# SEMANTICS FOR PROPOSITIONAL ATTITUDES

## I. THE CONTRAST BETWEEN THE THEORY OF REFERENCE AND THE THEORY OF MEANING IS SPURIOUS

In the philosophy of logic a distinction is often made between the *theory of reference* and the *theory of meaning*.[1] In this paper I shall suggest (*inter alia*) that this distinction, though not without substance, is profoundly misleading. The theory of reference is, I shall argue, the theory of meaning for certain simple types of language. The only entities needed in the so-called theory of meaning are, in many interesting cases and perhaps even in all cases, merely what is required in order for the expressions of our language to be able to refer in certain more complicated situations. Instead of the theory of reference and the theory of meaning we perhaps ought to speak in some cases of the theory of simple and of multiple reference, respectively. Quine has regretted that the term 'semantics', which etymologically ought to refer to the theory of meaning, has come to mean the theory of reference.[1] I submit that this usage is happier than Quine thinks, and that large parts of the theory of meaning in reality are – or ought to be – but semantical theories for notions transcending the range of certain elementary types of concepts.

It seems to me in fact that the usual reasons for distinguishing between meaning and reference are seriously mistaken. Frequently, they are formulated in terms of a first-order (i.e., quantificational) language. In such a language, it is said, knowing the mere references of individual constants, or knowing the extensions of predicates, cannot suffice to specify their meanings because the references of two individual constants or the extensions of two predicate constants 'obviously' can coincide without there being any identity of meaning.[2] Hence, it is often concluded, the theory of reference for first-order languages will have to be supplemented by a theory of the 'meanings' of the expressions of these languages.

The line of argument is not without solid intuitive foundation, but its implications are different from what they are usually taken to be. This

J. Hintikka, *Models for Modalities*

whole concept of meaning (as distinguished from reference) is very unclear and usually hard to fathom. However it is understood, it seems to me in any case completely hopeless to try to divorce the idea of the meaning of a sentence from the idea of the *information* that the sentence can convey to a hearer or reader, should someone truthfully address it to him.[3] Now what is this information? Clearly it is just information to the effect that the sentence is true, that the world is such as to meet the truth-conditions of the sentence.

Now in the case of a first-order language these truth-conditions cannot be divested from the references of singular terms and from the extensions of its predicates. In fact, these references and extensions are precisely what the truth-conditions of quantified sentences turn on. The truth-value of a sentence is a function of the references (extensions) of the terms it contains, not of their 'meanings'. Thus it follows from the above principles that a theory of reference is for genuine first-order languages the basis of a theory of meaning. Recently, a similar conclusion has in effect been persuasively argued for (from entirely different premises and in an entirely different way) by Donald Davidson.[4] The references, not the alleged meanings, of our primitive terms are thus what determine the meanings (in the sense explained) of first-order sentences. Hence the introduction of the 'meanings' of singular terms and predicates is strictly useless: In any theory of meaning which serves to explain the information which first-order sentences convey, these 'meanings' are bound to be completely idle.

What happens, then, to our intuitions concerning the allegedly obvious difference between reference and meaning in first-order languages? If these intuitions are sound, and if the above remarks are to the point, then the only reasonable conclusion is that our intuitions do not really pertain to first-order discourse. The 'ordinary language' which we think of when we assert the obviousness of the distinction cannot be reduced to the canonical form of an applied first-order language without violating these intuitions. How these other languages enable us to appreciate the real (but frequently misunderstood) force of the apparently obvious difference between reference and meaning I shall indicate later (see Section VI *infra*).

## II. FIRST-ORDER LANGUAGES

I conclude that the traditional theory of reference, suitably extended and developed, is all we need for a full-scale theory of meaning in the case of an applied first-order language. All that is needed to grasp the information that a sentence of such a language yields is given by the rules that determine the references of its terms, in the usual sense of the word. For the purposes of first-order languages, to specify the meaning of a singular term is therefore nearly tantamount to specifying its reference, and to specify the meaning of a predicate is for all practical purposes to specify its extension. As long as we can restrict ourselves to first-order discourse, the theory of truth and satisfaction will therefore be the central part of the theory of meaning.

A partial exception to this statement seems to be the theory of so-called 'meaning postulates' or 'semantical rules' which are supposed to catch non-logical synonymies.[5] However, I would argue that whatever non-logical identities of meaning there might be in our discourse ought to be spelled out, not in terms of definitions of terms, but by developing a satisfactory semantical theory for the terms which create these synonymies. In those cases in which meaning postulates are needed, this enterprise no longer belongs to the theory of first-order logic.

In more precise terms, one may thus say that to understand a sentence of first-order logic is to know its interpretation in the actual world. To know this is to know the interpretation function $\phi$. This can be characterized as a function which does the following things:

(1.1) For each individual constant $a$ of our first-order language, $\phi(a)$ is a member of the domain of individuals **I**.

The domain of individuals **I** is of course to be thought of as the totality of objects which our language speaks of.

(1.2) For each constant predicate $Q$ (say of $n$ terms), $\phi(Q)$ is a set of $n$-tuples of the members of **I**.

If we know $\phi$ and if we know the usual rules holding of satisfaction (truth), we can in principle determine the truth-values of all the sentences of our first-order language. This is the cash value of the statement made above that the extensions of our individual constants and constant predicates are virtually all that we need in the theory of meaning in an applied first-order language.[6]

These conditions may be looked upon in slightly different ways. If $\phi$ is considered as an arbitrary function in (1.1)–(1.2), instead of that particular function which is involved in one's understanding of a language, and if $\mathbf{I}$ is likewise allowed to vary, we obtain a characterization of the concept of interpretation in the general model-theoretic sense.

## III. PROPOSITIONAL ATTITUDES

We have to keep in mind the possibility that $\phi$ might be only a partial function (as applied to free singular terms), i.e., that some of our singular terms are in fact empty. This problem is not particularly prominent in the present paper, however.[7] If what I have said so far is correct, then the emphasis philosophers have put on the distinction between reference and meaning (e.g. between *Bedeutung* and *Sinn*) is motivated only in so far as they have implicitly or explicitly considered concepts which go beyond the expressive power of first-order languages.[8] Probably the most important type of such concept is a propositional attitude.[9] One purpose of this paper is to sketch some salient features of a semantical theory of such concepts. An interesting problem will be the question as to what extent we have to assume entities other than the usual individuals (the members of $\mathbf{I}$) in order to give a satisfactory account of the meaning of propositional attitudes. As will be seen, what I take to be the true answer to this question is surprisingly subtle, and cannot be formulated by a simple 'yes' or 'no'.

What I take to be the distinctive feature of all use of propositional attitudes is the fact that in using them we are considering more than one possibility concerning the world.[10] (This consideration of different possibilities is precisely what makes propositional attitudes propositional, it seems to me.) It would be more natural to speak of different possibilities concerning our 'actual' world than to speak of several possible worlds. For the purpose of logical and semantical analysis, the second locution is much more appropriate than the first, however, although I admit that it sounds somewhat weird and perhaps also suggests that we are dealing with something much more unfamiliar and unrealistic than we are actually doing. In our sense, whoever has made preparations for more than one course of events has dealt with several 'possible courses of events' or 'possible worlds'. Of course, the possible courses of events he considered were from his point of view so many alternative courses that the actual events might take. However, only one such course of events (at most) became actual. Hence there is a sense in which the others were merely 'possible courses of events', and this is the sense on which we shall try to capitalize.

Let us assume for simplicity that we are dealing with only one propositional attitude and that we are considering a situation in which it is attributed to one person only. Once we can handle this case, a generalization to the others is fairly straightforward. Since the person in question remains constant throughout the first part of our discussion, we need not always indicate him explicitly.

## IV. PROPOSITIONAL ATTITUDES AND 'POSSIBLE WORLDS'

My basic assumption (slightly oversimplified) is that an attribution of any propositional attitude to the person in question involves a division of all the possible worlds (more precisely, all the possible worlds which we can distinguish in the part of language we use in making the attribution) into two classes: into those possible worlds which are in accordance with the attitude in question and into those which are incompatible with it. The meaning of the division in the case of such attitudes as knowledge, belief, memory, perception, hope, wish, striving, desire, etc. is clear enough. For instance, if what we are speaking of are (say) *a*'s memories, then these possible worlds are all the possible worlds compatible with everything he remembers.

There are propositional attitudes for which this division is not possible. Some such attitudes can be defined in terms of attitudes for which the assumptions do hold, and thus in a sense can be 'reduced' to them. Others may fail to respond to this kind of attempted reduction to those 'normal' attitudes which we shall be discussing here. If there really are such recalcitrant propositional attitudes, I shall be glad to restrict the scope of my treatment so as to exclude them. Enough extremely important notions will still remain within the purview of my methods.

There is a sense in which in discussing a propositional attitude, attributed to a person, we can even restrict our attention to those possible worlds which are in accordance with this attitude.[11] This may be brought out e.g. by paraphrasing statements about propositional attitudes in terms of this restricted class of all possible worlds. The following examples will illustrate these approximate paraphrases:

> *a* believes that $p$ = in all the possible worlds compatible with what *a* believes, it is the case that $p$;
>
> *a* does not believe that $p$ (in the sense 'it is not the case that *a* believes that $p$') = in at least one possible world compatible with what *a* believes it is not the case that $p$.

## V. SEMANTICS FOR PROPOSITIONAL ATTITUDES

What kind of semantics is appropriate for this mode of treating propositional attitudes? Clearly what is involved is a set $\Omega$ of possible worlds or of models in the usual sense of the word. Each of them, say $\mu \in \Omega$, is characterized by a set of individuals $\mathbf{I}(\mu)$ existing in that 'possible world'. An interpretation of individual constants and predicates will now be a two-argument function $\phi(a, \mu)$ or $\phi(Q, \mu)$ which depends also on the possible world $\mu$ in question. Otherwise an interpretation works in the same way as in the pure first-order case, and the same rules hold for propositional connectives as in this old case.

Simple though this extension of the earlier semantical theory is, it is in many ways illuminating. For instance, it is readily seen that in many cases earlier semantical rules are applicable without changes. *Inter alia*, in so far as no words for propositional attitudes occur inside the scope of a quantifier, this quantifier is subject to the same semantical rules (satisfaction conditions) as before.

## VI. MEANING AND THE DEPENDENCE OF REFERENCE ON 'POSSIBLE WORLDS'

A new aspect of the situation is the fact that the reference $\phi(a, \mu)$ of a singular term now depends on $\mu$ – on what course the events will take, one might say. This enables us to appreciate an objection which you probably felt like making earlier when it was said that in a first-order language the theory of meaning is the theory of reference. What really determines the meaning of a singular term, you felt like saying, is not whatever reference it happens to have, but rather the way in which this reference is determined. But in order for this to make any difference, we must consider more than one possibility as to what the reference is, depending on the circumstances (i.e. depending on the course events will take). This dependence is just what is expressed by $\phi(a, \mu)$ when it is considered as a function of $\mu$. (This function *is* the meaning of $a$, one is tempted to say.) Your objection thus has a point. However, it does not show that more is involved in the theory of meaning for first-order languages than the references of its terms. Rather, what is shown is that in order to spell out the idea that the meaning of a term is the way in which its reference is determined we have to consider how the reference varies in different possible worlds, and therefore go beyond first-order languages, just as I suggested above. Analogous remarks apply of course to the extensions of predicates.

Another novelty here is the need of picking out one distinguished possible world from among them all, viz. the world that happens to be actualized ('the actual world').

## VII. DEVELOPING AN EXPLICIT SEMANTICAL THEORY: ALTERNATIVENESS RELATIONS

How are these informal observations to be incorporated into a more explicit semantical theory? According to what I have said, understanding attributions of the propositional attitude in question (let us assume that this is expressed by '*B*') means being able to make a distinction between two kinds of possible worlds, according to whether they are compatible with the relevant attitudes of the person in question. The semantical counterpart to this is of course a function which to a given individual person assigns a set of possible worlds.

However, a minor complication is in order here. Of course, the person in question may himself have different attitudes in the different worlds we are considering. Hence this function in effect becomes a relation which to a given individual *and to a given possible world* $\mu$ associates a number of possible worlds which we shall call the *alternatives* to $\mu$. The relation will be called the alternativeness relation. (For different propositional attitudes, we have to consider different alternativeness relations.) Our basic apparatus does not impose many restrictions on it. The obvious requirement that ensues from what has been said is the following:

(S.B.) $B_a p$ is true in a possible world $\mu$ if and only if $p$ is true in all the alternatives to $\mu$.

$B_a p$ may here be thought of as a shorthand for '$a$ believes that $p$'. We can write this condition in terms of an interpretation function $\phi$. What understanding $B$ means is to have a function $\phi_B$ which to a given possible world $\mu$ and to a given individual $a$ associates a set of possible worlds $\phi_B(a, \mu)$, namely, the set of all alternatives to $\mu$.[12] Intuitively, they are the possible worlds compatible with the presence of the attitude expressed by $B$ in the person $a$ in the possible world $\mu$.

In terms of this extended interpretation function, (S.B) can be written as follows:

$B_a p$ is true in $\mu$ if and only if $p$ is true in every member of $\phi_B(a, \mu)$.

## VIII. RELATION TO QUINE'S CRITERION OF COMMITMENT

The interesting and important feature of this truth-condition is that it involves quantification over a certain set of possible worlds. By Quine's famous criterion, we allegedly are ontologically committed to whatever we quantify over.[13] Thus my semantical theory of propositional attitudes seems to imply that we are committed to the existence of possible worlds as a part of our ontology.

This conclusion seems to me false, and I think that it in fact constitutes a counter-example to Quine's criterion of commitment *qua* a criterion of ontological commitment. Surely we must in some sense be *committed* to whatever we quantify over. To this extent Quine seems to be entirely right. But why call this a criterion of *ontological* commitment? One's ontology is what one assumes to exist in one's world, it seems to me. It is, as it were, one's census of one's universe. Now such a census is meaningful only in some particular possible world. Hence Quine's criterion can work as a criterion of *ontological* commitment only if the quantification it speaks of is a quantification over entities belonging to some one particular world. To be is perhaps to be a value of a bound variable. But to exist in an ontologically relevant sense, to be a part of the furniture of the world, is to be a value of a special kind of a bound variable, namely one whose values all belong to the same possible world. Thus the notion of a possible world serves to clarify considerably the idea of ontological commitment so as to limit the scope of Quine's dictum.

Clearly, our quantification over possible worlds does not satisfy this extra requirement. Hence there is a perfectly good sense in which we are not ontologically committed to possible worlds, however important their role in our semantical theory may be.

Quine's distinction between *ontology* and *ideology*, somewhat modified and put to a new use, is handy here.[14] We have to distinguish between what we are committed to in the sense that we believe it to exist in the actual world or in some other possible world, and what we are committed to as a part of our ways of dealing with the world conceptually, committed to as a part of our conceptual system. The former constitute our ontology, the latter our 'ideology'. What I am suggesting is that the possible worlds we have to quantify over are a part of our ideology but not of our ontology.

The general criterion of commitment is a generalization of this. Quantification over the members of one particular world is a measure of ontology, quantification that crosses possible worlds is often a measure of ideology. Quine's distinction thus ceases to mark a difference between two different types of studies or two different kinds of entities within one's universe. It now marks, rather, a distinction between the objects of reference and certain aspects of our own referential apparatus. Here we can perhaps see what the so-called distinction between theory of reference and theory of meaning really amounts to.

It follows, incidentally, that if we could restrict our attention to *one* possible world only, Quine's restriction would be true without qualifications. Of course, the restriction is one which Quine apparently would very much like to make; hence he has a legitimate reason for disregarding the qualifications for his own purposes.

Our 'ideological' commitment to possible worlds other than the actual one is neither surprising nor disconcerting. If what we are dealing with are the things people do – more specifically, the concepts they use – in order to be prepared for more than one eventuality, it is not at all remarkable that in order to describe these concepts fully we have to speak of courses of events other than the actual one.

## IX. SINGULAR TERMS AND QUANTIFICATION IN THE CONTEXT OF PROPOSITIONAL ATTITUDES

Let us return to the role of individual constants (and other singular terms). Summing up what was said before, we can say that what the understanding of an individual constant amounts to in a first-order language is knowing which individual it stands for. Now it is seen that in the presence of propositional attitudes this statement has to be expanded to say that one has to know what the singular term stands for in the different possible worlds we are considering.

Furthermore, in the same way as these individuals (or perhaps rather the method of specifying them) may be said to be what is 'objectively given' to us when we understand the constant, in the same way what is involved in the understanding of a propositional attitude is precisely that distinction which in our semantical apparatus is expressed by the function which serves to define the alternativeness relation. This function is what is 'objectively given' to us with the understanding of a word for a propositional attitude.

These observations enable us to solve almost all the problems that relate to the use of identity in the context of propositional attitudes. For instance, we can at once see why the familiar principle of the substitutivity of identity is bound to fail in the presence of propositional attitudes when applied to arbitrary singular terms.[15] Two such terms, say $a$ and $b$, may refer to one and the same individual in the actual world ($\phi(a, \mu_0) = \phi(b, \mu_0)$ for the world $\mu_0$ that happens to be actualized), thus making the identity '$a=b$' true, and yet fail to refer to the same individual in some other (alternative) possible world. (I.e., we may have $\phi(a, \mu_1) \neq \phi(b, \mu_1)$ for some $\mu_1 \in \phi_B(c, \mu_0)$ where $c$ is the individual whose attitudes are being discussed and $B$ the relevant attitude.) Since the presence of propositional attitudes means (if I am right) that these other possible worlds have to be discussed as well, in their presence the truth of the identity '$a=b$' does not guarantee that the same things can be said of the references of $a$ and $b$ without qualification, i.e., does not guarantee the intersubstitutivity of the terms $a$ and $b$.

Our observations also enable us to deal with quantification in contexts governed by words for propositional attitudes as long as we do not quantify *into* them. However, as soon as we try to do so, all the familiar difficulties which have been so carefully and persuasively presented by Quine and others will apply with full force.[16] An individual constant occurring within the scope of an operator like $B$ which expresses a propositional attitude does not specify a unique individual. Rather, what it does is to specify an individual in each of the possible worlds we have to consider. Replace it by an individual variable, and you do not get anything that you could describe by speaking of the individuals over which this variable ranges. There are (it seems) simply no uniquely defined individuals here at all.

It is perhaps thought that the way out is simply to deny that one can ever quantify into a non-extensional context. However, this way out does not work.[17] As a matter of fact, in our ordinary language we often quantify into a grammatical construction governed by an expression for a propositional attitude. Locutions like 'knows who', 'sees what', 'has an opinion concerning the identity of' are cases in point, and so is almost any (other) construction in which pronouns are allowed to mix with words for propositional attitudes. Beliefs about 'oneself' and 'himself' yield further examples, and an account of their peculiarities leads to an interesting reconstruction of the traditional distinction between so-called modalities *de dicto* and *de re*.[18]

Another general fact is that we obviously have beliefs about definite individuals and not just about whoever happens to meet a certain description. I want to suggest that such beliefs (and the corresponding attitudes in the case of other propositional attitudes) are precisely what one half of the *de dicto* – *de re* distinction amounts to.[19]

Furthermore, it does not do to try to maintain that in these constructions the propositional attitude itself has to be taken in an unusual extensional or 'referentially transparent' sense. Such senses can in fact be defined in terms of the normal senses of propositional attitudes. However, these definitions already involve the objectionable quantification into opaque contexts, and if one tries to postulate the defined senses as irreducible primitive senses, they do not have the properties which they ought to have in order to provide the resulting quantified statements with the logical powers they in fact have in ordinary language. For instance, Quine's attempt to postulate a sense of (say) knowledge in which one is allowed to quantify into a context governed by a transparently construed construction 'knows that' has the paradoxical result that

$$(Ex) \text{ Jones knows that } (x=a)$$

is implied by any (transparently interpreted) statement of the form

Jones knows that $(b=a)$

and even by a similarly interpreted sentence

Jones knows that $(a=a)$.

This I take to show that the first of these three sentences can scarcely serve as a formulation of 'Jones knows who (or what) $a$ is' in our canonical idiom. Yet no other paraphrase of this ubiquitous locution has been proposed, and none is likely to be forthcoming. (For what else can there be to Jones' knowing who $a$ is than his knowing of some well-defined individual that Jones is that very individual? And it is Quine who always insists as strongly as anyone else that the values of bound variables have to be well-defined individuals.) It is not much more helpful to try to maintain that no true sentences of the form

Jones knows that $(b=a)$

(with the transparent sense of 'knows') are forthcoming whenever Jones fails to know who $a$ is. The transparent sense in which this would be the case has never been explained in a satisfactory way, and I do not see how it can be done in a reasonable way without falling back to my own analysis. (What can it conceivably mean e.g. for Jones not to know in the transparent sense that an $a$, whom he knows to exist, is not self-identical? Can this self-identity fail to be true in a possible world compatible with everything Jones knows?)

Hence we have to countenance quantification into a context governed by an expression for an (opaquely construed) propositional attitude. Our semantical theory at once suggests a way of handling these problems. For instance, in order for existential generalization to be applicable to a singular term $b$ occurring, say, in a context where $a$'s beliefs are being discussed, it has to be required that $b$ refers to the *same* individual in the different possible worlds compatible with what $a$ believes (plus, possibly, in the actual world). This, naturally, will be expressed by a statement of the form

(*) $$(Ex)\,[B_a(x=b)\,\&\,(x=b)]$$

or, if we do not have to consider the actual world, of the form $(Ex)\,B_a(x=b)$.

## X. METHODS OF CROSS-IDENTIFICATION

This solution is simple, straightforward, and workable. It generalizes easily to other propositional attitudes. However, it hides certain interesting conceptual presuppositions. With what right do we speak of individuals in the different possible worlds as being *identical*? This is the problem to which we have to address ourselves.

It is not difficult to see what more there is given to us with our ordinary understanding of propositional attitudes that we have not yet dealt with. For instance, consider a man who has a number of beliefs as to what will happen tomorrow to himself and to his friends. Consider, on his behalf, a number of possible courses of events tomorrow. If I know what our man believes, I can sort these into those which are compatible with his beliefs as distinguished from those which are incompatible with them. But this is not all that is involved. Surely the same or largely the same individuals must figure in these different sequences of events. Under different courses of events a given individual may undergo different experiences, entertain different beliefs and hopes and fears; he may behave rather differently and perhaps even look somewhat different. Nevertheless our man can be (although he need not be) and usually is completely confident that, whatever may happen, he is going to be able to recognize (re-identify) his friends under these various courses of events, at least in principle. He may admit that courses of events are perhaps logically possible under which he would fail to do so; but these would not be compatible with his beliefs as to what will happen. Given full descriptions of two different courses of events tomorrow, both compatible with what our man believes ('believes possible', we sometimes say with more logical than grammatical justification), he will be able to recognize which individuals figuring in one of these descriptions are identical with which individual in the other, even if their names are being withheld. (Of course our man need not believe all this but my point is merely that he *can* and very often *does* believe it.)

The logical moral of this story is that together with the rest of our beliefs we are often given something more than we have so far incorporated into our semantical theory. We are given ways of *cross-identifying* individuals, that is to say, ways of understanding questions as to whether an individual figuring in one possible world is or is not identical with an individual figuring in another world.[20]

This is one point at which the obviousness of my claim may be partially obscured by my terminology. Let us recall what these 'possible worlds' are in the case of a propositional attitude. They are normally possible states of affairs or courses of events compatible with the attitude in question in some specified person. Now normally these attitudes may be attitudes toward definite persons or definite physical objects. But how is it that we may be sure, sight unseen, that the attitudes are directed toward the right persons or objects? Only if in all the possible worlds compatible with the attitude in question we can pick out the recipient of this attitude, i.e. the individual at its receiving end. Although in many concrete situations the possibility of doing so is obvious, it has not been built into our semantical apparatus so far. There is so far nothing in our semantical theory which enables us to relate to each other the members of the different domains of individuals $\mathbf{I}(\mu)$. In many, though not necessarily all, applications of such relations are given to us as a part of our understanding of the concepts involved. For such cases, we have to build a richer semantical theory.

The way to do so is to postulate a method of making cross-identifications. One possible way to do so is to postulate a set of functions $\mathbf{F}$ each member $f$ of which picks out at most one individual $f(\mu)$ from the domain of individuals $\mathbf{I}(\mu)$ of each given model $\mu$. We must allow that there is no such value for some models $\mu$. In other words, $f \in \mathbf{F}$ may be a partial function. Furthermore, we must often require that, given $f_1, f_2 \in \mathbf{F}$, if $f_1(\mu) = f_2(\mu)$ then $f_1(\lambda) = f_2(\lambda)$ for all alternatives $\lambda$ to $\mu$. In other words, an individual cannot 'split' when we move from a world to its alternatives. This question may seem to be a mere matter of detail, but it is easily seen that the question whether an individual can split in the sense just explained is tantamount to the question whether the substitutivity of identity can fail for bound (individual) variables, i.e. to the question whether a sentence

$$(Ux)(Uy)(x = y \supset B_a(x = y))$$

can fail to be logically true. This, again, is tantamount to the question whether a sentence of the form

$$(Ux)(Uy)(x = y \supset (Q(x) \supset Q(y)))$$

(with just one layer of operators for propositional attitudes in $Q$) can fail to be logically true.

In terms of the set **F**, the question whether $a \in \mathbf{I}(\mu)$ is identical with $b \in \mathbf{I}(\lambda)$ amounts to the question whether there is a function of $f \in \mathbf{F}$ such that $f(\mu) = a, f(\lambda) = b$.

## XI. THE ROLE OF INDIVIDUATING FUNCTIONS

Instead of speaking of a set of functions correlating to each other the individuals existing in the different possible worlds, it is often more appropriate to speak of these domains of individuals as being partly identical (overlapping). Then there would be no need to speak of correlations at all. This point of view is useful in that it illustrates the fact that the apparently different individuals which are correlated by one of the functions $f \in \mathbf{F}$ is just what we ordinarily mean by one and the same individual. It is the concrete individual which we speak about, which we give a name to, etc. In fact, the members of **F** might in fact be thought of as names or individual constants of a certain special kind, namely those having a unique reference in all the different worlds we are speaking of and hence satisfying formulas of the form (*). Indeed, I shall assume in the sequel that a constant of this kind can be associated with each function $f \in \mathbf{F}$.

However, emphasizing the role of the functions $f \in \mathbf{F}$ is useful for several purposes. First and foremost, it highlights an extremely important non-trivial part of our native conceptual skills, namely, our capacity to recognize one and the same individual under different circumstances and under different courses of events. What the set **F** of functions embodies is just the totality of ways of doing this. The non-trivial character of the possibility of this recognition would be lost if we should simply speak of the members of the different possible worlds as being partly identical.

For another thing, the structure formed by the relations of cross-world identity (David Kaplan calls them "trans world heir lines") may be so complex as to be indescribable by speaking simply of partial identities between the domains of individuals of the different possible worlds. Above, it was said that in the case of many propositional attitudes an individual cannot 'split' when we move from a world to its alternatives. Although this seems to me to be the case with all the propositional attitudes I have studied in any detail, it is not quite clear to me precisely why this should always be the case. At any rate, there seem to be reasons for suspecting that the opposite 'irregularity' can occasionally take place with some modalities: individuals can 'merge together' when we move from a world to its alternatives. An analogy with temporal modalities may be instructive here.[21] If we presuppose some suitable system of cross-identifications between individuals existing at different times which turn on continuity, it seems possible in principle that a singular term should refer to the same physical system at all the different moments of time we are considering although this system 'merges' with others at times and occasionally 'splits up' into several. Some of these complications seem to be impossible to rule out completely in the case of some propositional attitudes, and because of them the idea of partly overlapping domains seems to me seriously oversimplified.

An extremely important further reason why we cannot reify the members of **F** into ordinary individuals is the possibility of having two different methods of cross-identification between the members of the same possible worlds, i.e. two different sets of 'individuating functions' although we are dealing with precisely the same sets of possible worlds. I have argued elsewhere that this kind of situation is not only possible to envisage but is actually present in our own ways with perceptual contexts.[22] It would take us too far to show precisely what is involved in such cases. Suffice it to point out that this claim, if true, would strikingly demonstrate the dependence of our methods of cross-identification on our own conceptual schemes and hence on things of our own creation. The apparent simplicity of our idea of an 'ordinary' individual, safe as it may seem in its solid commonplace reality, is thus seen to be merely a reflection of the familiarity and relatively deep customary entrenchment of one particular method of cross-identification, which *sub specie aeternitatis* (i.e. *sub specie logicae*) nevertheless enjoys but a relative privilege as against a host of others.

The methods of cross-identification represented by the set **F** of 'individuating functions', as we might call them, also call for several further comments.

The main function of this part of our semantical apparatus is to make sense of quantification into contexts of propositional attitudes. The truth-conditions of statements in which this happens can be spelled out in terms of membership in **F**. As an approximation we can say the following: A sentence of the form $(Ex)\ Q(x)$ is true in $\mu$ if and only if there is an individual constant (say $b$) associated with some $f \in \mathbf{F}$ such that $Q(b)$ is true in $\mu$. This approximation shows, incidentally, how close we can stick to the simple-minded idea that an existentially quantified sentence is true if and only if it has a true substitution instance. The only additional requirement we need is that the substitution-value of the bound variable has to be of the right sort, to wit, has to specify the same individual in all the possible worlds we are speaking of in the existential sentence in question. This is what is meant by the requirement that $b$ has to be associated with one of the functions $f \in \mathbf{F}$.

This approximation, although not unrepresentative of the general situation, requires certain modifications in order to work in all cases. The set $\mathbf{F}$ has to be relativized somewhat in the same way the unrestricted notion of a possible world was relativized by the notion of alternative in the truth-criterion (S.B.) above. (Not everyone is in all situations 'familiar with' all the relevant methods of individuation, it might be said.) I shall not discuss the ensuing complications here, however, for they do not change the overall picture in those respects which are relevant in the rest of this paper.

## XII. STATEMENTS ABOUT DEFINITE INDIVIDUALS VS. STATEMENTS ABOUT WHOEVER OR WHATEVER IS REFERRED TO BY A TERM

The possibility of quantifying across an operator which expresses a propositional attitude enables us to explicate the logic of the locutions in which we need this possibility in the first place. Perhaps the most important thing we can do here is to make a distinction between propositional attitudes directed to whoever (whatever) happens to be referred to by a term and attitudes directed toward a certain individual, independently of how he happens to be referred to. This distinction was hinted at above. Now it is time to explain it more fully. For instance, someone may have a belief concerning the next Governor of California, whoever he is or may be, say, that he will be a Democrat. This is different from believing something about the individual who, so far unbeknownst to all of us, in fact is the next Governor of California.

In formal terms, the distinction is illustrated by the pair of statements

$$B_a\,(g \text{ is a Democrat})$$
$$(Ex)\,((x = g)\,\&\,B_a\,(x \text{ is a Democrat})).$$

Notice, incidentally, that my way of drawing this distinction implies that one can have (say) a belief concerning the individual who in fact is *a* only if such an individual really exists, whereas one can in principle have a belief concerning *a*, 'whoever he is', even though there is no such person. This, of course, is just as it ought to be.

The naturalness of our semantical conditions, and their close relation to the realities of actual usage, can be illustrated by applying them to what I have called a statement about a definite individual. As an example, we can use

> '*a* believes of the man who in fact is Mr. Smith that he is a thief',

in brief,

$$(Ex)\,(x = \text{Smith} \,\&\, B_a\,(x \text{ is a thief})).$$

In order for this to be true, there has to be some $f \in \mathbf{F}$ such that the value of $f$ in the actual world (call it $\mu_0$) exists and is Smith and that $f(\mu)$ has the property of being a thief whenever $\mu \in \phi_B(a, \mu_0)$, i.e. in all the alternatives to the actual world.

What the requirement of the existence of $f$ amounts to is clear enough. If it is true to say that *a* has a belief about *the particular individual* who in fact is Smith, then *a* clearly must believe that he can characterize this individual uniquely. In other words, he must have some way of referring to or characterizing this individual in such a way that one and the same individual is in fact so characterized in all the worlds compatible with what he believes. This is precisely what the existence of $f$ amounts to. If no such function existed, *a* would not be able to pick out the individual who in fact is Smith under all the courses of events he believes possible, and there would not be any sense in saying that *a*'s belief is *about* the particular individual in question.

## XIII. INDIVIDUATING FUNCTIONS VS. INDIVIDUAL CONCEPTS

One important consequence of my approach is that not every function which from each $\mu$ picks out an individual can be said to specify a unique individual. In fact, many perfectly good free singular terms fail to do so in the context of many propositional attitudes. Even proper names fail to do so in epistemic contexts, for one may fail to know who the bearer of a given proper name is.

Such arbitrary functions may be important for many purposes. They are excellent approximations in our theory to the 'individual concepts' which many philosophers have postulated.[23] (In Section VI above we already met a number of such 'individual concepts' in the form of the functions $\phi(a, \mu)$ with a fixed $a$.) Each such individual concept specifies or 'contains', as Frege would say, not just a reference (in the actual world) but also the way in which this reference is given to us. Each of them would thus qualify for a sense (*Sinn*) of a singular term *à la* Frege.[24] However, we do not need the totality of such arbitrary functions in the semantics which I am building up and which (I want to argue) is largely implicit in our native conceptual apparatus. Quine's criterion, however misleading it may be as a criterion of ontological commitment, still works as a criterion of commitment. If it is applied here, it shows that we are not committed (ontologically or 'ideologically') to these arbitrary functions, since we do not have to quantify over them, only over the members of the much narrower class **F**.

The other side of the coin is that in our semantical apparatus we do have to quantify over the members of **F**. Does it follow that they 'exist' or 'are part of our ontology'? An answer to this question can be given along the same lines as to the corresponding question concerning 'possible worlds'. The members of **F** are not members of any possible world; they are not part of anybody's count of 'what there is'. They may 'subsist' or perhaps 'exist', and they are certainly 'objective', but they do not have any ontological role to play. The need to distinguish between ontology and 'ideology' is especially patent here.

The functions that belong to **F** may of course be considered special cases of the 'individual concepts' postulated by some philosophers of logic or as special cases of Frege's 'senses' (*Sinne*). No identification is possible between the two classes, however, for we saw earlier that not every arbitrary singular term (say $b$) which picks out an individual from each $\mathbf{I}(\mu)$ we are considering goes together with an $f \in \mathbf{F}$, although every such term is certainly meaningful and hence has a Fregean 'sense' and perhaps even gives us an 'individual concept'. As I have put it elsewhere, members of **F** do not only involve a 'way of being given' as Frege's senses do, but also *a way of being individuated*.[25] The primary care is in our approach devoted to ordinary concrete individuals. Singular terms merit a special honorary mention only if they succeed in picking out a unique individual of this sort.

Let us say that an $f \in \mathbf{F}$ is (gives us) an individuating concept, and let us say that a term $b$ does *individuate* (in the context of discussing $a$'s beliefs) in so far as

$$(Ex)\ B_a\ (x = b)$$

is true. Then we could have individuation without reference and reference without individuation: Both

$$(Ex)\ B_a\ (x = b)\ \&\ \sim (Ex)\ (x = b)$$

and

$$(Ex)\ (x = b)\ \&\ \sim (Ex)\ B_a\ (x = b)$$

can be true. We could even have both, but without matching:

$$(Ex)\ (x = b)\ \&(Ex)\ B_a\ (x = b)\ \&\sim (Ex)\ ((x = b)\ \&\ B_a\ (x = b))$$

is satisfiable. Only if

$$(Ex)\ ((x = b)\ \&\ B_a\ (x = b))$$

is true does the successful individuation give us the individual which the term $b$ actually refers to.

## XIV. THE THEORY OF REFERENCE AS REPLACING THE THEORY OF MEANING

Here we are perhaps beginning to see what I meant when I said at the beginning of this paper that what is often called the theory of meaning is better thought of as the theory of reference for certain more complicated conceptual situations. Some of the most typical concepts used in the theory of meaning, such as Frege's *Sinne* and the 'individual concepts' of certain other philosophers of logic were in the first place introduced to account for such puzzles as the failure of the substitutivity of identity and the difficulty of quantifying into opaque contexts (e.g. into a context governed by a word for a propositional attitude). I have argued, however, that a satisfactory semantical theory which clears up these puzzles can be built up without using Frege's *Sinne* and without any commitment to individual concepts in any ordinary sense of the word. Instead, what we need are the individuating functions, i.e., the members of **F**. And what these functions do is not connected with the ideas of the traditional theory of meaning. What they do is precisely to give us the individuals which we naively think our singular statements to be about and which we think our singular terms as referring to. This naive point of view is essentially correct, it seems to me. The functions of $f \in \mathbf{F}$ are the prime vehicles of our references to individuals when we discuss propositional attitudes. What is not always realized, however, is how much goes into our ordinary concepts of an individual and a reference. These are not specified in a way which works only under one particular course of events. They are in fact specified in a way which works under a wide variety of possible courses of events. But, in order to spell out this idea, we are led to consider several possible worlds, with all the problems with which we have dealt in this paper, including very prominently the problem of cross-identifying individuals.

The function of our 'individuating functions', i.e. the members of the set **F**, is to bring out these hidden – or perhaps merely overlooked – aspects of our concept of an individual (definite individual). This close connection between the set **F** and the concept of an individual appears in a variety of ways. One may for instance think of the role which the membership in **F** plays in the truth-conditions which we set up above for quantification into modal contexts. When it is asked in such a context whether there exists an individual of a certain kind, a singular term specifies such an individual only if its references match the values of a unique member of **F** in all the relevant possible worlds. Thus it is these functions that in effect give us the individuals which can serve as values of bound variables. As we saw above, it is mainly the possible subtlety and multiplicity of relations of cross-identity that prevent us from simply making the domains of the different possible worlds partly identical and thus hypostatizing my individuating functions into commonplace individuals.

This connection between individuating functions and the concept of an individual is part of what justifies us in thinking that in the traditional dichotomy their theory would belong primarily to the theory of reference, in spite of the fact that their main function in our semantical theory is to solve some of the very problems which the traditional theory of meaning was calculated to handle. This role is perhaps especially clear in connection with the substitutivity of identity. As we have seen, this principle does not hold for arbitrary singular terms $a$, $b$. However, if it is required in addition that both of these terms specify a well-defined individual, i.e. satisfy expressions like (*), depending on the context, then the substitutivity of identicals is easily seen to hold, presupposing of course here the prohibition against splitting that was mentioned above. What this observation shows is clear enough. The failure of the substitutivity of identity poses one of the most typical problems for the treatment of which meanings, individual concepts and other paraphernalia of the theory of meaning were introduced in the first place. If the substitutivity of identity fails, clearly we cannot be dealing with ordinary commonplace individuals, it was alleged, for if two such individuals are in fact identical, surely precisely same things can be said of them. This is what prompts the quest for individuals of some non-ordinary sort, capable of restoring the substitutivity principle when used as references of our terms. (This is almost precisely Frege's strategy.) We have seen, however, that the (apparent) failure of the substitutivity is due simply to the failure of some free singular terms to specify the same individual in the different 'possible worlds' we have to consider. Moreover, we have seen that this apparent failure is automatically corrected in precisely those cases in which it ought to be corrected, viz. in the cases where the two terms in question really do specify a *unique* individual. (That this depends on certain specific requirements concerning our methods of cross-identification, viz. on a prohibition against 'splitting', does not affect my point.) Substitutivity of identity is restored, in brief, not by requiring that our singular terms refer to the entities postulated by the so-called theory of meaning, but by requiring (in the form of an explicit premise) that they really succeed in specifying uniquely the kind of ordinary individual with which the theory of reference typically deals. One can scarcely hope to find a more striking example of the breakdown of the distinction between a theory of meaning and a theory of reference.[26]

## XV. TOWARDS A SEMANTIC NEOKANTIANISM

The aspect of my observations most likely to upset many contemporary philosophers is the ensuing implicit dependence of our concept of an individual on our ways of cross-identifying members of different 'possible worlds'. These 'possible worlds' and the supply of individuating functions which serve to interrelate their respective members may enjoy, and in my view do enjoy, some sort of objective reality. However, their existence is not a 'natural' thing. They may be as solidly objective as houses or books, but they are as certainly as these created by men (however unwittingly) for the purpose of facilitating their transactions with the reality they have to face. Hence my reasoning ends on a distinctly Kantian note. Whatever we say of the world is permeated throughout with concepts of our own making. Even such *prima facie* transparently simple notions as that of an individual turn out to depend on conceptual assumptions dealing with different possible states of affairs. As far as our thinking is concerned, reality cannot be in principle wholly disentangled from our concepts. A *Ding an sich*, which could be described or even as much as individuated without relying on some particular conceptual framework, is bound to remain an illusion.

## REFERENCES

[1] See e.g. W. V. Quine, *From a Logical Point of View*, Harvard University Press, Cambridge, Mass., 1953 (2nd ed.: 1961), pp. 130–132.

[2] For a simple recent argument of this sort (without a specific reference to first-order theories), see e.g. William P. Alston, *Philosophy of Language*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964, p. 13. Cf. also Quine, *op. cit.* pp. 21–22.

[3] In more general terms, it seems to me hopeless to try to develop a theory of sentential meaning which is not connected very closely with the idea of the information which the sentence can convey to us, or a theory of meaning for individual words which would not show how understanding them contributes to appreciating the information of the sentences in which they occur. There are of course many nuances in the actual use of words and sentences which are not directly explained by connecting meaning and information in this way, assuming that this can be done. However, there do not seem to be any obstacles in principle to explaining these nuances in terms of pragmatic, contextual, and other contingent pressures operating on a language-user. For remarks on this methodological situation, see my paper 'Epistemic Logic and the Methods of Philosophical Analysis', *Australasian Journal of Philosophy* **46** (1968) 37–51.

[4] Donald Davidson, 'Truth and Meaning', *Synthese* **17** (1967) 304–323.

[5] See Quine, *op. cit.*, pp. 32–37.

[6] The main reason why the truth of these observations is not appreciated more widely seems to be the failure to consider realistically what the actual use of a first-order language (say for the purpose of conveying information to another person) would look like.

[7] The basic problems as to what happens when this possibility is taken seriously are discussed in my paper, 'Studies in the Logic of Existence and Necessity I', *The Monist* **50** (1966) 55–76, reprinted (with changes) in the present volume under the title 'Existential Presuppositions and Their Elimination' (above pp. 23–44).

[8] This is certainly true of Frege. His very interest in oblique contexts seems to have been kindled by the realization that they cannot be handled by means of the ideas he had successfully applied to first-order logic.

[9] The term seems to go back to Bertrand Russell, *An Inquiry into Meaning and Truth*, George Allen and Unwin, London, 1940.

[10] An important qualification here is that for deep logical reasons one cannot usually distinguish effectively between what is 'really' a logically possible world and what

merely 'appears' on the face of one's language (or thinking) to be a possibility. This, in a sufficiently sharp analysis, is what destroys the pleasant invariance of propositional attitudes with respect to logical equivalence. Even though $p$ and $q$ are equivalent, i.e. even though the 'real' possibilities concerning the world that they admit and exclude are the same,

| | | |
|---|---|---|
| | knows | |
| | believes | |
| *a* | remembers | that *p* |
| | hopes | |
| | strives | |

and

| | | |
|---|---|---|
| | knows | |
| | believes | |
| *a* | remembers | that *q* |
| | hopes | |
| | strives | |

need not be equivalent, for the apparent (to $a$) possibilities admitted by $p$ and $q$ need not be identical.

I have studied this concept of an 'apparent' possibility and its consequences at some length elsewhere (especially in the second and third paper printed in *Deskription, Analytizität und Existenz* (ed. by Paul Weingartner), Pustet, Salzburg and Munich 1966, in 'Are Logical Truths Analytic?', *Philosophical Review* **74** (1965) 178–203, in 'Surface Information and Depth Information' forthcoming in *Information and Inference* (ed. by K. J. J. Hintikka and P. Suppes), D. Reidel Publishing Co., Dordrecht, 1969), and in 'Are Mathematical Truths Synthetic A Priori?', *Journal of Philosophy* **65** (1968) 640–651.
It is an extremely interesting concept to study and to codify. However, it is not directly relevant to the concerns of the present paper, and would in any case break its confines. Hence it will not be taken up here, except by way of this *caveat.*

[11] There is a distinction here which is not particularly relevant to my concerns in the present paper but important enough to be noted in passing, especially as I have not made it clear in my earlier work. What precisely are the worlds 'alternative to' a given one, say $\mu$? A moment's reflection on the principles underlying my discussion will show, I trust, that they must be taken to be worlds compatible with a certain person's having a definite propositional attitude in $\mu$, and not just compatible with the content of his attitude, for instance, compatible with someone's knowing something in $\mu$ and not just compatible with what he knows. I failed to spell this out in my *Knowledge and Belief* (Cornell Univ. Press, Ithaca, N.Y., 1962), as R. Chisholm in effect pointed out in his review article, 'The Logic of Knowing', *Journal of Philosophy* **60** (1963) 773–795.

[12] As the reader will notice, I am misusing (in the interest of simplicity) my terminology systematically by speaking elliptically of 'the person $a$' etc. when 'the person referred to by $a$' or some such thing is meant. I do not foresee any danger of confusion resulting from this, however.

[13] E.g. Quine, *op. cit.*, pp. 1–14, W.V. Quine, *Word and Object*, The MIT Press, Cambridge, Mass., and John Wiley, New York and London, 1960, pp. 241–243. It is not quite clear from Quine's exposition, however, precisely how much emphasis is to be put on the word 'ontology' in his criterion of ontological commitment. My discussion which focuses on this word may thus have to be taken as a qualification to Quine's criterion rather than as outright criticism.

[14] Quine, *From a Logical Point of View*, pp. 130–132.
[15] For a discussion of the problems connected with the substitutivity principle, see my exchange with Føllesdal: Dagfinn Føllesdal, 'Knowledge, Identity, and Existence', *Theoria* **33** (1967) 1–27; Jaakko Hintikka, 'Existence and Identity in Epistemic Contexts', *ibid.* 138–147.
[16] See Quine, *From a Logical Point of View*, ch. 8; *Word and Object*, ch. 6; *The Ways of Paradox and Other Essays*, Random House, New York, 1966, chapters 13–15.
[17] Some arguments to this effect were given in *Knowledge and Belief* (ref. 11 above), pp. 142–146. The only informed criticism of this criticism that I have seen has been presented by R. L. Sleigh, in a paper entitled 'A Note on an Argument of Hintikka's', *Philosophical Studies* **18** (1967) 12–14. As I point out in my reply, 'Partially Transparent Senses of Knowing' (forthcoming), Sleigh's argument turns on an ambiguity in my original formulation which is easily repaired. Neither the ambiguity nor its elimination provides any solace to the adherents of the view I have criticized, however.
[18] One thing at which this old distinction aims is obviously the distinction (which I am about to explain) between statements about whoever or whatever meets a description and statements about the individual who in fact does so. For the distinction, cf. Jaakko Hintikka, 'Individuals, Possible Worlds, and Epistemic Logic', *Nous* **1** (1967) 32–62, especially 46–49, as well as '"Knowing Oneself" and Other Problems in Epistemic Logic', *Theoria* **32** (1966) 1–13.
[19] Cf. below (Section XII).
[20] Cf. here my paper, 'On the Logic of Perception' in *Perception and Personal Identity* (ed. by N. Care and R. Grimm), Case Western Reserve Univ. Press, Cleveland, 1969. (Reprinted in the present volume, pp. 151–183.)
[21] For temporal modalities, see e.g. A. N. Prior, *Past, Present and Future*, Clarendon Press, Oxford, 1967. – I am not saying that our actual methods of cross-identification in the case of temporal modalities (i.e. on ordinary methods of re-identification) turn on continuity quite as exclusively as I am about to suggest. It suffices for my purposes to present an example of methods of cross-identification that allows both 'branching' and 'merging', and it seems to me at least conceivable that temporal modalities might under suitable circumstances create such a situation.
[22] This is argued in 'On the Logic of Perception' (ref. 20 above).
[23] Cf. e.g. R. Carnap, *Meaning and Necessity*, University of Chicago Press, Chicago, 1947 (2nd ed.: 1956), pp. 41, 180–181, and Section VI *supra*.
[24] Cf. Gottlob Frege, 'Über Sinn und Bedeutung', *Zeitschrift für Philosophie und philosophische Kritik* **100** (1892) 25–50, especially p. 26, last few lines.
[25] Cf. 'On the Logic of Perception' (ref. 20 above).
[26] Views closely resembling some of those which I am putting forward here (and in some cases anticipating them) have been expressed by David Kaplan, Richard Montague, Dagfinn Føllesdal, Stig Kanger, Saul Kripke, and others. Here I am not trying to relate my own ideas to theirs. It is only fair, however, to emphasize my direct and indirect debts to these writers.