102 (1957).