CONTENTS

PART THREE: REFERENCE, ESSENCE, MIND, AND METAPHYSICS

CHAPTER 1

The Prehistory of Modal Logic

C. I. Lewis, Ruth Marcus, W.V.O. Quine

Every philosophical problem, when it is subjected to the necessary analysis and purification, is found to be not really philosophical at all, or else to be, in the sense in which we are using the word, ***logical***.
—Russell (1914a), p. 33

Philosophical propositions . . . must be ***a priori***. *A philosophical proposition must be such as can neither be proved or disproved by empirical evidence. . . . Philosophy is the science of* ***the possible***. *. . . Philosophy, if what has been said is correct, becomes indistinguishable from* ***logic***.
—Russell (1914b)

1. LOGIC AND LOGICAL TRUTH: FREGE, RUSSELL, AND TARSKI

These two passages were written by Russell thirty-five years after Frege's invention of modern symbolic logic in the service of reducing first arithmetic, and then mathematics in general, to logic, and four years after Russell and Whitehead attempted a similar reduction. Although those attempts failed to achieve their ambitious goals, they did bring logic, the philosophy of mathematics, and the practice of logical analysis to the forefront of philosophical inquiry. By 1914, the practice was advancing in broad areas of epistemology and metaphysics.

The first of Russell's two remarks suggests that all true answers to philosophical questions are, in some sense, *truths of logic*. The second adds that these answers are knowable a priori. Russell adds that philosophy's job is to explore the limits of the *possible*, which requires distinguishing it from the *impossible*, from which the *necessary* is definable. Like logical truths, Russell seemed to think that philosophical truths must be both necessary and knowable a priori, if knowable at all. In suggesting this, he was trafficking in logical, epistemic, and metaphysical modalities without clearly distinguishing them. A central task of this and parts of the next two volumes will be to trace the struggle in mid- to late twentieth-century philosophy to understand and distinguish these modalities.

We may begin by comparing Frege's conception of logic to the now standard Tarskian conception, which is a mixture of formalized linguistic elements—sentences, formulas, and their constituents—plus constraints on their interpretations. These contrast with Fregean concepts and propositions, made up of intuitively understood epistemic and metaphysical elements. The rules of logic were, for Frege, rules of guaranteed truth preservation, some of which were more cognitively fundamental than others. The truths of logic were said to be entirely general in the sense that our knowledge of them doesn't depend on any special subject matter.

Logic was, for Frege, our most powerful tool for deriving new knowledge from prior knowledge.

> I became aware of the need for a *Begriffsschrift* [concept script] when I was looking for the fundamental principles or axioms upon which the whole of mathematics rests. Only after this question is answered can it be hoped to treat successfully the springs of knowledge upon which this science thrives.[1]
>
> The firmest proof is obviously the purely logical kind, which, prescinding from the particularity of things, is based solely *on the laws on which all knowledge rests*. Accordingly, we divide all truths that require justification into two kinds: those whose proof can be given purely logically and those whose proof must be grounded on empirical facts.[2]

Here Frege suggests (i) that all knowledge rests to some degree on knowledge of fundamental logical laws and principles; (ii) that knowledge of them and their logical consequences *isn't* grounded in empirical facts, and so, presumably, is a priori; and (iii) that all other knowledge is a posteriori.

In volume 1 of *Basic Laws of Arithmetic* he adds:

> *Now the question of why and with what right we acknowledge a logical law to be true, logic can only answer by reducing it to another logical law.* Where that is not possible, logic can give no answer. Leaving aside logic, we may say: we are forced to make judg-

1. Frege (1897–91), p. 3.
2. Frege (1964), preface, section 3, p. 48, my emphasis.

> ments by our nature and external circumstances; and if we make judgments, we cannot reject this law—of identity, for example; we must recognize it if we are not to throw our thought to confusion and in the end renounce judgment altogether. I do not wish to either dispute or endorse this view and only remark that what we have here is not a logical implication. What is given is not a ground [reason] for [something's] being true, but of our holding [it] as true.[3]

Here Frege acknowledges an epistemic hierarchy of logical laws and principles. Those on higher levels are justified by deriving them from those on lower levels, culminating in the most fundamental logical laws, for which no further justifying reason can be given, but without which we can scarcely reason at all. Because he took arithmetic to be derivable from the most fundamental logical laws, he claimed that the same is true of "the fundamental propositions of the science of number." He says, "We have only to deny any one of them and complete confusion ensues. *Even to think at all seems no longer possible.*"[4]

These passages suggest that Frege was an epistemic foundationalist about logical truths, the most basic of which we know with a priori certainty, which, he thought, could be extended to arithmetic and most mathematics.[5] Although he doesn't talk explicitly about metaphysical necessity or possibility, he takes fundamental logical laws to be true, while recognizing that any logical consequence of a truth *must be true*, seeming, thereby, to implicitly recognize the *necessity* of certain conditionals. Since he also thinks that fundamental logical truths *must* be true, he seems to accept their necessity. For Frege, it is not sentences but thoughts (propositions) that are logical truths, all of which are both necessary and a priori. Whether or not the converse holds is less clear. Like Russell in 1914, he intermingled logical, epistemic, and metaphysical modalities.[6]

The conception of logical truth and logical consequence descending from Tarski (1935, 1936) is different. Rather than propositions, logical truths are sentences of a formalized language (e.g., of the first-order predicate calculus) that are *true in all models*—where a model is a nonempty domain D of objects the language is used to talk about plus an assignment of objects, sets of objects, and functions from objects to objects as denotations of its nonlogical symbols. Names denote individual objects, n-place predicates denote sets of n-tuples of objects, and n-place function symbols designate n-place functions from n-tuples of objects to individual objects. The denotation of $\ulcorner f\ (t_1 \ldots t_n) \urcorner$ is the object that the function denoted by f assigns to the denotations of $t_1 \ldots t_n$. Logical truths are sentences the truth of which survives all reinterpretations of its nonlogical symbols over domains of any size.

3. Frege (1997), preface, section xvii, p. 204, my emphasis.
4. Frege (1884 [1950]), section 14, p. 21, my emphasis.
5. Frege made a special exception for Euclidian geometry, about which he was a Kantian. See p. 41 of volume 1 of this work.
6. A fuller discussion can be found in section 4 of chapter 1 of volume 1 of this work.

Truth in a model M is then defined as follows.[7]

An atomic formula ⌈$Pt_1 \ldots t_n$⌉ is true in a model M iff the n-tuple of denotations of $t_1 \ldots t_n$ in M is a member of the denotation of P in M.

⌈~ S⌉ is true in M iff S is not true in M.

⌈S & R⌉ is true in M iff S and R are both true in M.

Similar clauses are given for ⌈S v R⌉, ⌈S ⊃ R⌉, and ⌈S ↔ R⌉.

⌈∃v S(v)⌉ is true in M iff there is an object o in the domain of M such that S(v) is true in M, when the variable v is temporarily used as a name of o. ⌈∀v S⌉ is true in M iff for every object o in the domain of M, S(v) is true in M no matter which o of the domain is temporarily assigned to v as referent.

A *logical truth* is a sentence that is true in all models, and a *logical consequence* of a set P of sentences is a sentence that is true in every model in which all members of P are true.

These definitions don't mention what is provable or derivable, which varies with the richness of the underlying language and the associated proof theory. For standard first-order logic there are proof theories that allow one to derive (in finitely many steps) every logical truth and no others from the rules plus the designated axioms of the system. Which sound and complete proof theory one uses isn't a matter of logic itself. Thus, no Fregean epistemic hierarchy of justification and knowledge is presupposed. We do, however, get laws of guaranteed truth preservation that encode many of our argumentative commitments (thought of as propositions one is committed to by accepting or asserting other propositions).

This claim rests on three Frege-friendly assumptions: (i) sentences express propositions we assert when we assertively utter sentences; (ii) the propositions *that S* and *that S is true* are necessary, a priori consequences of one another; and (iii) any warrant for believing one of those propositions is warrant for believing the other. Suppose I assert propositions p1, p2, p3 in an argument by assertively uttering sentences S1, S2, S3—knowing that the S's express the p's and that my audience also knows this. In so doing, I commit myself to p1, p2, p3. By (ii) and (iii), I also commit myself to the truth of each, plus the truth of S1–S3. Since I am committed to anything trivially inferable from what I *explicitly assert*, I am also committed to the conjunction of p1, p2, and p3 and to the claim that it is true. Similarly for the sentence ⌈S1&S2&S3⌉. Finally, suppose S+, which expresses p+, is a *logical consequence* of my (sentential) premises, and hence of their conjunction. Then, since my premises can't be true unless S+ is too, *I can't deny the truth of S+* without undermining my warrant for *the truth of* the conjunction of p1, p2, p3 and for the conjunction itself. Since I under-

7. The definition is a simplification in which variables are allowed (under certain conditions) to function as temporary names. For Tarski's original formalization, see section 4 of chapter 9 of volume 2 of this work. For interpretation of the ⌈ . . . ⌉ notation, see "A Word about Notation" preceding chapter 1.

stand that S+ expresses p+, I am *implicitly* committed to the truth of p+ and to p+ itself (whether I recognize it or not).

In short, our post-Tarskian conception of *logical consequence* can be used to track many (though not all) *propositions* to which we commit ourselves in reasoned argument. This isn't a claim of post-Tarskian logic, which, because it is fully formalized, is a meta-mathematical discipline. It is a claim about why, despite making no cognitive claims about what we believe, assert, or know, that discipline retains a recognizably logical role in guiding proper *reasoning*. This is not to say that other, nonlogical, relations between propositions based on the objects and concepts they concern are not needed to more fully account for proper reasoning. They are. These include nonlogical but a priori connections, necessary connections, and probabilistic connections. However, it would be many decades before these would be fleshed out in useful detail.

This brings us back to Frege and Russell. Do first-order logical truths like $S \supset S$ and $\sim (S \& \sim S)$ express *propositions* that are *necessary a priori truths*? Do their negations express *impossibilities*? Special complications aside—involving vagueness and/or partially defined predicates/properties—the natural answer to both questions is 'Yes'.[8] With this we return to the mixture of logical and seemingly nonlogical modalities found in Frege and Russell. What exactly are these modalities, and how do they bear on one another? This will be one of the main questions underlying most of this chapter, most of this volume, and some of the next two.

2. C. I. LEWIS: A LOGIC OF NECESSITY AND POSSIBILITY?

Lewis was concerned with logical implication. Before illustrating his concern, it will be helpful to summarize how we understand *logical implication* today. Let P and Q be sentences of the propositional or first-order predicate calculus. To say that P *logically implies* Q is to say that Q is a (Tarskian) logical consequence of P, which is a metalogical remark *about* a logical system, rather than a remark expressible in it. The fact that the system contains no 2-place sentential connective holding between sentences (and no 2-place predicate true of them) if and only if the second is a logical consequence of the first is no defect, because formal logic doesn't require any such connective (or predicate) in order to play its role in guiding reasoning when we use its sentences. The fact that the nonlogical vocabulary is fully formalized and capable of varying interpretations, while the logical vocabulary is fixed, makes it possible to define appropriately general definitions of logical truth and logical consequence while investigating them mathematically.

These points are relevant to Lewis's "Implication and the Algebra of Logic" published in 1912, in which he comments on the logic of *Principia Mathematica*,

8. For some discussion of these complications, see the interchange between Tim Williamson and me in the symposium including Soames (2002b). See also Soames (2009c, 2018b).

volume 1, published two years earlier. In his article, he focuses on what *Principia* calls 'implication'—namely, any instance of the schema '$P \supset Q$', defined as '$\sim P \vee Q$', taking a disjunction to be true whenever either disjunct is true. Since this makes every sentence Q an "implication" of any falsehood P, Lewis's determination to read '$\supset$' as *implies* led him, via the trivial theorem schema [$\sim P \supset (P \supset Q)$], to characterize *Principia* as committed to the "absurd" conclusion that "a false proposition implies anything."[9] Although what Lewis suggests by this is absurd, it is tempting to suppose that his criticism is simply the result of confusing what is standardly called "material implication" with logical implication. That is just a confusion. Still, Lewis may have had something more in mind.

The following paragraph gives us an inkling of what it may have been. It starts out this way.

> The existence of these two theorems [that "a false proposition implies anything" and that "a true proposition is implied by any proposition"] in the algebra [of *Principia*] brings to light the most severe limitation of the algebraic or *material implication.* One of the important practical uses of implication is the testing of hypotheses whose truth or falsity is problematic [or uncertain]. The algebraic implication [material implication] has no application here. If the hypothesis happens to be false, it implies anything you please. If one finds facts, *x, y, z,* otherwise unexpected but suggested by the hypothesis, the truth of these facts is implied by the hypothesis, whether that hypothesis is true or not—since any true proposition is implied by all others. In other words, no proposition could be verified by its *logical* consequences. If the proposition happens to be false, it has these consequences anyway.[10]

This seems to embody a confusion, or an oversight, or both. Of course, one must not confuse logical implication (the converse of logical consequence) with the relation between P and Q when the former materially implies the latter. As I pointed out earlier, logical consequence can be used to track our argumentative commitments, including those confirming or falsifying a hypothesis. Lewis may have wished for a clear and mathematically tractable definition of logical implication/consequence of this sort that would not be available until the advent of model theory, while wrongly imagining that the relevant implication relation *must itself be expressible in the language of the hypothesis being tested.* If so, he was wrong. But that isn't the end of the matter.

The paragraph quoted above continues as follows:

> Similarly, no contrary-to-fact supposition could have logical significance, whether one happens to know that it *is* contrary to fact or not. For if the fact *is* otherwise, the proposition which states the supposition [materially] implies everything. In the ordinary and "proper" use of implies certain conclusions can validly be inferred from contrary to fact suppositions, while certain others cannot. Hypothe-

9. Lewis (1912), p. 529.
10. Ibid., my emphasis.

> ses whose truth is problematic [unknown] have consequences *which are independent of their truth or falsity.*[11]

Apart from beating a dead horse, Lewis is here setting the stage for the introduction of a new notion of implication. To illustrate the idea, let's start with a contemporary philosophical conception of *metaphysical necessity* according to which *Necessarily water is* H_2O is widely regarded to be true. With this in mind, consider the contrary-to-fact supposition, *Suppose the bottle was/were/had been filled with water*, and the inference, *then it would have contained* H_2O. The related counterfactual conditional could be tested by looking for chemical residue. The idea is that the conditional *if the bottle was/were/had been filled with water, then it would have contained* H_2O expresses a useful "intensional" sense of implication, distinct from material implication, in which the antecedent *implies* the consequent. Presumably, Lewis didn't have this example in mind, but he may have imagined something similar.

Next consider a notion of *logical necessity* according to which S is logically necessary iff it is a Tarskian logical truth or it can be turned into one by putting synonyms for synonyms. (We leave open how *synonymy* is defined.) Lewis's point may be that some such notion of *strict implication—It is (logically) necessary that if P, then Q*—is more useful than anything we are given in *Principia Mathematica.*[12] As we will see in the next few chapters, there was something right about this, even though it is questionable how much of what we need is a stronger *logic.* The important point here is the initial steps he took in what we can, in retrospect, see to have been in a positive direction.

Lewis's disappointment with the logic of *Principia Mathematica* leads him to suggest a new "intensional" sense of disjunction, 'or_i', in which a disjunction is true if and only if the ordinary "extensional" disjunction is a necessary truth. Strict implication—$P \Rightarrow Q$—is then defined as an intensional disjunction, $\sim P$ *or* Q, which amounts to *It is (logically) necessary that* $(P \supset Q)$. Thus, we get a sense of implication in which Q cannot be false when P is true. Let's not quibble over the fact that we could, in principle, construct a strong implication relation in the metalanguage using the ordinary Tarskian notion of *logical consequence* plus names of object-language sentences. Lewis's project of including necessary conditionals in the object-language was more ambitious.

Perhaps too ambitious. He closes his 1912 article by suggesting a system containing both intensional and extensional disjunctions, including all *Principia* theorems plus new ones involving the strict conditional. He says:

> The primary advantage [of the new system] over any present system lies in the fact that its meaning of implication [strict implication] is precisely that of ordinary inference and proof.[13]

11. Ibid., my emphasis.
12. Ibid., p. 530.
13. Ibid., p. 531.

Not quite. It is no part of our ordinary practices of reasoning or constructing proofs to infer any truth Q that happens to be (logically) necessary given any arbitrary premise P. Still, one gets the general idea.

Lewis (1918, 1923) continue the attempt to provide more powerful logics than *Principia Mathematica*. The idea was to embed the extensional logic of *Principia* in a broader intensional system built around a notion of strict implication designed to capture what he took to be the proper ordinary notion *implies*. This, he believed, could be defined equally well as $\Box (P \supset Q)$ or $\sim\Diamond (P \,\&\, \sim Q)$—where '$\Box$' is read *necessarily*, or *it is a necessary truth that* while '$\Diamond$' is read *possibly* or *it could be, or have been, true that.* In this sense of implication, necessary falsehoods *imply* every proposition and necessary truths are *implied by* every proposition. Although one might doubt that this sense of implication captures any notion of inference commonly employed in ordinary reasoning, Lewis thought otherwise. In Lewis and Langford (1932, p. 250) he argues that necessary falsehoods are contradictory, and that any conclusion can be derived from, and hence is implied by, a contradiction.

The appendix to that work contains five modal systems of the propositional calculus in ascending order of strength, which have become well-known. Systems S1–S3 draw from three possible rules of modal inference and four possible modal axioms.

Rules

R1 *Weak Necessitation.* This allows one to derive $\Box \Phi$ from any theorem Φ of the nonmodal propositional calculus.
R2 Derive $\Box(\Box\Phi\supset\Box\Psi)$ from $\Box(\Phi\supset\Psi)$.
R3 *Substitutivity of Strict Equivalents*, which allows one to derive Φ from any theorem strictly equivalent to it. (P, Q are strictly equivalent if each strictly implies the other.)

Axioms

T. $\Box P \supset P$
K. $\Box (P \supset Q) \supset (\Box P \supset \Box Q)$
A1. $\Box (P \supset Q) \supset \Box(\Box P \supset \Box Q)$.
A2. $[\Box (P \supset Q) \,\&\, \Box (Q \supset R)] \supset \Box (P \supset R)$.

Lewis's system S1 can be formalized using R1 and R3 plus axioms T and A2. S2 is formalizable using R1 and R2 plus axioms T and K. Adopting R1 plus T and A1 gives us S3.

The next group of systems, which have received the most attention, employ *Strong Necessitation*, which allows one to derive $\Box\Phi$ from any theorem Φ. The weakest of these new systems (system K) adopts K as its only modal axiom, while the system T adds axiom T to K. S4 adds A3 to system T. S5 adds a new axiom, B, to S4.

A3. $\Box P \supset \Box \Box P$
B. $P \supset \Box \Diamond P$

Are any of these systems correct? Given that these were supposed to be systems of *logic*, one wonders which (if any) is the correct system of *logic*. Answering this question requires some prior notion of what logic amounts to. To begin, we note that logic provides norms of correct reasoning that apply to all subject matters. Following those norms must guarantee that whatever we derive from truths must themselves be true. Thus, in the case of the standard nonmodal propositional or predicate calculi, we look to logic for fully general rules guaranteeing truth preservation—rules that are sound in the sense of deriving only Tarskian model-theoretic logical consequences from the premises of any argument. In the propositional and first-order predicate calculi we can provide rules that are also complete, in allowing every logical consequence of any premises to be derived.

Next, suppose we add □, understood as *it is logically true that*, to these standard Tarskian systems, taking it to be a new logical symbol *the interpretation of which does not vary from model to model*. With this, we may argue, following Burgess (2009, section 3.8), that the system will have at least the power of S4. The argument requires understanding logical truth in the propositional and first-order predicate calculus in the way just sketched—as truth in all models. First consider axiom T, $\Box P \supset P$. Since its antecedent *says that P is true in all models*, for the antecedent to be true P must also be true. Hence T is true. Next consider axiom K, $\Box (P \supset Q) \supset (\Box P \supset \Box Q)$. By the same reasoning, for the antecedent of K to be true is for $(P \supset Q)$ to be true under all interpretations of P, Q. Next, for the antecedent of the consequent of K to be true is for P to be true under all interpretations. It then follows that Q is true under all interpretations, which means that the consequent of K, and so K itself, must be true. Next consider the characteristic axiom of S4, $\Box P \supset \Box \Box P$. Since the box—*it is a logical truth that*—is a logical symbol with a fixed interpretation which doesn't vary from model to model, the truth of all reinterpretations of P guarantees both the truth of the antecedent and the truth of the consequent, which come to the same thing. We might put this by saying that the logic of *it is a logical truth that* must be at least as strong as S4.

What about S5, which comes from S4 by adding axiom B, $P \supset \Box \Diamond P$? If the antecedent of B is true, then $\sim P$ isn't a logical truth, which gives us $\sim \Box \sim P$ and hence $\Diamond P$. Since this holds for all interpretations of P, and all interpretations of $\Diamond P$, it holds for all interpretations of $\Box \Diamond$ P, thus giving us axiom B. So, we may conclude, that the logic of *it is a logical truth that* must be at least as strong as S5. In the case of the modal propositional calculus a further argument can be given that for any formula A of the calculus that isn't a theorem of S5, there will be interpretations in which A is not true. Since this means that A is not a logical truth of the calculus, it suggests that S5 is the strongest acceptable system in which the box is interpreted as *truth in all models/interpretations*.[14] However, it is not clear that we learn very much from this.

14. Burgess (2009), pp. 65–66.

One worry concerns what we take the operator □—*it is logically true that*—to operate on. Because the presence of *that* in English standardly signals indirect discourse, this puts the operator in a class with the following, which are naturally taken to operate on propositions, rather than sentences or truth values.

it is knowable a priori that
it is a necessary truth that
it is a logical truth that
it is universally believed that
it is true that

There is nothing comparable in standard, nonmodal first-order logic, where all the operators are defined on sentences or formulas. Consider what it would mean to add to standard logic *it is a logical truth that.* There is something troubling about this right off the bat. Consider *For all x, x is a square iff x is a rectangle with 4 equal sides.* This is not a logical truth in the standard Tarskian sense—one which would turn out true in all domains under all interpretations of its nonlogical constants. However, it can (we may suppose) be turned into one by putting synonyms for synonyms.

With this in mind, consider two interpretations of *it is a logical truth that S.* According to the first it is true if and only if S is true under all reinterpretations of its nonlogical vocabulary. What is declared to have this property is sentence S itself. Under the second interpretation both S and any sentence resulting from putting synonyms for words or phrases of S count as logically true. Both interpretations make sense. But what is the point in adding extra synonymous sentences to standard Tarskian logical truth? The answer, I suppose, is that the extra sentences all (are used to) *say the same thing* as the original logical truth they are synonymous with. They all express the same proposition, where propositions are the things said, asserted, believed, and known. Consequently, if we think it is a central task of anything we properly call "logic" that it encode our argumentative commitments, extending the notion of a logical truth to apply to propositions may fall under a defensible conception of what logic is. Under this conception, we take the common locution, *that S*, of *indirect discourse*, to refer to the proposition commonly expressed by all sentences (in whatever language) that are strictly translatable into S. In so doing we, in effect, extend the notion *logical truth* to apply to propositions (in the manner sketched above) as well as sentences.

Let S be some standard Tarskian logical truth—one that comes out true under any interpretation of its nonlogical symbols. Then, *it is logically true that S* will be true. Will *it is logically true that S* be logically true? Presuming that the operator is a logical symbol, the interpretation of which doesn't vary from model to model, it will. So, we can iterate necessity operators as much as we want. But nothing interesting is going on. No new claim is made. The same is true if we substitute ~ *Possible* ~ for any such operator. Finally, the notion of *logical implication*—*it is logically true that* $(P \supset Q)$—we get is no advance over one that simply notes that Q is a logical consequence of P.

Although these results involving S4 and S5 are, in themselves, not very interesting, the possibility of interpreting other modal systems as *in some way* logical in our now modern sense is illustrated in *The Unprovability of Consistency: An Essay in Modal Logic*, published by George Boolos in 1979. There, formal proofs in a certain modal system called 'G', after Gödel, are used to prove significant extensions of his two incompleteness theorems.[15] This illustrates that there are ways of understanding modal systems in which they can be sources of significant metalogical results for standard nonmodal logical systems. Like the weak systems, K and T, G is *normal*—i.e., it incorporates both Strong Necessitation and axiom K. However, very much unlike other well-known normal systems it doesn't incorporate axiom T, *which indicates that, in G, the box isn't an attempt to explicate any ordinary sense of necessity, logical or otherwise.* Still, it may encode *provability* (in a given system of proof). In other words, the box can coherently be read *it is provable (in such and such system) that.* The other axioms of G are sentences of the form $\Box\,(\Box P \supset P) \supset \Box\,P$, from which it follows that axiom T can't, on pain of inconsistency, be a theorem of G.[16]

Having come this far, we have moved well beyond the initial ideas of C. I. Lewis. In doing so, I hope to have shed some light on his conviction that intensional object-language operators may contribute to the norms governing our reasoning about various subjects, while leaving it open how much, or in what ways, the development of systems containing such operators track what we feel comfortable calling 'logic'. I will return to this topic after the model theory of modal logic has been presented more fully.

3. QUANTIFIED MODAL LOGIC: MARCUS, QUINE, SMULLYAN, FITCH, AND CHURCH

3.1. Marcus: The First Systems of Quantified Modal Logic

The first systems of quantified modal logic to receive serious attention in the period after World War II were presented in three papers published in 1946 and 1947 by Ruth Barcan Marcus.[17] Although no semantic theory interpreting its formulas was given, axiomatic proof-theoretic systems were specified, and interesting theorems were proved. The system in Marcus (1946a) is a quantified version of Lewis's S2, encompassing rules R1 and R2 plus axioms T and K. The 1947 paper, which employs a version of Lewis's S4 system, builds on Marcus (1946a), purporting to prove the necessity of identity, $\forall y\ [x=y \supset \Box\ x=y]$. This turned out to be more challenging than one might imagine in the absence non-proof-theoretic semantic principles for understanding modal

15. These theorems are explained in chapter 8 of volume 2 of this work.
16. Boolos (1979). T can't be provable because of Gödel's second incompleteness theorem, the unprovability of consistency. If $\Box(\Box P \supset P)$ were a theorem—i.e., if we could prove that the theorems of a first-order theory T formulated in the predicate calculus were true—then we could prove that T was consistent, which, Gödel showed, cannot be.
17. Barcan (1946a, 1946b, 1947).

logics. The key idea employed in the now standard semantics is that variables are treated as temporary names for objects in the domain of quantification at a given world-state. Quantified sentences are assigned truth conditions relative to assignments of objects as values of the relevant variables. So, if 'x' and 'y' are assigned the same object—making the semantic content of $x=y$ relative to an assignment the same as that of $x=x$—then, of course, $x=y$ will be true, relative to that assignment, at all world-states possible from the original world-state, thus validating *Necessarily* $x=y$ (assuming that variables are rigid designators). Because the Marcus systems were merely proof-theoretic, this semantic explanation was unavailable.

The proof of the necessity of identity offered in Marcus (1947) employs a pair of 2-place identity predicates I_m and I. The former is called "material identity" (relating all and only pairs consisting of an object and itself) and one is called "strict identity" (which necessarily relates all and only such pairs). Theorem 2.33 of Marcus (1947) establishes the *strict*—i.e., *necessary equivalence*—of the two. Marcus establishes this proof-theoretically using three decidedly nontrivial assumptions: (i) that $\Box\Box$ *S is strictly equivalent to* $\Box S$, (ii) *the (2nd order) Barcan Formula*, and (iii) the *(2nd order) Converse Barcan Formula.* The version of the Barcan formula in question is that $\Box\exists\Phi(\ldots)$ *strictly (i.e., necessarily) implies* $\exists\Phi\Box(\ldots)$. The converse formula reverses the order.

These formulas arise in her systems due to the interaction of the rule of strong necessitation (which generates the necessitation of any theorem) with a proof theory that allows open formulas (understood as equivalent to the result of adding universal quantifiers binding the variables) to be generated as theorems. Thus, whenever we can derive Φx in Marcus's systems, we can derive $\Box\ \Phi x$, $\forall x\ \Phi x$, $\Box\ \forall x\ \Phi x$, and $\forall x\ \Box\ \Phi x$. However, the question "*Should these formulas be derivable?*" can't be seriously addressed prior to giving model-theoretic semantic systems for interpreting modal logic. Because the Marcus proof-theoretic derivation of the necessity of identity implicitly relies on unexamined semantic elements, it can at most be regarded as suggestive.

Early exponents of modal logic including Marcus, Carnap, and others were, of course, guided by the idea that $\Box$ expresses some notion of necessary truth. Typically, the targeted truths were said to be *analytic—i.e., sentences thought to be true solely in virtue of meaning and capable of being known to be true simply by understanding and reasoning a priori them.* These were often taken to include, but not be limited to, ordinary logical truths. Unfortunately, which sentences fall under this heading was far from clear. In addition, this rough-and-ready characterization leaves the interpretation of formulas in which the box operates not on (closed) sentences but on formulas containing free variables—like the Barcan and converse Barcan formulas and the theorem expressing the necessity of identity—unexplained.

To achieve a modicum of understanding, I will begin with the idea that sentences of modal languages are evaluated for truth or falsity at indices called 'world-states', which are candidates for being actual (i.e., as being a state the universe is really in). Sentences and formulas relative to assignments of values to variables are evaluated at world-states at which singular terms denote objects of which predicates are true or false. We may begin by keeping an open mind about whether the domains of quantification can differ from world-state to world-state. If we provisionally assume that the domains of objects at different world-states are always the same, making the range of quantifiers used at any world-state the same as their range at other world-states, then the two famous Barcan formulas will be correct, and the proof that she offers of a version of the necessity of identity will be sound. If, on the other hand, different possible objects can exist at different world-states, the offered proof won't go through. Whether or not we should want it to go through depends on what we take modal sentences—containing the necessity and possibility operators—to say.

3.2. Quine: Quantifying into Modal Contexts

The problem of interpreting quantified modal logic was taken up by Willard van Orman Quine in two classic papers, "Notes on Existence and Necessity" in 1943 and "The Problem of Interpreting Modal Logic" in 1947. Always skeptical of apriority and necessity, he was particularly opposed to quantified modal logic, as espoused by Marcus and Carnap. Since both seemed to equate necessity with analyticity, Quine was willing (for the sake of argument) to take analyticity for granted—defining it as a sentence that can be turned into a logical truth by replacing synonyms with synonyms. Interpreting necessity as analyticity and taking S to be possible if and only if its negation isn't analytic, he could make provisional sense of claims like (1) and (2), as instances of what he called *the first-grade of modal involvement.*

1. '9 is an odd number' is necessary.
2. 'The number of planets is even' is possible.

His second grade was illustrated by (3) and (4).

3. It is necessary that 9 is an odd number.
4. It is possible that the number of planets is even.

The strategy was to reduce the second grade of modal involvement to the first. Thus, Quine tried to reduce the truth conditions of (3) and (4) to those of (1) and (2). When modal operators are iterated, Quine assigned sentences to a hierarchy, depending on how many modal operators are embedded under such operators. Each level is governed by a definition of logical truth and analyticity, with the truth conditions of ⌜It is necessary that S⌝—which is of level n+1 when S is of level n—being given in terms of the definitions of logical truth and analyticity at level n.[18]

At the third grade of modal involvement '□' and '◊' are operators (expressing necessity and possibility) that can be prefixed to open formulas, allowing quantifying into the scope of the operator, as in (5) and (6).

18. Quine (1947), section 2.

5. ∃x (x is the number of planets in the solar system & □x is odd).
6. ∃x (x is an even number & ◊x is the number of planets in the solar system).

Quine rightly points out that understanding both analyticity and objectual quantification doesn't guarantee that we can assign intelligible truth conditions to sentences like these. If necessity is analyticity, it is a property of *sentences.* To make sense of (5) and (6) we must decide whether an open formula *relative to an assignment of an object to a variable* is a logically true sentence or one that can be turned into a logical truth by replacing synonyms with synonyms. Since open formulas are *not* sentences and variables relative to assignments are *not* terms with meanings or definitions, it is puzzling what the truth conditions of (5) and (6) are supposed to be. It is tempting to think that this was a problem of Quine's own making. But it wasn't. It arose precisely because early modal logicians like Marcus and Carnap did take necessity to be analyticity—i.e., truth in virtue of meaning. What Quine showed was that if quantified modal logic was to progress, then it had to either solve this puzzle or develop nonlinguistic understandings of necessity and possibility.

His next move was to close both escape routes. His most interesting argument purported to show that, the interpretation of necessity aside, quantified modal logic violates fundamental logical and semantic principles, and so must be rejected. In addition to definitions D1–D3, Quine's argument for this depended on A1, which is true, plus A2 and A3, which, unfortunately for him, turned out to be false.[19]

A1. The modal operators '□' and '◊' are referentially opaque.
A2. Occurrences of objectual variables in the scope of referentially opaque operators are not purely referential.
A3. Bindable occurrences of objectual variables must be purely referential.
D1. An occurrence of a term in a formula or sentence S is *purely referential* if and only if what it contributes to the truth or falsity of S (relative to an assignment) is simply what it designates or denotes (relative to the assignment).
D2. A position in a sentence S is *referentially transparent* if and only if for any pair of terms t and t*, the results S(t) and S(t*) of substituting these terms into that position in a sentence will have the same truth values (relative to an appropriate assignment) if and only if ⌈t = t*⌉ is true (relative to that assignment). A position is *referentially opaque* if and only if it is not referentially transparent.
D3. A sentential operator is *referentially transparent* if and only if any referentially transparent position in a sentence remains so when

19. The explication given below is of reasoning in Quine (1943, 1953b). The category of "terms" in D1, D2 includes singular definite descriptions.

the operator is prefixed to the sentence. A sentential operator is *referentially opaque* if and only if it is not referentially transparent.

The idea behind A2 is this: Let 'O' be a referentially opaque operator, let ⌈O F(x)⌉ be a formula in which a variable 'x' occurs free (in position p), and let ⌈O F(t)⌉ and ⌈O F(t*)⌉ be sentences that differ in truth value which arise from substituting distinct terms t and t* designating the same object o for 'x' (at p). There must be such terms if 'O' is referentially opaque. The truth value of ⌈O F(x)⌉ relative to an assignment A of o to 'x' differs from the truth value of *one* these two sentences even though ⌈t = t* = x⌉ is true relative to A. Suppose ⌈O F(t*)⌉ differs in truth value from ⌈O F(x)⌉ (relative to A). Then, Quine concludes, occurrences of t* in the former and 'x' in the latter *both* fail to be purely referential, verifying A2.[20]

As noted in section III of Kaplan (1986) and Kazmi (1987), this argument is fallacious. From the fact that ⌈O F(t*)⌉ differs in truth value from ⌈O F(x)⌉, we can conclude that either the occurrence of t* in the former or the occurrence of 'x' in the latter is not *purely referential*, but we cannot conclude that both aren't, or that the occurrence of 'x' isn't. Furthermore, one can construct opaque operators, as Kaplan does in sections IV, VII, and VIII–XIII, for which occurrences of variables in their scope *are* purely referential, while being coherently bindable from outside by objectual quantifiers. So A2 is false.

Quine (1947) uses different reasoning in attempting to establish A3. He notes that if O is a referentially opaque operator, there will be truths ⌈t = t* & O (S(t) & ~ O S(t*)⌉. If t and t* occupy positions open to objectual quantification, and if existential generalization is universally truth preserving, then ⌈∃x ∃y (x = y & O Sx & ~O Sy⌉ must also be true. Since this violates what Quine calls "the law of the substitutivity of identity for variables" and requires some occurrences of variables to be non-purely referential, he thinks it is impossible. As we will see in a moment, Quine was (understandably) wrong about this. The other flaw was his incorrect assumption that existential generalization is fundamental to objectual quantification. Although it is always truth preserving in certain contexts, it fails to be so in others.

In Kaplan's insightful discussion of Quine, he introduces the notion of the *valuated sentence* associated with F(x) relative to an assignment of object o to 'x'. It is what one gets by substituting *the object o itself* for *'x'* in the syntactic structure F(x). Given this, one can define referentially opaque operators that allow quantifying in while being truth preserving in both ordinary and valuated sentences. For example, we might define an operator O1 that maps an ordinary sentence S onto truth iff Ralph utters or accepts S, while mapping a valuated sentence VS onto truth if and only if he utters or accepts any complete sentence that results from replacing an occurrence of an object o in VS with any occurrence of *any* proper name of o. So understood, occurrences of variables

20. For interpretation of the ⌈ . . . ⌉ notation, see "A Word about Notation" preceding chapter 1.

under O1 are *purely referential*, and the standard "law" (7) of quantification theory is retained.[21]

7. $\forall x,y\ [x = y \supset (O1\ (Fx) \supset O1\ (Fy))]$

Although this argument is sound, it goes beyond what is needed. *Nothing in the nature of quantification* requires (7) to be true. Let a *finely valuated sentence* be just like a valuated sentence except that instead of replacing the variable 'x' with an object o, we replace 'x' with <'x',o>. Now we stipulate that *O2* maps a finely valuated sentence FVS onto truth if and only if Ralph utters or accepts any complete sentence that results from replacing all occurrences of each variable/object pair <'v', o> in FVS with occurrences of a proper name of o, *provided that different occurrences of the same pair are replaced by occurrences of the same name*. Quantification into contexts governed by O2 is as intelligible as quantification into contexts governed by O1, even though (7) fails when O2 replaces O1. So (7) *isn't* really a *law* of quantification, and bindable occurrences of variables need not always be purely referential.

To understand this one must *not* confuse schema (8a) with the indiscernibility principle that may be formulated by (8b) or (8c).[22]

8a. $\forall x,y\ [x = y \supset (S(x) \supset S(y))]$
8b. $\forall x,y$ ($x = y \supset$ every property of x is a property of y)
8c. $\forall x,y\ [x = y \supset \forall P(Px \supset Py)]$

For some opaque operators 'O', instances of (8a) that arise from replacing '$S(x) \supset S(y)$' with '$O(x \neq y) \supset O(y \neq y)$' are false, if Ralph utters or accepts 'Hesperus ≠ Phosphorus' but doesn't utter or accept $\ulcorner n \neq n \urcorner$ for any name designating Venus. This is consistent with the truth of (8b) and (8c), since the property Venus must have iff '$O(x \neq y)$' is true (relative to an assignment A of Venus to 'x', 'y') is *being designated by some pair of names* t_1 *and* t_2 *such that Ralph utters* or accepts $\ulcorner t_1 \neq t_2 \urcorner$, while the property Venus must have iff '$O(x \neq x)$' is true (relative to A) is *being designated by some name t such that Ralph utters or accepts* $\ulcorner t \neq t \urcorner$. In short, the failure of Quine's principle A3 does not, in and of itself, threaten the indiscernibility of identicals.[23]

Nevertheless, the failure of Quine's argument A1–D2—that quantifying into referentially opaque constructions violates fundamental semantic and logical principles—doesn't finally put to rest his worries about the quantified modal logic of his day. To do that, one must make sense of quantifying into modal contexts *when necessity is identified with analyticity*. Quine argues in (1947, 1953) that this is impossible because the truth conditions of sentences of the *third*

21. 'F' is a schematic letter in (10).
22. See Kazmi (1992).
23. Two interesting analyses of propositional attitude verbs that lead to violations of (8a) are Mark Richard (1987) and Kit Fine (2007). These are critically discussed in Soames (1987b, 2012) and chapter 7 of Soames (2002a). See Soames (2015) for a conception of propositions as objects of attitudes like belief that also allows for violations of (8a).

grade of modal involvement can't be specified in terms of the truth conditions of those of the *second grade*.

As he notes, it is natural to appeal to (i) and (ii) in attempting to do so.

(i) $\exists x \ldots x \ldots$ is true ***only if*** $\ldots a \ldots$ is true for some term a.
(ii) $\exists x \ldots x \ldots$ is true ***if*** $\ldots a \ldots$ is true for some term a.

Principle (i) is potentially problematic because there will often be no guarantee that unnamed, or even unnamable, objects (if such there be) might be the only ones underwriting the truth of an existence claim. Principle (ii) is also problematic. Suppose there are two names, 'a' and 'b', such that (9a) and (10a) are both true when R can be any reflexive relation (including identity). Then, by (ii), (9b) and (10b) must also be true. (Remember that in 1947 necessity was standardly equated with analyticity.)

9a. $a = b \ \& \ \Box aRa$
b. $\exists x\ [x = b \ \& \ \Box xRa]$
10a. $b = b \ \& \ \sim\Box bRa$
b. $\exists x\ [x = b \ \& \ \sim\Box xRa]$

Since (9b) and (10b) are contraries, they can't both be true. So, to prevent (ii) from being falsified, one must restrict the terms used to specify the truth conditions of quantified sentences to members of a class T of terms ***coreferential members of which are analytically equivalent***—where analytically equivalent terms are those substitution of which always preserves *analyticity*. Call this restriction *principle (iii)*. If (iii) is observed, (9a) and (10a) can't be jointly true, which will block the erroneous characterization of (9b) and (10b) as jointly true.

However, to adopt (iii) as the means of specifying the truth conditions of third-grade modal sentences in terms of second-grade sentences drastically limits the domain of objects and the class of terms designating them. Since it seems obvious that one can understand both the name 'Hesperus' and the name 'Phosphorus', or the name 'Cicero' and the name 'Tully', without having any basis to conclude that they name the same thing, examples like these seem to show that ordinary proper names of empirically given objects must be excluded. That's not all. As Quine notes in (1947, 1953b), the severity of the needed restrictions would undercut any significant philosophical interest in quantified modal logic. Nor does there seem to be another way of specifying truth conditions of the third grade in terms of those of the second. Thus, Quine was right to this extent: *if necessity is nothing more than analyticity* (and a sentence is analytic iff understanding what it means and reflecting on it is sufficient to come to know that it is true), then quantified modal logic is of little, if any, interest. His error was in taking it for granted, along with most of those against whom he argued, that if there is such a thing as necessity, it must be analyticity.[24]

24. Thanks to Ali Kazmi for useful discussions of this material.

3.3. Smullyan, Fitch, and Church: Response to Quine

Despite the understandable, but ultimately unfounded, skepticism of Quine (1947), one of its virtues was that it attracted responses from three leading philosophers of his day—Arthur Smullyan, Frederic Fitch (who was Ruth Marcus's dissertation advisor), and Alonzo Church, who edited the *Journal of Symbolic Logic*. The responses to Quine in Smullyan (1947, 1948) and Fitch (1949) focused on the difference between proper names and singular definite descriptions (which Quine had ignored). Their idea can be illustrated by contrasting 'Hesperus' and 'Phosphorus', thought of as genuine proper names, with a pair of descriptions, *the planet visible in the evening sky (from certain places at certain times)* and *the planet visible in the morning sky (from other places at other times)*. Assume, for the sake of argument, that the semantic content of a genuine name is its referent so that coreferential names have the same semantic content, making them substitutable in sentences in modal sentences without changing truth value. Assume, by contrast, that the semantic contents of the singular definite descriptions corresponding to the names are, respectively, the property *being unique in being visible in the evening sky (from certain places at certain times)* and the property *being unique in being visible in the morning sky (from certain places at certain times)*. Under these assumptions, they will embed differently than names do under modal operators.

Now think again about Quine's examples (9) and (10), taking the two names to be 'Hesperus' and 'Phosphorus' and the relation R to be identity.

9a*. Hesperus = Phosphorus and it is a necessary truth that Hesperus is Hesperus
b*. ∃x [x = Phosphorus and it is a necessary truth that x = Hesperus]
10a*. Phosphorus = Phosphorus and it's not a necessary truth that Phosphorus = Hesperus
b*. ∃x [x = Phosphorus and it's not a necessary truth that x = Hesperus]

If, as we may assume, *necessarily Hesperus = Hesperus* is true, then (given our assumption that coreferential names have the same semantic content) we will conclude that *necessarily Hesperus is Phosphorus* is also true. Hence (9*a,b) will be true, while (10*a,b) will be unproblematically false, and underivable from (9*a,b), thus undermining Quine's argument. When we substitute corresponding *descriptions* for 'a' and 'b', even (9a*) will fail since it won't be a necessary truth that any planet is visible, either morning or evening.

The common thread in the response of Smullyan and Fitch to Quine is their focus on the need to distinguish proper names, which they thought of as contributing merely their referents to the modal truth conditions of sentences, from descriptions, which contribute properties presumed to be uniquely instantiated, while often picking out different objects at different possible circumstances. Whether or not the descriptions are analyzed as Fregean complex singular terms or are eliminated via Russell's theory of descriptions (which Smullyan and Fitch favored) doesn't materially affect their responses.

The weaknesses of their responses were (i) their silence about what, other than analyticity, necessity might be; and (ii) their failure to note **that merely understanding** coreferential proper names is not sufficient to allow one to recognize that they make the same contributions to the semantic contents of sentences containing them. They were not in a position to see this because their conception of meaning/semantic content, and of a sentence being true in virtue of meaning as one that can be known to be true simply by understanding it, presupposed **that sameness of meaning is cognitively transparent** to competent speakers. The fact that ordinary proper names don't conform to this presupposition was thus at odds with their linguistic notion of necessity.

Alonzo Church was sensitive to these problems in his 1950 review of Fitch's article. Church says that Fitch holds

> (with Smullyan) that two proper names of the same individual must be synonymous. It would seem to the reviewer [Church] that, as ordinarily used, 'the Morning Star' and 'the Evening Star' cannot be taken to be proper names in this sense; for it is possible to understand the meaning of both phrases without knowing that the Morning Star and the Evening Star are the same planet. Indeed, for the same reasons, ***it is hard to find any clear example of a proper name in this sense***. (my emphasis)[25]

Although Church had a good point, which threatened the attempt by Smullyan and Fitch to disarm Quine's attack on quantified modal logic, it took two more decades before the beginning of a systematic response to Church began to be constructed, during which time the problem of interpreting quantified modal logic remained pressing.

For now, one may notice that Church's point could have been made with examples like 'Hesperus' and 'Phosphorus', which have the ordinary syntax of proper names, without appealing directly to special cases like 'the Evening Star' and 'the Morning Star', which might appear to be abbreviated descriptions. Suppose (i) that the semantic contents of 'Hesperus' and 'Phosphorus' (like that of ordinary names in general) are their referents, (ii) that substitution of nonlogical constants with the same semantic contents in sentences doesn't change semantic content, and (iii) that modal operators on sentences operate on the semantic contents of those sentences. Given all this, substitution of one of the two names for the other under a modal operator will preserve semantic content, and hence, truth value. This may be so, even if *understanding* the two names typically involves associating them with nontrivially different information. Generally, anyone who uses these two names is expected to know that uses of one typically presuppose that it stands for something visible in the evening while uses of the other presuppose that it stands for something visible in the morning. In ordinary life, one who mixes this up would be taken to *misunderstand* the names.

25. Church (1950), p. 63.

With this in mind, consider A's use of sentence (11), addressing B, each presupposing they *understand* the names.

11. Hesperus is Phosphorus.

Suppose that in so doing A asserts that o = o, where o = Venus. Nevertheless, this would standardly not be taken to be all that A asserts. Taking it for granted that A presupposes that they both understand the names, B reasons that A knows that A will be taken to be committed to the claim that *the unique object that is both Hesperus and visible in the evening is the unique object that is both Phosphorus and visible in the morning*. Knowing that A expects B to so reason, B *correctly* concludes that A did *assert* the descriptively enriched proposition, which is contingent, even though A also asserted the necessary truth which, by assumption, is the semantic content of (11).

On this analysis, the extra representational content carried by A's utterance arises from the semantic content of the sentence A assertively uttered, the presupposition that A and B understand the names, and the information that comes with that. What does understanding 'Hesperus' and 'Phosphorus' require? Presumably, it requires knowing that most agents who use them take, and expect others to take, 'Hesperus' to stand for something seen in the evening and 'Phosphorus' to stand for something seen in the morning. Presupposing that A and B understand the names (in this sense), both add descriptive content to what A asserts by uttering (11).

What if, at some later point, A were to make the further comment (12)?

12. Necessarily Hesperus is Phosphorus.

Since taking the names to refer to things *actually* seen at certain times tells one *nothing* about when they are seen at other *possible* world-states, speaker-hearers don't descriptively enrich *under the modal operator*, making the proposition asserted by A's use of (12) true, even though the asserted content of A's original use of (11) is contingent.[26]

These complications involving semantic content, presupposition, understanding, and assertion are the tip of an iceberg that was invisible to the pioneers of modal logic after World War II. Thus, it is not surprising that initial investigators, who were trying to forge the semantic and logical ideas needed by sophisticated modal systems, found themselves pulled in different directions.

26. See Soames (2015), pp. 84–88. This distinction between semantic content, conditions of understanding, and how they contribute to asserted content can be used to improve on the account of "partially descriptive names" given in chapter 5 of Soames (2002a)—e.g., names *like Professor Saul Kripke, Mr. Terry Thomas, Miss Ruth Barcan, Mrs. Ruth Barcan Marcus, Princeton University, New York City, Park Avenue, Woodland Park, Mount Rainier, Lake Crescent, Puget Sound, The Columbia River, Whidbey Island, The Empire State Building, The Eiffel Tower, Yankee Stadium, Fort McHenry, Seattle Washington, Princeton Township, Princeton Borough*, etc.

INDEX OF NAMES

SUBJECT INDEX