ON THE LOGIC OF ATTRIBUTIONS OF SELF-KNOWLEDGE TO OTHERS

My immediate purposes here are: (a) to discuss some of the peculiar logical features of statements attributing self-knowledge to others, and (b) to use those features as criteria for an examination of the existing theories on the logical form of such statements. My remote purposes are: (c) to add to and correct, as well as to provide an introduction to, the very unreadable study of the logic of self-knowledge presented by me in two earlier papers, "'He'"¹ and "Indicators and Quasi-indicators,"² and (d) to locate the real issue between the competing theories on the logic of self-knowledge.

We shall use single quotes around a sentence (or sentence form) to produce a name of the sentence (or sentence form) in question, and double quotes around it to produce a name of a statement (or statement matrix) formulated with the sentence (or sentence form). We shall not discuss how the statements so named are determined, yet we shall assume that a given sentence within double quotes names, throughout the ensuing discussion, one and the same statement. We shall employ a numeral within parentheses to refer either to a sentence that follows the numeral on its first appearance or to the statement (or statement form) formulated by that sentence, leaving it to the context to specify which one is meant. Following Nelson Goodman, we shall refer to demonstrative and personal pronouns, adjectives, and adverbs as indicators, when and only when they are used to make a demonstrative reference.

1. ATTRIBUTIONS OF SELF-KNOWLEDGE

1. The first-person quasi-indicator. Here we shall not discuss second-person statements that attribute self-knowledge to others, like

(1) You know that you are brilliant.

The study of (1) requires an investigation of indexical reference in oratio obliqua, i.e., the reference that is part and parcel of a statement, like "you are brilliant" in (1), which lies in the scope of an attribution of knowledge (belief, conjecture, assertion, etc.) to someone. Such an investigation has been carried out in the above-noted two papers. But two principles about indexical reference in oratio obliqua should be noted for future use. Consider (1) again: the second occurrence of the indicator 'you' in (1) formulates a second-person reference made not by the person to whom knowledge is attributed by (1), but by the speaker. Likewise, in

(2) Anthony knows that I (you, this man) am (is, are) a millionaire.

the indicator 'I' ('you', 'this man') expresses an indexical reference made, normally, not by Anthony, but by the asserter of (2), whoever he may be. Now, in (1) the second indicator 'you' refers to the man to whom knowledge is attributed by (1) and, furthermore, attributes to that man a first-person reference; in (2), on the other hand, the indicator 'I' ('you', 'this man') neither refers to Anthony nor attributes to Anthony any specific way of referring to the person to whom the asserter of (2) refers indexically. It is easy to see that, because (2) is not a statement attributing self-knowledge to anyone, the indexical reference made by the asserter of (2) leaves it unrevealed how Anthony refers to the person to whom the speaker refers indexically. Obviously, similar considerations apply to the other indicators 'that', 'here', 'there', and 'now'; so in general we have:

(I.1) An indexical reference in oratio obliqua is made by the speaker and is presented by the oratio obliqua as being made by him and not as being made by the person to whom knowledge (belief, conjecture, assertion, etc.) is attributed; an indicator always has the largest scope in a sentence containing it.

and

(I.2) An indexical reference R in oratio obliqua, that is not part of an attribution of self-knowledge (self-belief, self-conjecture, etc.) leaves it by itself wholly unspecified how the person to whom knowledge (belief, etc.) is attributed refers to the person or object to whom R is made.

In the sequel we shall be concerned almost exclusively with third-person statements that ascribe self-knowledge to others, like

(3) The Editor of Soul knows that he (himself) is a millionaire.

and

(4) The Editor of *Soul* knows that Mary knows that her niece knows that he (himself) is a millionaire.

In these cases the attribution of self-knowledge is made by means of the third-person pronoun 'he (himself)' to be abbreviated 'he*', which has here the following characteristics: (i) it does not express an indexical reference made by the speaker; (ii) it appears in *oratio obliqua*; (iii) it has an antecedent, namely 'the Editor of *Soul*', to which it refers back; (iv) its antecedent is outside the *oratio obliqua* containing 'he*'; (v) 'he*' is used to attribute, so to speak, implicit indexical references to the Editor of *Soul*; that is, if the Editor were to assert what, according to (3) and (4), he knows, he would use the indicator 'I' where we, uttering (3) and (4), have used 'he*': he would assert, respectively,

(3a) I am a millionaire.

and

(4a) Mary knows that her niece knows that I am a millionaire.

Because of (i), by (I.1), and because of (v), by (I.2), 'he*' is not an indicator. I have called it 2 the quasi-indicator corresponding to the first person.

Thus, the problem of the logic of third-person attributions of self-knowledge is simply the problem of the logic of the quasi-indicator 'he*'. More specifically, the question of the logical form of such attributions is the question of the logical form of statements formulated with sentences containing occurrences of 'he*', and this is the question whether 'he*' is analyzable in terms of indicators, proper names, descriptions, or quantification. This question of analysis is precisely the question we are primarily concerned with here.

2. 'He*' and individual constants. Logicians speak of *individual constants* to refer to both proper names and definite descriptions. Here we shall use the term to refer to indicators as well. The first thing we want to examine is the relation between the quasi-indicator 'he*' and individual constants. To start with, note that statement (3) above neither entails nor is entailed by

(3b) The Editor of *Soul* knows that the Editor of *Soul* is a millionaire.

Palpably, the Editor of *Soul* may just have been appointed to the editorship in an as yet unpublicized secret meeting of the journal's editorial board, so that our Editor does not know that he* is the Editor of *Soul*. Yet he may, on the one hand, know both that he* has been poor all along and that the Editor, whoever he may be at
that time, has been bequeathed two million dollars, so that (3) is false while (3b) is true. On the other hand, our Editor may both know that he* has always had millions of pounds and believe that the Editor of *Soul* is old penniless Doctor Jones, so that (3) is true while (3b) is false.

More generally, if ‘a’ stands for a definite description of the Editor of *Soul* that predicates a nonanalytic property 𝜙-ness to the Editor, then (3) fails to entail the corresponding statement of the form:

(3c) The Editor of *Soul* knows that a is a millionaire.

for if our Editor does not know that he* is 𝜙, then (3) may be true even if the Editor believes that a is not a millionaire.

If ‘a’ stands for a definite description of the Editor of *Soul* not containing occurrences of ‘he*’ that refer to the Editor, then (3) does not entail the corresponding statement of the form (3c), for such description must attribute to the Editor some nonanalytic property that (uniquely) characterizes him.

Now, let ‘a’ stand for a name of the Editor of *Soul*. Clearly, to know that he* is a the Editor will have to know not only that he* is named ‘a’, but also that he has some of the properties predicated to him by means of the third-person descriptions supporting the name ‘a’. Undoubtedly, the Editor may very well either fail to know that ‘a’ is his name or fail to know that one of those supporting descriptions applies to him. Thus, if the Editor possesses one million dollars and believes that he* is not a, it is possible that (3) be true without the corresponding statement of form (3c) being true.

Let ‘a’ stand for the indicator ‘this (man)’. Then (3) fails to entail the corresponding statement:

(3d) The Editor of *Soul* knows that this (man) is a millionaire.

By (I.1) above, the indexical reference made by the occurrence of ‘this (man)’ in (3d) is made by the speaker, who may very well be other than the Editor. If the Editor does not know (as may be the case if, e.g., he is not paying attention) that he* is the man being demonstratively referred to by the asserter of (3d), (3) may be true without (3d) being true, and vice versa. Obviously, similar considerations apply to the other indicators.

Summing up, then:

(E.1) There is no individual constant ‘a’ containing no occurrence of the quasi-indicator ‘he*’ such that: “The Editor of *Soul* knows that he* is a millionaire” either (i) entails or (ii) is entailed by its corresponding statement of the form “The Editor of *Soul* knows that a is a millionaire.”
Of course, if (3), "The Editor of Soul knows that he* is a millionaire," is true, then the proposition known by the Editor is true. This proposition is (3a), namely, the proposition he would assert by saying "I am a millionaire." Clearly, (3) entails (3a); so, even though there is an individual constant, namely the first-person pronoun, that refers to the person whom according to (3) the Editor knows something about, that constant cannot be used to formulate the proposition of form (3c) above; for "The Editor of Soul knows that I am a millionaire" is an entirely different proposition containing first-person references by the speaker, who need not be the Editor of Soul. Thus,

(E.2) There is no individual constant 'a' such that: both "The Editor of Soul = a" is true and "The Editor of Soul knows that he* is a millionaire" either (i) entails or (ii) is entailed by "Both the Editor of Soul knows that a is a millionaire, and a is a millionaire."

Let us go back to (4) "The Editor of Soul knows that Mary knows that her niece knows that he* is a millionaire." If (4) is true, the Editor can truly assert (4a), "Mary knows that her niece knows that I am a millionaire." But, by (I.2) above, the presence of the indicator 'I' in (4a) hides the way in which Mary's niece refers to the Editor. However, since she cannot refer to him in the first person, if she were to express what according to (4) she knows, Mary's niece must have some name or third-person description of the Editor that she can use to refer to him. Now in its turn, (4a) entails the statement that the Editor can express by saying:

(4b) Mary knows that I am a millionaire.

Again, Mary cannot refer to the Editor in the first-person way, but must have some name or description that she can use to refer to him. Yet it does not follow from (4) that Mary and her niece refer to the Editor in exactly the same way, or even that they have a common way of referring to him. Thus,

(E.3) For no individual constant 'a' referring to the Editor of Soul does (4) entail the corresponding statement of the form "Mary knows that a is a millionaire, and Mary's niece knows that a is a millionaire."

Indeed, the Editor of Soul may know that (4b) is true because his friend Gaskon, who is known by him to be incapable of telling a lie, said to him: "I know that Mary knows that you are a millionaire." This statement contains a second-person reference by Gaskon, but leaves it unspecified how Mary refers to the Editor, and the Editor may very well fail both to know Mary and to know how she refers to
him. Conversely, suppose that ‘a’ is a purely third-person expression with no occurrences of ‘he*’ such that our Editor knows that Mary knows that a is a millionaire. Let a = the Editor of Soul, but suppose that the Editor does not know that he* is a. Clearly, all of this is compatible with the falsity of “The Editor knows that Mary knows that he* is a millionaire.” Thus, we have the stronger property:

(E.4) There is no individual constant ‘a’ which does not contain occurrences of ‘he*’ and which refers to the Editor of Soul such that: (4) either (i) entails or (ii) is entailed by the corresponding statement of the form “The Editor of Soul knows that Mary knows that a is a millionaire.”

3. Degrees of ‘he*’. In (3) the quasi-indicator ‘he*’ has its antecedent across just one occurrence of ‘knows that’, and that antecedent is the subject of that occurrence of ‘knows’. On the other hand, in (4), ‘he*’ occurs in the scope of ‘her niece knows’ which in turn occurs in the scope of ‘Mary knows’, which finally occurs in the scope of ‘the Editor of Soul knows’, where neither Mary nor her niece is attributed first-person reference by ‘he*’. We shall say that in (4) the subjects ‘Mary’ and ‘her niece’, sandwiched between ‘he*’ and its antecedent ‘the Editor of Soul’ in a chain of inclusive scopes, are pseudo-antecedents of ‘he*’. We shall also say that the occurrence of ‘he*’ in (3) is of degree 1, whereas the occurrence of ‘he*’ in (4) is of degree 3, the idea being roughly that the degree is the number of pseudo-antecedents plus one (antecedent).

Consider now

(5) The Editor of Soul knows that he* knows that Mary knows that he* is a millionaire.

In (5) the two occurrences of ‘he*’ have ‘the Editor of Soul’ as their antecedent, but only the first occurrence from the left refers back to it immediately; the second occurrence of ‘he*’ refers back to the preceding occurrence of ‘he*’ and only through this occurrence does it refer to ‘the Editor of Soul’. Thus, while ‘the Editor of Soul’ is the remote antecedent of both occurrences of ‘he*’, the second occurrence of ‘he*’ has the first as its immediate antecedent. Only for the first occurrence do the remote and immediate antecedents coincide. We shall take the distance of an occurrence of ‘he*’ to its immediate antecedent in a sentence S as determining the degree of that occurrence of ‘he*’ in S. Thus, we shall say that the first occurrence of ‘he*’ in (5) is of degree 1, and so is the second, because there is no pseudo-antecedent in between it and its immediate antecedent.
Still intuitively, consider

(6) The Editor of *Soul* knows that Gaskon knows that he* [Gaskon] knows that he* [the Editor] is a millionaire.

Clearly, the first occurrence of 'he*' referring immediately, with no pseudo-antecedents in between, to 'Gaskon', is of degree 1. Now, the second occurrence of 'he*' referring back to 'the Editor of *Soul*', is of degree 3, because 'Gaskon' and the 'he*' referring back to it are two different pseudo-antecedents of the second 'he*'. This is exactly the view I took in "Indicators and Quasi-indicators" and later in "On the Logic of Self-Knowledge." 4

In general, if an occurrence \( O \) of 'he*' in a sentence \( S \) lies in the scope of an expression of the form 'X knows that' which in its turn lies in the scope of an expression of the form 'Y knows that', where \( Y \) is the immediate antecedent of \( O \), then the expression for which 'X' stands is a *pseudo-antecedent* of \( O \). Let \( m \) be the number of pseudo-antecedents of \( O \) in \( S \). Then the degree of \( O \) in \( S \) is \( m + 1 \). Here \( m \) is exactly the number of cases in which the occurrence \( O \) of 'he*' in \( S \) appears subordinated to an expression \( e \) represented by 'X' such that \( O \) does not attribute a first-person reference to whom \( e \) refers qua being referred to by \( e \): \( m \) represents the number of possibly different ways that, according to \( S \), the person referred to by \( O \) is referred to.

(E.5) No occurrence \( O \) of degree higher than 1 of the quasi-indicator 'he*' in a sentence \( S \) attributes a first-person reference to its pseudo-antecedents in \( S \); furthermore, \( O \) leaves it unspecified how the persons referred to by its pseudo-antecedents in \( S \) refer to the person \( O \) refers to.

4. *A theoretical notation.* Given the preceding distinctions, we are in a position to provide a theoretical notation that can eliminate the multiple ambiguities that can result from having one and the same noise or string of marks 'he (himself)' represent not only different degrees of the quasi-indicator 'he*', but also references to different antecedents and even different persons. We simply supplant each occurrence \( O \) of 'he*' with an expression of the form ' \((a)k\)' , where 'a' stands for the remote antecedent of \( O \) and ' \( k \) ' for the degree of \( O \). Thus, (6) above can be unambiguously formulated as:

(6a) The Editor of *Soul* knows that Gaskon knows that (Gaskon)i knows that (the Editor of *Soul*)2 is a millionaire.

To illustrate the power of this notation to extirpate ambiguities,

---

consider a sentence with the quasi-indexical cross-references indicated by the arrows:

\[(7) A \text{ knows that } A \text{ knows that } he^* \text{ knows that } B \text{ knows }
\]
\[
\text{that } he^* \text{ knows that } he^* \text{ knows that } he^* \text{ is } R \text{ to him}.)
\]

This somewhat complicated pattern of arrows is perspicuously captured by our notation thus:

\[(7a) A \text{ knows that } A \text{ knows that } (A)_{1} \text{ knows that } B \text{ knows that } (A)_{1} \text{ knows that } (B)_{8} \text{ is } R \text{ to } (A)_{1}.
\]

5. 'He*' and 'knows who'. A sentence like 'Peter knows who his neighbor is' really formulates different statements, not only because of the multiplicity of persons named 'Peter', but also because of the different criteria involved in identifying a certain person, i.e., knowing who a certain person is. The criteria vary from case to case depending on purposes and circumstances. What we must do in discussing the logic of 'knows who' is to assume that some set of criteria has been chosen and that it remains fixed throughout an investigation of entailments. A valid entailment should hold for any set of criteria kept fixed.

Consider, for instance, the case of a man, to be called "Quintus," who is brought unconscious to a military tent, but on gaining consciousness suffers from amnesia, and during the next months becomes a war hero and gets lost in combat and completely forgets the military chapter of his life. Later on Quintus studies all accounts of the war hero and discovers that he ((the hero), not he*) was the only one wounded 100 times. Quintus becomes fascinated by the hero's accomplishments (for some mysterious reason unknown to him) and comes to write the most authoritative biography of the hero. Clearly, for most normal situations, regardless of shifts in the criteria for identifying a person, Quintus knows who the hero was much better than most people, even though Quintus does not know that he* is the war hero. Since the hero is the same person as Quintus and so must know exactly what Quintus knows, we have:

\[(E.6) \text{ The statement "The war hero wounded 100 times knows who the war hero wounded 100 times is (was)" does not entail "The war hero wounded 100 times knows that he* is [tenselessly] the war hero wounded 100 times."}
\]

Given the thorough study that Quintus has made of the feats of the war hero, and given, as we are, that the war hero exists, we also have:
(E.7) The statement "There exists [tenselessly of not, as one pleases] a person known to the war hero wounded 100 times such that the war hero wounded 100 times knows that that person is identical with the war hero wounded 100 times" does not entail "The war hero wounded 100 times knows that he* is the war hero wounded 100 times."

Since the war hero exists and is in fact that same man known as an existent by Quintus (= the war hero) and known by him to be identical with the war hero, we also have:

(E.8) The statement "There exists a person known to the war hero wounded 100 times, who is in fact identical with the war hero wounded 100 times and is known by the war hero wounded 100 times to be identical with the war hero wounded 100 times" does not entail "The war hero wounded 100 times knows that he* is the war hero wounded 100 times."

II. THEORIES ABOUT THE LOGIC OF SELF-KNOWLEDGE

1. Unanalyzability theories. Several different theories about the logical form of third-person statements ascribing self-knowledge to others have been proposed. Jaakko Hintikka has proposed one theory whose fundamental claim is, to put it in my terminology, that the quasi-indicator 'he*' is fully analyzable in terms of one existential quantifier. I have proposed two theories that are variations on the one claim embodied in the following two theses:

T1. The occurrences of 'he*' of degree 1 are unanalyzable; they constitute a peculiar and irreducible mechanism of reference to persons.

T2. Each occurrence of 'he*' of degree greater than 1 is analyzable in terms of both occurrences of 'he*' of degree 1 and occurrences of one existential quantifier per pseudo-antecedent.

Because of T1 I shall refer to my two theories as unanalyzability theories. The first theory, theory I, proposed in "Indicators and Quasi-indicators," includes the thesis

T3.1. The existential quantifiers introduced in accordance with T2 are quantifiers over properties.

The description (op. cit., p. 100) of the general method of elimination behind T3 is rather complicated, but we can illustrate it by applying it to (4) above:

(4) The Editor of Soul knows that Mary knows that her niece knows that he* is a millionaire.

\footnote{Knowledge and Belief: An Introduction to the Logic of the Two Notions (Ithaca, N.Y.: Cornell, 1962), esp. chapter vi, and more particularly §§6.3 and 6.12. This theory has been modified and elaborated as noted below.}
As we said, the occurrence of ‘he*’ in (4) is of degree 3. We eliminate it in favor of one occurrence of degree 2, obtaining:

(4.1') The Editor of Soul knows that Mary knows that (there is a property $\phi$-ness such that he* = the only one who is $\phi$ and her niece knows that the only one who is $\phi$ is a millionaire).

Then we eliminate this ‘he*’ of degree 2, producing:

(4.1") The Editor of Soul knows that (there is a property $\psi$-ness such that both he* = the only one who is $\psi$ and Mary knows that (there is a property $\phi$-ness such that both the only one who is $\psi$ = the only one who is $\phi$ and her niece knows that the only one who is $\phi$ is a millionaire)).

My unanalyzability theory II differs from theory I in having, instead of T3.1:

T8.1. The existential quantifiers introduced in accordance with T2 range over persons.

Since theory II was put, in “On the Logic of Self-knowledge,” as a development for incorporation into Hintikka’s system of epistemic logic, T8.1 was meant to require, besides, that the persons over whom the quantifiers range be known to the persons to whom knowledge is attributed. The method of analysis (given on p. 206) is parallel to the method behind T3.1, and it can be illustrated with the final analysis of (4):

(4.1") The Editor of Soul knows that \[ (3 y) (he^* = y \text{ and Mary knows that } (3 x) (x = y \text{ and her niece knows that } x \text{ is a millionaire})) \].

Obviously, there is a third unanalyzability theory, namely, one according to which all degrees of ‘he*’ are unanalyzable. This may be further developed by preserving some of the entailments provided by T3.1 and T3.11. It is a possibility that deserves some attention. I myself have some inclination to adopt it.

In favor of the unanalyzability thesis T1, there is the evidence of the type illustrated by (E.1) and (E.2). Unanalyzability is not the only way to satisfy (E.1) and (E.2), but it certainly is an adequate way. In “‘He’” and in “Indicators and Quasi-indicators” there are other pieces of evidence, and in the latter there is a general intricate argument for T1, which still seems to me to be valid; it depends on properties of indicators, which we have ignored here for the most part.

Granting the analyzability of degrees of ‘he*’ greater than 1, there is in favor of thesis T2 the powerful evidence of the type illustrated by (E.3)–(E.5). Whereas (E.5) unmistakably shows that there is a sort of misplacement with an occurrence of ‘he*’ in the scope of a pseudo-
antecedent, (E.3) and (E.4) very firmly require that a different way of referring to the antecedent of an occurrence of 'he*' be available to each person referred to by a pseudo-antecedent.

I have not reached a definite decision about T3. Even though in "The Logic of Self-knowledge" I expressed my preference for T3.1, I also have some inclination to prefer T3.11, especially in its pure form, i.e., without the qualification introduced in it in order to adjoin it to Hintikka's calculus. My reason for preferring that the qualification be deleted will be apparent from the ensuing discussion.

2. Analyzability theories. Hintikka's analyzability theory is formulated within his formal system of epistemic logic, and we must say something about that system. Take a first-order functional calculus with individual constants and introduce into it an operator 'Ka', which is to be read "a knows." If 'p' stands for a complete sentence, i.e., with no free (bindable) variables, then 'Ka(p)' is to be read as "a knows that p" (where the double quotes function as something like Quine's corners and so we must suppose that after 'that' we have, without quotes, the sentence 'p' represents). If 'fx' stands for a sentential function, having, say, the free variable 'x' ranging over persons, then 'Ka(fx)' is to be read in either one of two ways, much better conveyed with an example:

(A) '(3x)Ka(fx)' is to be read as "a knows who is f'.

(B) '(3x)Ka(fx)' is to be read as "There is a person known to a whom a knows to be f."

Hintikka introduces a good number of very interesting rules of entailment governing formulas with the operator 'Ka'. Within that background of notation and rules his theory can be stated very simply: Let 'a knows that (. . . he* . . .)' contain the one occurrence, as deployed, of 'he*' that has as its immediate antecedent the deployed occurrence of 'a'. Then we have:

(AT.ii) "a knows that . . . he* . . ." is fully analyzable as "'(3y)(y = a & K_a( . . . y . . . ))'."

This is both a simple and a beautiful theory. (AT.ii) was put forward by Hintikka in Knowledge and Belief, of course, not quite in the form above, since before my 1966 paper in Ratio the idea of the quasi-indicator 'he*' with all its complexities of degrees and kinds of antecedents had not been discussed at all. Hintikka was, aside from Geach, who had studied one important aspect of it, the first

to consider seriously the formal logic of self-knowledge; in *Knowledge and Belief* he discussed only simple examples involving one occurrence of what I later called the first degree of ‘he*’. On July 14, 1965, Geach wrote me that he had discussed my article “‘He’” with Hintikka. Evidence of that discussion appears in what Hintikka later called “Geach’s ingenious puzzle,” which Geach had found very exciting in “‘He’.” Nevertheless, Hintikka had not seen this paper when he wrote “‘Knowing Oneself’”; at the end of his discussion of “Geach’s puzzle” Hintikka examined his only example with two occurrences of ‘he*’. He analyzed each occurrence in terms of one existential quantifier by means of what seems an immediate application of (AT.ii). But in fairness to Hintikka it must be said that he does not formulate a general principle such as (AT.ii), being content to illustrate the power of his notation with simple examples.

In my review of *Knowledge and Belief* I pointed out that the statement “There exists a person such that a knows that that person is identical with a but a does not know that he himself is that person” is not self-contradictory, but either it cannot be symbolized in Hintikka’s calculus (because of interpretations (A) and (B) of the free bindable variables in the scope of ‘Ka’) or it has a self-contradictory symbolization. I was then making use of something like (E.7) above. Hintikka replied in “‘Knowing Oneself’,” defending his analysis of ‘he*’, but rejecting some of his rules. Unaware of Hintikka’s reply, I repeated my criticism in an article in *Noûs* (see fn 4), showed anew some of the complexity of ‘he*’ and proposed analyzability theory ii. Hintikka replied, also in *Noûs*, confirming his rejection of those rules. He also added the new rule (C.ind¬).

I want to discuss (AT.ii) in two stages. First I want to argue that the schema of analysis included in (AT.ii) is valid at most just for ‘he*’s of degree 1 and that it has to be supplemented with theses T2 and T3.ii and something like the method described in my *Noûs* paper. Then I want to make a case for T1.

Naturally, among the rules of Hintikka’s calculus are several rules of existential instantiation. One of these rules, after his revisions (op. cit., p. 37), is this:

(C.E”). If “∃(x)(x = a & KaKbKc(x is f))” is in a model set μ, then, for at least one constant ‘h’, “h = a & KcKbKc(h is f)” and “∃(x)(x = h & KcKbKc(x = h))” are both in μ.

A model set $\mu$ is interpreted in Hintikka's calculus as an epistemic possible world for certain person(s). That is, a model set $\mu$ represents a possible complete state of knowledge of a person in question. The rules characterizing the membership of a model set $\mu$ characterize the structure of that person's knowledge. Clearly, a model set is by definition a consistent set. Now, Hintikka defines indefensibility (Knowledge and Belief, p. 40), which is his counterpart to inconsistency for the case of epistemic propositions, as follows: a formula $\phi$ is indefensible if and only if it cannot be embedded in a model set $\mu$. Consider a rule of the form "If $\phi$ is in $\mu$, then $\phi'$ is in $\mu"$ characterizing model sets. Any model set containing $\phi$ contains, therefore, $\phi'$. Hence, a set containing both $\phi$ and $\neg \phi'$ is not a model set. Hence, a rule of the preceding form is tantamount to a principle of entailment of the form "$\phi$ entails $\phi'". Thus, \((C.E)\) seems to yield:

\[(C.E'). \quad \langle \exists x \rangle (x = a \& K_x K_x K_x (x is f)) \text{ entails, for some individual constant 'h', the corresponding statement of the form "h = a \& K_x K_x K_x K_x (h is f)."
}\]

Note that \((C.E')\) does not specify the individual constant 'h'.

The symbolization of "The Editor of Soul knows that Mary knows that her niece knows that he* is a millionaire" is, in accordance with \((A.T.ii)\), as illustrated by a formula of Hintikka's (Nou's, p. 50), "\((\exists x)(x = the Editor and K_{editor} K_{Mary} K_{her niece} (x is a millionaire))." To this \((C.E')\) applies, yielding, for some individual constant 'h' referring to the Editor of Soul, the entailed statement that goes back into English as "The Editor of Soul knows that Mary knows that her niece knows that h is a millionaire." This violates (E.4) by violating part (i). By a rule of transitivity of knowledge,\(^{10}\) from the latter statement Hintikka's system allows, correctly, the derivation of the statements "The Editor of Soul knows that h is a millionaire," "Mary knows that h is a millionaire," and "Mary's niece knows that h is a millionaire." Hence, \((E.3)\) is also violated. Clearly, one over-all existential quantifier for the whole of \((4)\) cannot do.

A first reply to the preceding argument is this. It might be aduced that Hintikka's system does not really have individual constants in the sense discussed in connection with \((E.1)-(E.4)\), that his quantifiers do not have the substitution interpretation, and that his variables of instantiation are purely algorithmic devices to figure

\(^{10}\) This rule is, roughly, "if a knows that b knows that p, a knows that p." But this formulation won't do if the sentence 'p' stands for contains quasi-indicators. This is discussed in my "Omniscience and Indexical Reference," this JOURNAL, Lxiv, 7 (April 13, 1967): 207 ff. This paper discusses a theological problem involving indexical and quasi-indexical reference raised by Norman Kretzmann in "Omniscience and Immutability," this JOURNAL, Lxiii, 14 (July 14, 1966): 409-421.
out entailments. It is doubtful that Hintikka would adopt this reply. At any rate, whether Hintikka’s variables are purely algorithmic devices or not, (C.E")' implies, incorrectly, that Mary, her niece, and the Editor of *Soul* share one way of referring to the Editor, even if we cannot tell which one it is. To put it somewhat differently, (C.E") and (AT.ii) make (4) imply that there is one and the same statement or proposition, so far unspecified, which those three persons have as common object of knowledge—and this is an implication that cannot be extracted from (4) by purely logical means.

A second reply to the above argument is this: by deriving (C.E") from Hintikka’s rule (C.E)" we have misinterpreted the sense of (C.E)"; this is the only rule of the form “If $\phi$ is in $\mu$, then $\phi'$ is in $\mu$” that must be construed not as “$\phi$ entails $\phi'$,” but the other way around; surely one can know a proposition of the form “$(\exists x)(x \text{ is } f)$” without knowing that John, or Smith, or . . . is $f$; thus, from (C.E") we should derive the converse of (C.E’); hence, datum (E.3) part (i) above is irrelevant to the analyzability theory (AT.ii). Hintikka has endorsed this reply in correspondence.

This interpretation of (C.E") does not, however, save (AT.ii). Now (AT.ii) fails to satisfy (E.4) by failing to satisfy part (ii) of (E.4). But one other comment is in order. If (C.E") is not to yield (C.E’), then (C.E") becomes a rather mysterious rule within an extremely illuminating semantical approach to the logic of knowledge. (C.E") runs, then, counter to the grain of the rules characterizing the membership of model sets, which rules yield direct entailment principles. Yet (C.E") cannot be like the other rules. Undoubtedly, in a pure calculus of quantification, which is extensional, we have the rule “If “$(\exists x)(x \text{ is } f)$” is in $\mu$, then for some constant ‘$h$’ “$h \text{ is } f$” is in $\mu$”; for we can always invent a name for, or produce a unique description of, the object which is $f$ and which we want to discuss in some context. Thus, in pure quantification we must have the analogue of (C.E’): “$(\exists x)(x \text{ is } f)$” implies, for some constant ‘$h$’, “$h \text{ is } f$.” But in the case of epistemic logic, we cannot invent names or descriptions to instantiate ‘$(\exists x)$’. For instance, we cannot invent a name for the object that satisfies “$(\exists x)(K_a(x \text{ is } f))$”, and put “$K_a(h \text{ is } f)$”; for then we would be claiming that $a$ knows that object under the name ‘$h$’, which would be outrageous. In short, existential instantiation needs a special treatment in epistemic logic.12

11 Nevertheless, he disclaims the substitution interpretation of the quantifiers in *Noûs*, op. cit., p. 38.

I conclude, therefore, that (AT.ii) must yield to another theory. The most natural suggestion in line with Hintikka's analyzability idea is one that adopts thesis T2 together with something like T3 and the claim that one existential quantifier is required for the analysis of occurrences of 'he*' of degree 1. But even though Hintikka simply attaches his two interpretations (A) and (B) to free (but bindable) variables inside the scope of 'K', it is not clear that they should be mixed without more ado. Furthermore, the objection raised above against (AT.ii) depends on (C.E''), which is independent of Hintikka's interpretations (A) and (B). (C.E'') requires merely that, when we apply existential instantiation, we replace the quantified variable with the instantional variable (or constant) throughout; and this requirement is one that general quantificational logic imposes, before the introduction of epistemic operators. Thus, we should discuss three replacements for (AT.ii) which are prima facie different from one another, all having in common T2 and something like T3.ii, but differing in the theses replacing my T1. To formulate these theses, let 'he*_{a1}' be an occurrence of 'he*' of degree 1 having the expression 'a' deployed in the formula as its immediate antecedent. Then we have:

T1'.0. "a knows that he*_{a1} . . ." is fully analyzable as "((\exists x)(x = a & K_a (\ldots x \ldots)))", where 'K_a(\ldots x \ldots)' is simply read as "a knows that \ldots x \ldots," on a par with the reading of 'K_a(p)."

T1'.A. "a knows that he*_{a1} . . ." is fully analyzable as "((\exists x)(x = a & K_a (\ldots x \ldots)))," where this formula is to be interpreted in accordance with (A).

T1'.B. "a knows that he*_{a1} . . ." is fully analyzable as "((\exists x)(x = a & K_a (\ldots x \ldots)))," where this formula is to be interpreted in accordance with (B).

One objection against T1'.0, T1'.A, and T1'.B all at once is that they violate (E.1) and (E.2), since in all three cases the existential instantiation of the analysans yields "h = a & K_a(\ldots h \ldots)" for some individual constant referring to the person 'a' refers to. Again, it might be adduced that in Hintikka's calculus there are no real individual constants, but only algorithmic counters for reckoning entailments. The rejoinder is, again, that, by the rule of existential instantiation, "a knows that he* is \phi" is made to entail that there is a proposition of the form "h is \phi" which both is known by the person referred to as 'a' and can be detached from "a knows that he* is \phi," whereas such entailment does not hold.
Another objection against T1’.A is that on interpretation (A) of Hintikka's calculus his analysis of 'he*' fails to satisfy criterion (E.6). Given (A) and (AT.ii), (9) below is symbolizable as (9a) and (10) below as (10a):

(9) The war hero knows who the war hero is (tenselessly):
   (9a) \( (\exists x)K_{\text{hero}}(x = \text{the hero}) \)

(10) The war hero knows that he* is the hero:
   (10a) \( (\exists x)(x = \text{the hero} \& K_{\text{hero}}(x = \text{the hero})) \)

In Hintikka's original calculus in Knowledge and Belief (9a) and (10a) are demonstrably equivalent. This equivalence clearly violates (E.6). In "Individuals, Possible Worlds, and Epistemic Logic" Hintikka grants that (9) may be true without (10) being true. He points out (p. 51) that nothing he has said in this article requires that (9a) imply (10). He contends that there is a strong sense of ‘know’ in which it does not hold. His new view is that (E.6) is true for the weak sense of ‘know’. But I am not convinced. At any rate, since his new rules do not suffice to show that (9a) is equivalent to (10a), Hintikka owes us a rule that restores this equivalence for the strong sense of ‘know’.

As far as I can see, (E.6) holds for a strong sense of ‘know’ in which (9a) implies (10a). Hintikka engages (in his Nous article) in a protracted discussion of the entailment of (10a) by (9a). He says (p. 53): "Castañeda is presupposing a standard of knowing who on which one can know who h is without being able to locate him in the actual world." I have never presupposed such a thing. My statement above about the war hero's case presupposes exactly the opposite, that the war hero must be able to locate the hero in the actual world in order to know who the hero is (was). Quintus's authoritative biography of the hero places the hero in the actual world.

In the same vein, Hintikka writes (p. 52): "The implication would hold [from (9a) to (10a)] if h’s knowing who h is gave him the ability to say of an actual flesh-and-blood individual that he is h." Precisely this is the implication involved in the claim that Quintus knows who the war hero is, since Quintus by his careful study of the actual war hero’s actual exploits has acquired the ability to say of an actual flesh-and-blood individual that he (that individual, not he himself) is [tenselessly] the war hero. Remember that the existential quantifier is a tenseless operator.

Indeed, because the hero-biographer has written an actual book about an actual flesh-and-blood individual whom he has placed in his actual place in the actual world, T1’.B seems to run into the same
difficulties. On interpretation (B) Hintikka’s system satisfies neither (E.7) nor (E.8).

There is a sense, of course, in which

\[(11)\] The war hero knows (at time t) that he himself was wounded 100 times.

does entail

\[(12)\] \((\exists x)(x = \text{the hero} \& K'_{\text{hero}} (x \text{ was wounded 100 times}))\)

\textit{provided that} (12) is not tied up with individual constants for instantiation. Surely, if (11) is true, then there is a person (namely the person to whom the war hero refers at t by means of the first-person pronoun) who is in fact identical with the war hero and is such that the war hero knows at t that that person was wounded 100 times. But note that here the expression 'the person to whom the war hero refers at t by means of the first-person pronoun' is \textit{not} an individual constant into which we are to instantiate the existential quantifiers of (12), for such an instantiation would be

\[(13)\] The person to whom the war hero refers at t by means of the first-person pronoun = the war hero, and the war hero knows at t that the person to whom the war hero refers at t by means of the first-person pronoun was wounded 100 times.

Clearly, (11) may be true, even if (13) is false, for it may very well be the case if the war hero does not know at t that he \textit{himself} is the war hero. Now, even though (11) entails (12) and still satisfies (E.1) above, (11) is compatible with

\[(14)\] There exists a person who is in fact (known to the war hero and) identical with the war hero, such that the war hero fails to know at t that such a person was wounded 100 times.

Indeed, again without being able to instantiate the existential quantifier of (14), we can name one person that makes (14) true: e.g., the hero’s biographer is an actual person known to the war hero, who happens to be in fact identical with the war hero, and yet is not known by the hero to have been wounded 100 times. But if, according to interpretation (B) of Hintikka’s calculus, we symbolize (11) and (14) we find that their conjunction entails a self-contradiction by the rules of his revised calculus.\(^\text{13}\)

3. \textit{Conclusion.} Hintikka’s analyzability theory CAT.II), though it is beautiful, is much too simple. It cannot do justice to the complex

\(^{13}\) The deduction is a trivial matter but uses his new rule (C.ind=), given \textit{ibid.}, p. 55.
logical structure of third-person attributions of self-knowledge. The new theories formulated above do indeed provide machinery sufficient to handle adequately the difficulties raised by the different degrees of quasi-indexical reference; but these theories do not seem refined enough to do justice to the subtle features (E.1) and (E.2) of attributions of self-knowledge to others, and they fail to have (9a) entail (10a). Perhaps there are some reasons for identifying interpretations (A) and (B), and, furthermore, theses T1'O, T1'A, and T1'B; but these identifications do involve a misrepresentation of the logic of self-knowledge, because they violate (E.6)-(E.8). At any rate, the fundamental issue between the current theories on the logic of self-knowledge is the issue whether the occurrences of the quasi-indicator 'he*' of degree 1 are analyzable or not. The only alternative, if the complexity of my analysis of the higher degrees is deemed unpalatable (see ibid., p. 49), is the very simple theory that no degree of 'he*' is analyzable. This theory may indeed state its case in terms of the technical notation introduced above in section 1.4.14

HECTOR-NERI CASTAÑEDA

Wayne State University

BOOK REVIEWS


Numerous writers, among them some authentic epigoni of the existentialists themselves, have claimed that existentialism has not produced any "ethics." The reasons behind this claim are varied. For one thing existentialists prefer to take their examples and paradigms from "extreme" and "limit" cases, from pathology and literature, rather than from the "normal" medium range of everyday human experience; this especially upsets British moralists. Very few would deny that the existentialists make frequent and heartfelt ethical assertions or that they are extremely sensitive to moral qualities and situations, but many would claim that they bind such judgments and feelings to a previously conceived metaphysic of human existence, an extreme "ontological individualism," of which their ethical ejaculations are just so many unanalyzable and metaphorical illustrations. More seriously, the charge is often made that they assert rather than argue, that one cannot distinguish their premises from

14 I am grateful to Lawrence Powers for having overhauled the syntax of this paper.