Notre Dame Journal of Formal Logic
Volume 23, Number 2, April 1982

Intention-Based Semantics

STEPHEN SCHIFFER*

We are physical objects in a physical world; our bodies, collections of molecules, move, and among the myriad products of these movements are marks and sounds. These physical phenomena have physical explanations, forthcoming, in principle, from the physical sciences, from physics, at the most fundamental level, to the neuro-biological sciences at more specialized levels. So much for the unassailable.

At the same time, we are apt, pretheoretically, to suppose that some of these marks and sounds have semantical properties, and that those who produce them have psychological states, notably beliefs, desires, and intentions. The sequence of marks 'Mitterand defeated Giscard', for example, is a *sentence* of a *language*, it has *meaning*, viz., that Mitterand defeated Giscard, it is *true*, it contains *names* that *refer* to people, and a *predicate* that is *true of* pairs of things. One producing this sequence of marks is not unlikely to *believe that Mitterand defeated Giscard*, and to *intend*, in producing those marks, to instill the same belief in another.

The subject matter of the philosophy of language, if it has one, is the nature of the semantical properties of linguistic items. But no complete account of those properties will leave unanswered these two questions:

(1) How is the semantic related to the psychological?
(2) How are the semantic and the psychological related to the physical?

I believe, for familiar reasons, later briefly to be touched on, that (2) is the urgent question, in this sense: that we should not be prepared to maintain that there *are* semantic or psychological facts unless we are prepared to maintain that such facts are completely determined by, are nothing over and above, physical facts. I am also inclined towards a reductionist response to

*I am grateful to Brian Loar for running conversations on these topics.

Received August 17, 1981; revised September 23, 1981

question (1), to seeing the semantic as reducible to the psychological; more exactly, but not yet to dwell on the distinctions implied, to seeing public language semantical properties as reducible to psychological properties, or, as in the case of a property like truth, to semantical properties of psychological states. These two reductionist tendencies are not unrelated, for I believe that *the only viable reduction of the semantic and the psychological to the physical is via the reduction of the semantic to the psychological.*

My topic is a research program in progress, which I shall call *intention-based semantics*, that was given its incipient expression in H. P. Grice's seminal article "Meaning" [17] and which argues for the reduction of the semantic to the psychological.[1] My purpose in this article is to do what I can, within the confines of these pages, to clarify the nature of this program, and to enhance its credibility.

1 Certain intention-theoretical writings have, unwittingly, tended to foster the misleading impression that the program was an exercise in conceptual analysis, the aim and the end of which was the definition of various ordinary language semantical idioms in terms of certain complexes of propositional attitudes, as if, having attained these definitions, there were no further questions to ponder. One seeing the project in this dim light was likely to be struck by the intention-theorist's accounting for the content of a sentence in terms of its being related in a certain way to a certain propositional attitude *with the same content*, as for example, the theorist would explain the fact that 'Mitterand defeated Giscard' means that *Mitterand defeated Giscard* in terms of its being conventionally correlated in a certain way with the belief that *Mitterand defeated Giscard*; this perception then likely to produce the complaint that the content of beliefs cannot be assumed to be unproblematic, itself in need of no explication.

In fact, the program need have no truck with conceptual analysis, and of course does not treat mental content as being unproblematic. The intention-theorist seeks to reduce the having of content of marks and sounds to the having of content of psychological states. Then, having reduced all questions about the semantical features of public language items to questions about mental content, he sees his task as having to answer those further questions, but free now to pursue those answers without any further appeal to public language semantical properties, which is perhaps the most *attractive* feature of this reductionist enterprise.

Possibly the least tendentious way of understanding intention-based semantics is as a certain claim about a certain nested sequence of *stipulative* definitions.

1. The first stipulative definition is of *speaker's meaning**, the asterisk a reminder that the expression is, in the first instance, merely being introduced as the label for a kind of act of which there may or may not be instances. Speaker's meaning* is identified by the theorist with a species of intentional behavior the specification of which involves nothing overtly semantical. Theorists will differ on details, but my own approach is first to define:

> *S* means* something in uttering *x* iff for some *p*, *S* means* in uttering *x* that *p*, or, for some person *A*, that *A* is to make it the case that *p*;

meaning* that *p* is then defined around the essential condition that:

> *S* means* that *p* in uttering *x* only if, for some person *A* and relation *R*, *S* utters *x* intending it to be mutual knowledge* between *S* and *A* that *xRp* and, on the basis of this, that *S* uttered *x* with the intention of activating in *A* the belief that *p*.

In the imperatival case, where *S* means* that *A* is to make it the case that *p*, the primary response aimed at is *A*'s making it the case that *p*.[2]

A few comments are in order.

(1) The above really represents the definition of speaker's meaning* as it applies to the paradigmatic case where the speaker has a particular audience in mind; the proper definition would entail no such quantification over audiences.

(2) The completed definition would impose a requirement on the way in which *S* must intend his utterance of *x* to activate in *A* the belief that *p*. Grice's original account required that this be achieved by way of *A*'s recognition of *S*'s intention to produce in *A* the belief that *p*, but that condition is better suited to an account of *telling*. My preference is to have the definition require merely that, for some relation R' (which may be the same as *R*), *S* intends his utterance to active *A*'s believing that *p* via *A*'s belief that $xR'p$.

(3) The success of communicative intentions depends on one's primary intention being entirely "out in the open", in a way that needs to be made precise, and mutual knowledge*, another stipulatively defined notion, is intended to accommodate that need (as well as others). In *Meaning* [39], '*x* and *y* mutually know* that *p*' was defined as the infinite continuation of the sequence '*x* knows that *p*, *y* knows that *p*, *x* knows that *y* knows that *p* . . . ', but there is a serious (but I think unanswered) question about the psychological reality of so strong a concept, and a weaker understanding of mutual knowledge*, such as that suggested by Bennett or by Loar,[3] seems preferable.

(4) As all other semantical* notions are to be defined in terms of speaker's meaning*, its definition must not itself presuppose anything semantical*. Consequently, *S*'s utterance *x* need not have meaning*, and the relation *R* which *S* intends to obtain between *x* and *p*, and to provide the evidential basis for the requisite mutual knowledge*, may be any relation which *S* thinks will get the job done. In a given case, for example, *x* might be the sound 'grrr', *p* the belief-object that *S* is angry, and *R* some complicated associational relation which obtains between the two by virtue of the fact that 'grrr' resembles the sound dogs make when they are angry. Clearly, however, some relations will do the job better than others, and, given that there are acts of speaker's meaning*, we should expect that

there is some relation R^* such that typically a speaker will utter x to mean* that p only if xR^*p, and this because the fact that xR^*p provides the best possible evidence that an utterance of x was produced with those intentions definitive of meaning* that p. Now, to anticipate a later elaboration, the intention-theorist's strategy as regards the definition of sentence-meaning* is: (a) to specify R^* in wholly psychological terms, and (b) to define sentence-meaning* in terms of R^*, thereby identifying a sentence's having a certain meaning* with its being a maximally efficacious device for performing acts of speaker's meaning* of a certain kind. The definition below of 'σ is a conventional device* in o for meaning* that p in a population G' will eventually emerge as the intention-theorist's candidate for R^*.

(5) The letter 'p' occurs in the definition as a quantified variable whose values are things believed. But what are they? This daunting question defines one of the issues of this paper, and is pursued in Section 4.

2. Also on the speaker side, and in terms of the definition of speaker's meaning*, it proves useful to define the notion of an illocutionary* act, and various notions of speaker's reference*.[4] Illocutionary* acts are directly definable in terms of speaker's meaning*, the different types of such acts–telling*, requesting*, warning*, etc.–distinguishable in terms of the constraints they impose on the values of one or another of the variables occurring in the definition of speaker's meaning*. For example, in uttering x, S was *ordering** A to make it the case that p just in case S meant*, in uttering x, that A was to make it the case that p, A's reason for compliance to be A's awareness that S stands in such and such a position of authority with respect to A. The leading concept of speaker's reference* is defined in terms of meaning* something *of* an object, and requires a prior treatment of *de re* propositional attitudes. These ancillary agent semantical constructs find their main employment in accounting for the semantics of various particular kinds of expressions.

3. Various ways in which behavior may be *conventional** are definable, as various kinds of self-perpetuating regularities in behavior, which definitions, like that of speaker's meaning*, are wholly in terms of propositional attitudes which overtly presuppose nothing semantical or linguistic (cf. [22], [27], and [39], Ch. 5). Especially pertinent to matters at hand is the notion of something's being a conventional device* for meaning* something. Letting 'σ' range over utterance-types and 'o' over possible occasions of their utterance, we may say, quite roughly, that:

σ is a conventional device* in o for meaning* that p in a group or population G iff

(a) it is practicable for a member of G to mean* that p in o by uttering σ, if it is possible for him to mean* that p in o,

and this just by virtue of its being the case that

(b) it is mutual knowledge* in G that (a).

The sentence 'I have no nose' is, it might be suggested, a conventional device* among speakers of English for meaning* that one has no nose; but it requires a very special set of circumstances in order for it to be possible for one to mean* that one has no nose. Further, if there is a context which renders such an act possible, then we do not want to require that it be performed by uttering the words 'I have no nose', but only that the act of meaning* *can* be performed, in the context, by an utterance of those words. Hence, the form of condition (a). Condition (b) captures the conventional* aspect of meaning* that p in o by uttering σ: the relevant feature of σ by virtue of which one may in uttering σ mean* that p *is just the fact* that it is mutually known* in G that one can mean that p by uttering σ.[5]

4. This brings us finally to *public language**, and the intuitive idea to be captured here is that a system of marks or sounds is a *public language** provided that it is a conventional* system for performing acts of speaker's meaning*; and that to know the meaning* of any item, or sequence of items, of that system is just to know its role in the system.

Now a neat way of being more precise is achieved by a selective mimicking of Lewis [27]. Let us, in the first instance, pretend that the only languages we are concerned to capture contain indexical expressions, but no ambiguous expressions or constructions, and no mood other than the indicative. Then we would first define a *language** as any function L from strings of marks or sounds σ—the complete sentences of L—and possible occasions o of their utterance to objects of belief, and next stipulatively define:

> A language* L is a *public language** *of a population* G iff $(\sigma)(o)(\exists p)\sigma$ is a conventional device* in G for meaning* that p in o if $L(\sigma,o) = p$.

Mood is an abstraction from syntactical form the semantical significance of which resides in its correlation with a certain kind of speech act. To allow for various moods in an unambiguous language we should define the language* function as having in its range ordered pairs of the form $\langle\Psi,p\rangle$, where Ψ is a kind of illocutionary* act (definable, it will be recalled, as a species of speaker's meaning*). To accommodate ambiguity, the function should be to sets of such ordered pairs, and in the end we should say—letting 'A' range over such sets—that:

> A language* L is *a public language** *of a population* G iff $(\sigma)(o)(\exists\Psi)(\exists p)(\exists A)\sigma$ is a conventional device* in G for Ψing that p in o if: (1) $\langle\Psi,p\rangle \in A$ and (2) $L(\sigma,o) = A$.

Then the meaning* of σ in L, and in G, may be identified with that function M_σ, determined by L, such that, for each possible occasion o of the utterance of σ, $M_\sigma(o) = L(\sigma,o)$, other semantical* properties—refers* to, is true*, is true of*, etc.—definable in a similar vein.

As stipulative definitions the foregoing definitions are indefectible; after all, they might not be realized. Intention-based semantics is the claim that

> public language semantical properties *are* semantical* properties, and that, in this way, those semantical properties *reduce* to psychological properties.

(By a 'public language semantical property' I mean any property of a linguistic item which entails that it has meaning in a public, communicative language; the reason for so much precision in formulation presently to be clarified.)

A few smaller claims may help to make the plausibility of this big claim more readily evident.

> 1. *People typically, or often, perform acts of speaker's meaning* when they speak.*

The claim here is not that speaker's meaning* defines some antecedently clear concept of speaker's meaning, nor that language is not used properly or literally unless something is meant*; merely that it is not uncommon for a speaker to mean* something when he or she utters a complete sentence.

If there is a problem with this claim it is that the definition of speaker's meaning*, owing to the complexity of intention it requires of a speaker, lacks psychological reality. Benson Mates, in reference to Grice's definition of speaker's meaning, once *assured* me that, when *he* spoke, he did not have all of those intentions, and I have heard the sentiment of this remark echoed in the more subtle criticisms of other philosophers. The issue may in fact be joined over Grice's definition, for it does come close to what I should accept as a definition of *telling**, no doubt the central species of speaker's meaning*.

For S to tell* A that p in uttering x is, essentially, for S to utter x with the primary intention of informing A that p, further intending this to be achieved via A's believing: (1) that x is related in a certain way to p, and, at least partly on the basis of this, (2) that S uttered x with the intention of informing A that p, and, at least partly on the basis of this, (3) that S believes that p, and, at least partly on the basis of this, (4) that p.

The question is: Do speakers have such complicated intentions?

Few nowadays would be seduced by the super-Cartesian presupposition of the implicit Matesian argument, and yet, I am not aware of any better *specific* reason for doubting that speakers have such intentions. On the other side, something can be said in support of the psychological reality of speaker's meaning*.

Suppose that S did x with the primary intention of thereby bringing about an effect of kind E, i.e., that that was his motive or reason for doing x. In the event, S will have desired the occurrence of an E-event, and believed, to some nonnegligible degree, that his doing an act of a certain kind K, to which he believed x to belong, would cause an event of kind E; and his action will have resulted from this belief and desire. Now it might also have been the case that S had certain beliefs as to how his doing an act of kind K would cause an event of kind E; perhaps he believed that such an act would cause an E-event only by first causing an event of a certain kind E'. This belief, too, would figure into the explanation of S's action, in that, had he lacked the belief that his action would cause an event of kind E', and all other things been equal, then he would not have done what he did. Here we may say–at least to give the phenomenon a label–that S did x with the primary intention of bringing about an effect of kind E, *also intending* to bring about an effect of kind E', and intending the E'-event to cause the E-event.

For example, a dog threatened to menace Janet's garden. She smacked her hand against a book, with the intention of thereby getting rid of the dog. For she believed that her smacking the book would produce a sharp noise, which would startle the dog and cause it to run off. Of course, none of those thoughts occurred consciously to Janet, raced, as it were, through her mind. She simply had the book in her hand, saw the dog, and acted spontaneously, without conscious thought. Nevertheless, she had these beliefs and they, together with her desire to get the dog out of her garden, explain her action. I describe this in my perhaps philosophically tainted idiolect by saying that Janet's primary intention in smacking the book was to get rid of the dog, but that she also intended to produce a sharp noise, intended the noise to startle the dog, and intended the dog's being startled to cause it to run off. But nothing whatever hangs on the word 'intend', and, as regards the definition of speaker's meaning*, I am perfectly content to replace all references to secondary intentions by the indicated belief locutions. The issue, then, concerning the psychological reality of speaker's meaning* amounts just to this: Is it plausible to suppose that speakers typically have the relevant beliefs about their audiences?

I believe that it is plausible, that we have ever so many beliefs about communication which enter into, and explain, our communicative acts, but that, computer-like, we calculate instantaneously, unconsciously, and are not ready articulators of these mental processes.

Consider now a typical case of *telling*, with an eye toward discerning if it is also a case of telling*.

Emily knows that it has stopped raining, knows that Toby does not know this but would want to, and says 'It's stopped raining' with the intention of imparting this knowledge to him. She has told him that it has stopped raining, but has she told* him this? Surely, Emily believes that Toby will recognize that an utterance of 'It's stopped raining' is for them a reliable device for conveying the information that it has stopped raining and that, at least partly as a result of this, he will believe that she uttered the sentence intending to inform him that it had stopped raining. For in the circumstances she clearly would not have uttered the sentence had she not had both beliefs, had she thought, for example, that Toby did not understand English, or that thinking that she was teasing him, he would fail to recognize her intention to inform him. Furthermore, to continue, Emily believes that Toby will believe—at least partly on the basis of his recognition of her intention—that she believes that it has stopped raining, and, partly as a result of this (for he will know that, in the circumstances, she would be very unlikely to have this belief unless it had stopped raining) that it has indeed stopped raining. Now it is evident that Emily believes that Toby's belief that it has stopped raining will be based on his believing that she believes this; for otherwise we shall have no way of accounting for the fact that she would not have spoken had she thought that Toby would take her to be lying. Likewise, the fact that she again would not have spoken had she thought that he would not take her intention to be to inform him is explained by her knowledge that, in the circumstances, Toby will believe that she believes that it has stopped raining only if he thinks that she uttered the sentence with the intention of producing that belief in him.

Can there really be any serious doubt as to whether such a speaker would have such beliefs? How else can we plausibly account for Emily's belief that her utterance would cause Toby to believe that it had stopped raining?[6]

2. *Every public, spoken or written language is in fact a public language**.

The intention-theorist's account of speaker's meaning* is *explanatory* of his account of utterance-meaning*, in that, given that there are acts of speaker's meaning*, together with certain commonplaces about human rationality and the human condition, it can be deduced that there will inevitably come to be utterance-types with meaning*–that is to say, conventional devices* for performing acts of speaker's meaning*; for having the properties circumscribed by these labels makes an utterance-type maximally efficacious as a means for performing acts of speaker's meaning* (cf. [39], Ch. 5, and [43], Sec. IV). Hence, if, as I have argued, there are acts of speaker's meaning*, then we should expect there to be conventional devices* for performing acts of speaker's meaning*. What might they be? Natural language sentences come to mind, and well they should. Let us say that a sentence σ and a belief-object p are *M*-compatible if one speaking literally can assert that p in uttering σ. I submit that, refinements aside, if σ and p are *M*-compatible, then σ is a conventional device* for meaning* that p (and likewise, *mutatis mutandis*, for nonindicative sentences).

The sentence 'She is playing the flute' is *M*-compatible with *Tanya is playing the flute.* In uttering the sentence Jim might be telling* Susan that Tanya is playing the flute, and if there is no difficulty in supposing that there are acts of speaker's meaning*, then there can be no difficulty in supposing that it was, even prior to Jim's utterance, mutually believed* by him and Susan that an utterance of 'She is playing the flute' would in the circumstances be a practicable means for telling* Susan that Tanya was playing the flute, and that that is precisely why Jim uttered the sentence to tell* Susan that. How else could Jim have hoped to tell* Susan that Tanya was playing the flute by uttering 'She is playing the flute'?

But if public language sentences are conventional devices* for meaning* that with which they are *M*-compatible, then it follows trivially that spoken and written natural languages are in fact public languages*.

3. *Semantical* properties weakly define the public language semantical properties they entail.*

In the first place, the semantical* properties of linguistic items entail their corresponding semantical properties. Necessarily, if L is a public language*, then L is a language; if σ means* p, then σ means p; if t refers* to x, then t refers to x; and so on. The psychological properties bearing semantical* labels may not, contrary to what I believe, be instantiated, but *surely* it is impossible to conceive of a possible world in which a system of marks or sounds is a public language* but not a language.

I do not, however, claim that all semantical properties entail semantical* properties. I can easily imagine a world like ours but for the fact that strings indistinguishable from English sentences are never produced for communicative purposes of any sort, but only for "thinking on paper," and I should not jib at ascribing semantical properties to those strings. Nor do I object to calling formal languages, uninterpreted or interpreted, languages, and am at ease with talk, however unprototypical, of formulas having meaning in a language of thought. These are, in the first instance, verbal matters, and it cannot be assumed that, in themselves, they carry philosophical import.

At the same time I do want to claim that the *public language semantical properties* of linguistic items entail their corresponding semantical* properties. A public language semantical property, I have stipulated, is any property which entails that a bearer of it has meaning in a public language used, *inter alia*, for communication, in the sense in which we humans typically communicate with one another. This is not a trivial claim, even granting the communicative* nature of human communication. For it means, in effect, that there is no single, nondisjunctive "meaning property" common to whatever items might, in one sense or another, be said to have meaning; no univocal sense in which items may represent the world—which of course suits the intention-theorist, whose claim it will soon be that the representational properties of public language items *reduce to* the representational properties of mental (i.e., neural) states.

Although not trivial, the claim that public language semantical properties entail their corresponding semantical* properties is, I believe, immensely plausible. For typical human communicative behavior is communicative* behavior, which entails that the devices used for such behavior, public language sentences, are conventional devices* for performing acts of speaker's meaning*, and so have meaning* and similar semantical* properties; and from all this the claim in question follows.

I say that Ψ *weakly defines* Φ provided that, necessarily, something is Φ iff it is Ψ. I call this sense of definability *weak* because it does not follow from the fact that Ψ weakly defines Φ that Φ is *identical* to Ψ; they may be quite distinct but necessarily co-extensional properties, an important point should a purported reduction be in the offing.

Now it follows from the foregoing that public language semantical properties are at least weakly defined by their semantical* equivalents.[7] But this does not yet yield the claim which defines intention-based semantics—namely, that:

4. *Public language semantical properties are identical to their semantical* equivalents, and thus reduce to psychological properties.*

Semantical* properties are, by stipulative definition, psychological properties, and they weakly define the semantical properties of public language items, qua items in a public language. But from the fact that semantical properties are weakly definable in terms of psychological properties, it does not follow that they are reducible to them. It is consistent with the truth of the modal biconditionals which are these definitions that, for example, they characterize relations between quite distinct properties of equal conceptual status, neither of which is reducible to the other, nor explicable without reference to the other. Yet the weak definability of meaning in terms of thought, without the reducibility of meaning to thought, is barely of passing interest, a curious fact in need of explanation, certainly no account of what meaning *is*. Intention-based semantics is precisely the claim that public language semantical properties *are* semantical* properties, and in this way reduce to the psychological.

There is, I believe, but one ground for resisting the reduction thesis given weak definability, and that is the suspicion, or stronger sentiment, that propositional attitudes cannot themselves be explicated, or understood, without reference to the public language semantical properties of linguistic items. To all intents and purposes, intention-based semantics is the thesis that

(a) public language semantical properties are weakly definable in terms of semantical* properties,

and that

(b) propositional attitudes are explicable without any recourse to those properties;

for, if (a) and (b) are true, it will then be undeniable that public language semantical properties *are* psychological properties. The remainder of this paper is about (b), about what it comes to, what motivates it, and, especially, what makes it plausible.

2 Intention-based semantics, claiming, as it does that, public language semantical properties are identical to complex psychological properties defined in terms of belief, desire, and intention, entails, on pain of a vicious circularity, that those psychological states are not themselves to be accounted for, even partly, in terms of those semantical properties. The best way to argue for this entailed claim would be actually to present a complete and credible theory of belief (and the other attitudes) which had the desired property. However, notwithstanding the existence of at least one attractive candidate (cf. [32], and Sections 4 and 5 below), I am not in a position to proffer such a theory (the difficulty, later to be mentioned, in no way the result of the need for a semantically untainted theory). So I shall try to do the next best thing—namely, to give reasons for supposing that the correct theory of belief, whatever it turns out to be, will be as the intention-theorist requires it to be.

A few preliminaries will be useful before getting down to my case for thinking that thought is explicable independently of meaning.

(a) To simplify the discussion, I shall nearly always address the issue in terms of whether *belief* can be explicated independently of public language semantical properties, it being understood that what applies to belief also applies, *mutatis mutandis*, to desire, and to any other propositional attitudes not reducible to belief and desire (and often, again for simplicity, I shall write as if intention were definable in terms of belief and desire).

(b) 'Public language semantical property' is fast becoming unwieldy; so let us replace it with 'semantical$_o$ property', the subscript '*o*' for '*outer*, public language' later to provide a useful contrast to the subscript '*i*' in talk of the 'semantical$_i$ properties of formulas in the *inner* language of thought'.

(c) Our issue is whether or not belief is explicable without recourse to semantical$_o$ properties. But that, too, is unwieldy, especially as it will prove convenient to state a condition for belief not being explicable independently of semantical$_o$ properties, and I shall use the phrase 'belief is semantic$_o$ dependent' as shorthand for the tortuous phrase.

(d) It would be naive to suppose that this issue was a priori precise, and a mistake to try to make it so. Nevertheless, the issue can become precise when presented relative to certain assumptions, and my intention is to argue from assumptions–functionalist ones, to glance ahead–such that, relative to them, there is only one clear thing that could be meant by the claim that belief was (or was not) semantic$_o$ dependent.

(e) Finally, and again for convenience, I shall for a while assume that we really do have beliefs with determinate content, and that, therefore, belief-desire explanations of behavior are sometimes literally true. A note of pessimism on this score will not be sounded until the final pages.

My case for thinking that belief is semantic$_o$ *in*dependent is a two-stage enterprise. The first, were I to spell it all out, would consist in arguing for, and explaining, the claim, stated now without refinement, that:

(1) Belief is semantic$_o$ dependent only if there is a theory T such that belief and some semantical$_o$ construct are correctly explicable only via their theoretical definitions with respect to T.

The second stage would consist in showing that:

(2) For a variety of reasons, it is not plausible to suppose that there is a theory which satisfies the condition of (1).

It will now be seen that one renders (1) acceptable by showing (refinements again aside) that:

(i) Belief cannot be construed as an irreducibly mental property or relation.

(ii) The only plausible nonmentalist account of belief, which makes determinate sense of it, is (what may be called) an *extended functionalist* account, which requires that there be a theory T such that, for each p in the domain of T, there is an extended functional role F, determined, *inter alia*, by the role that the construct belief plays according to the generalizations of T, such that one believes that p just in case one is in a state of a type which has F.

(iii) Whether or not semantical$_o$ properties reduce to psychological properties, it is at least clear that they cannot be explicated without reference to them.

Now virtually nothing needs to be said about (iii); it is obvious. The fact that the sequence of marks

Brooke Shields doesn't suffer from acne

has the semantical$_o$ properties it happens contingently to have has, essentially, *something* to do with the beliefs and intentions of those who understand this sentence, and there can be no hope of explicating those semantical$_o$ properties without reference to those psychological states.

Evidently, the case for (i) still needs to be made, although it must perforce be mostly assumed in the present context. Still, it might clear the air a little if I merely note, without arguing for it, the basis of my own materialist *parti pris.*

In the first place, I believe that:

(a) For every bodily movement *m*, there is a complete physical explanation, which provides causally necessary and sufficient conditions for the occurrence of *m*.

In other (but no more precise) words, there is within the agent an unbroken causal chain of events such that, in the typical case, each event in the chain is causally necessary and sufficient for the occurrence of the movement *m*; and the properties of these events by virtue of which they are subsumable under the applicable causal laws are wholly physical properties. Needless to say, this is a bet, as no one can give these causal explanations, but, I should think, a reasonable bet, based as it is on the well-known fact that we are not hollow inside, but house billions of neurons sending forth and receiving electrochemical impulses, from receptor cells to effector cells, trillions of times each instant, from even before the moment of one's birth until the moment of one's death.

I further believe that:

(b) Beliefs—or the events of coming to have them—are causes of bodily movements.

For example, when I step back to the curb because I see that a car is coming, my coming to believe that a car is coming is a cause of my stepping back. Moreover:

(c) The property of being a belief is a causally necessary property of those beliefs which are, in the way typical of beliefs, causes of behavior.

Consider that state of mine which was my belief that a car was coming: had that state not been a belief it would not have been a cause of my stepping back. (The palpable plausibility of this point is evidently overlooked by those who are attracted to a *Verstehen* construal of belief-desire explanations of behavior.) And, finally, I also believe that:

(d) One's beliefs are supervenient on physical reality, in that there cannot be two possible worlds which are physically alike, in one of which, but not the other, one has a given belief.

In devising an account of belief that is consonant with (a)-(d) it is not enough merely to suppose that particular beliefs are physical states; one needs to account for what it is about a particular physical state-token that makes it a belief. The reason that I cannot believe that *being a belief* is an irreducible property—that is to say, that that property is not identical to some property intrinsically specifiable in a nonmentalistic idiom—is that, quite simply, *I* can make no coherent sense of that hypothesis given (a)-(d).

This brings me to the middle claim, (ii), which is also in need of elaboration.

To begin, let '*B*'' denote the property expressed by the open sentence

x is a belief that snow is white;

i.e., the property of being a belief that snow is white, which property, and all others like it, I will call a *belief property*. (To be sure, this is a composite property, derived from the relation expressed by the two-place open sentence 'x is a belief that p', but that is not something we need yet to dwell on.) The states which have belief properties are, we had better suppose, physical states, states of the central nervous system, tokens of neural state-types. Suppose, then, that s is my present belief that snow is white. Then the question arises: What makes s a belief that snow is white? Which is to ask: In what does its having the property B' consist?

Were we mentalists with respect to belief properties, we should refuse the question, protesting its implicit presupposition that it has a correct reductionist response, and insist that what makes s a belief that snow is white is, quite simply, that it stands in the belief relation to the belief object denoted by 'that snow is white', that relation itself primitive and irreducible. Rejecting this mentalist response, I am committed to finding a property Φ, however complex or composite, which is not intrinsically psychological, such that $B' = \Phi$.

A *functional role* is a second-order property of first-order state-types. Having a functional role is a matter of being causally or counterfactually related in a certain way to sensory inputs, other internal states, and behavioral output. It is plausible to suppose that there is a functional role F such that my state s (which happens to be my present belief that snow is white) is *a belief* by virtue of its being a token of some neural state-type that has F. But is it also plausible to suppose that there is a functional role F' which entails, but is not entailed by, F, the functional role common to all beliefs, such that s is a belief *that snow is white* by virtue of being a token of a neural state-type which is F'? That depends on how we understand the notion of a functional role, none too precisely characterized above, or in the literature.

A *narrow* functional role, I shall say, is any understanding of functional role which would make one functionally equivalent to any cell for cell duplicate of oneself. It is, I believe, doubtful that there is any *narrow* functional role F' such that s has B' by virtue of being a token of a state-type which has F'. For, to mention the most evident of reasons, it is plausible to suppose that my Twin Earth *Doppelgänger* does not believe that *snow* is white, as the stuff he calls 'snow' is not formed by the sublimation of water (i.e., H_2O) vapor, but of the notorious XYZ vapor.

By an *extended* functional role I shall mean any property which entails a narrow functional role (thus every narrow functional role is an extended functional role). If F is an extended functional role, then it may be that one is in a state of a type which has F only if one is in a state the causal ancestry of which includes objects of a certain kind; perhaps F entails causal relations regarding the *acquisition* of tokens of state-types having the narrow functional role entailed by F.[8] My Twin Earth duplicate will be a narrow functional equivalent of mine, but not an extended functional equivalent.

I submit that the only viable way of being a materialist, relative to the assumption that there are determinate belief properties, is to maintain that, for every p, there is an extended functional role F, such that the belief property

being a belief that p

is *identical to* (what may be called) the *extended functional property*

being a token of a type which has F.[9]

But what determines the extended functional property with which a given belief property is identical? Here is where we need recourse to the determination of an extended functional property by some theory in which the construct belief occurs. Simplifying somewhat, and deviating only a little from well-trodden paths (cf. [38], [23]-[25], [13], and [32]), I shall briefly explain this as follows.

First, let T be a theory, and the predicates $A_1, \ldots, A_n$ those terms occurring in T–T's theoretical terms–whose extensions are determined by their roles in T, and let

$$T(A_1, \ldots, A_n)$$

be T written out as a single sentence. Then *the theoretical definition of A_i with respect to T* (supposing A_i to be a k-ary predicate) is:

$$A_i x_1, \ldots, x_k =_{df} (\exists \Phi_1, \ldots, \Phi_n)(T(\Phi_1, \ldots, \Phi_n) \wedge \Phi_i x_1, \ldots, x_k).^{10}$$

In effect, this says that A_i stands for the property (or relation) of having some property (or relation) which is the i^{th} member of an n-ary sequence of properties and/or relations which satisfies the n-place open sentence

$$T(\Phi_1, \ldots, \Phi_n).$$

Suppose, for example, that two one-place predicates A and B are the only terms defined by a theory T', written as the sentence

$$T'(A,B).$$

Then the theoretical definitions of A and B with respect to T' are:

(i) $Ax =_{df} (\exists \Phi)(\exists \Psi)(T'(\Phi, \Psi) \wedge \Phi x)$.
(ii) $Bx =_{df} (\exists \Phi)(\exists \Psi)(T'(\Phi, \Psi) \wedge \Psi x)$.

Thus, something satisfies A just in case it has some property which is the first member of an ordered pair which satisfies the open sentence

$$T'(\Phi, \Psi),$$

and something satisfies B just in case it has some property which is the second member of such an ordered pair.

Equivalently, one may view the matter thus. The one-place open sentence

(iii) $(\exists \Psi) T'(\Phi, \Psi)$

expresses a second-order property, the property which a first-order property has just in case it conforms to the role determined for A by the generalizations of T'. If T' is a psychological theory, and A a psychological predicate, then this property might be a functional role. Then the open sentence which defines A,

(iv) $(\exists\Phi)(\exists\Psi)(T'(\Phi,\Psi) \wedge \Phi x)$

expresses the property of having some property which satisfies (iii). Consequently, if the property expressed by (iii) is a functional role, then that expressed by (iv) is a functional property. In the event, the theoretical definition of A with respect to T' says, in effect, that A stands for that functional property.

If belief properties are extended functional properties, then, subject to a certain qualification, there is some theory T which theoretically defines *belief* in tandem with other theoretical primitives of T, such that T indirectly determines, for each belief property, the extended functional property with which it is identical. The determination will be indirect, of course, because, e.g., 'is a belief that snow is white' will not be a primitive of T (a matter to which I shall later return). The qualification alluded to, one safely ignored for the immediate exposition, is that T may not actually contain any single construct *belief*, but may theoretically define other constructs in terms of which belief may, from without the theory, be defined.[11]

There is something *prima facie* puzzling in the idea that there might be two concepts, neither of which reduces to the other and neither of which is explicable independently of the other. And yet, there do seem to be such pairs of concepts, *belief* and *desire* no doubt being the most illustrious example.

And indeed there is no puzzle, if we think of the two concepts as being explicated in tandem via their theoretical definitions with respect to some theory in which they both occur. The point is in fact nicely illustrated in the recent example of the theoretical definitions of A and B with respect to T'. For, had either been definable in terms of the other, it would not have been a theoretical primitive of T'; at the same time, neither can be defined independently of the other, as they are both defined with respect to their interlocking roles in T'; definitions (i) and (ii) providing the resolution.

Our issue is whether belief can be explicated independently of any semantical$_o$ notion, and we are assuming: (a) that semantical$_o$ notions cannot be explicated independently of belief, and (b) that belief can be explicated—that is, functionally reduced—only via its theoretical definition with respect to some theory. Relative to (a) and (b), the claim that belief is semantic$_o$ dependent, that it cannot be explicated without recourse to the semantic$_o$, can amount only to this: that there is some semantical$_o$ construct M and some theory T such that the extensions of M and the construct *belief* can be determined only via their theoretical definitions with respect to T; that, in other words, the explications of meaning$_o$ and belief must proceed in tandem, via their theoretical definitions with respect to some psycho-semantical$_o$ theory in which they occur as interlocking constructs; that meaning$_o$ is to belief, from a functionalist point of view, as belief and desire are to one another.

I do not believe that there is any theory with respect to which belief and some semantical$_{o}$ construct are (as I shall say) co-definable, and therefore I believe that belief is, as the intention-theorist claims, semantic$_{o}$ independent. I have two sorts of reasons for this position. One is based on an argument from the weak definability of the semantic$_{o}$ in terms of the psychological, while the other proceeds from reflection on the sorts of theories which might provide the wherewithal for a functionalist reduction of belief. I turn now to the general argument, the next section to broach considerations of the second sort.

I have already argued for the weak definability of the semantic$_{o}$ in terms of the psychological, but have also admitted that that does not entail the reduction of the former to the latter. The upshot of the considerations I am presently to adduce is that, to all intents and purposes, weak definability *plus (extended) functionalism* does yield the reduction of the semantic$_{o}$ to the psychological. My idea is to show that, if weak definability is true, then there is no theory and no semantical$_{o}$ construct such that belief is co-definable with that construct with respect to that theory. Given (extended) functionalism, this means that belief is semantic$_{o}$ independent, and, as we have seen, the semantic$_{o}$ independence of belief together with the weak definability thesis is tantamount to intention-based semantics.

There is something a little vulgar about plunging straight off into an argument that lays claim to a certain degree of rigor, so let me set the proper mood by beginning with a fairly simple and intuitive point.

Let T be any theory in which one or more semantical$_{o}$ terms commingle with 'believes', 'desires', and 'intends', and let T^* be that theory that is formed from T by replacing all occurrences of semantical$_{o}$ terms in T with their intention-theoretical definientia. Assume that the semantic$_{o}$ is weakly defined by the psychological in the way argued for. Then T and T^* are logically equivalent. But what objective, determinate sense could it now make to suppose that belief is explicable only via its theoretical definition with respect to T? Since T^* is by hypothesis logically equivalent to T, surely it would do just as well for determining the functional reduction of belief. But T^*, by construction, contains nothing semantic$_{o}$, and, as T is any arbitrary psycho-semantical$_{o}$ theory, this strongly suggests that, if weak definability obtains, then the functional theory which defines belief–assuming that there is such a theory–will contain no semantical$_{o}$ constructs.

There is, however, as I have suggested, a more rigorous route to this conclusion.

Once again, let T be any theory which contains some semantical$_{o}$ construct along with belief and the rest, and T^* that theory obtained from T by replacing T's semantical$_{o}$ terms with their intention-theoretical definientia. To simplify matters a little it will be useful to pretend that the one-place predicate M is T's only semantical$_{o}$ term, and that B and D are one-place predicates representing belief and desire (realistically, of course, they would be represented as relational predicates), the only psychological constructs occurring in T and T^* (thereby further pretending–or is it assuming?–that intention reduces to belief and desire). Then

(α) $T(B,D,M)$

and

(β) $T^*(B,D)$

may represent T and T^* written out as single sentences.

The conclusion I wish to reach is that:

(†) If weak definability obtains, then it is not the case that B, D, and M are theoretically definable with respect to T,

which, as T represents any arbitrary psycho-semantical$_o$ theory, would show that, given weak definability, belief is not theoretically co-definable with any semantical$_o$ construct. To this end we may begin with the following argument:

(1) If weak definability obtains, then T and T^* are logically equivalent.
(2) But if B, D, and M are theoretically definable with respect to T, then T and T^* are not logically equivalent.
(3) Therefore, (†).

This argument is valid, and (1) is obviously true. It remains to show that (2) is true.

Now, as I have defined 'the theoretical definition of A_i with respect to $T(A_1, \ldots, A_n)$', it would follow that the *presumptive* theoretical definitions of B, D, and M with respect to T—that is, the definitions that would define them were they theoretically defined with respect to T—are, respectively:

(i) $Bx =_{df} (\exists\Phi)(\exists\Psi)(\exists\Omega)(T(\Phi,\Psi,\Omega) \wedge \Phi x)$.

(I.e., belief is the property of having some property that is the first member of an ordered triple of properties which satisfies $\ulcorner T(\Phi,\Psi,\Omega)\urcorner$.)

(ii) $Dx =_{df} (\exists\Phi)(\exists\Psi)(\exists\Omega)(T(\Phi,\Psi,\Omega) \wedge \Psi x)$.
(iii) $Mx =_{df} (\exists\Phi)(\exists\Psi)(\exists\Omega)(T(\Phi,\Psi,\Omega) \wedge \Omega x)$.

Thus, the proper demonstration of premise (2) would show that, if B, D, and M are defined by (i)-(iii), then T and T^* are not logically equivalent. However, it turns out that, for formal reasons, the proper proof, while perfectly available, is a little complicated. Consequently, I shall first prove, not (2), but (2′), which differs from (2) only in a respect that is irrelevant to the issue at hand, and then sketch how the proof of (2) will resemble that of (2′). The replacement premise is

(2′) If B, D, and M are L-theoretically definable with respect to T, then T and T^* are not logically equivalent,

where an 'L-theoretical definition' abbreviates a 'Lewis-style theoretical definition' (cf [23]-[25]). The L-theoretical definition of A_i with respect to $T'(A_1, \ldots, A_n)$ in effect construes the predicate A_i as a property-name whose reference is fixed by a contingent description, roughly of the form 'the property that has such and such functional role'. That is to say, more exactly,

$A_i =_{df}$ the i^{th} member of a unique realization of $\ulcorner T'(\Phi_1, \ldots, \Phi_n)\urcorner$, so that A_i refers to a property or relation P just in case: (a) P is the i^{th} member of an n-ary sequence of properties and/or relations which satisfies $\ulcorner T'(\Phi_1, \ldots, \Phi_n)\urcorner$, and (b) there is no other n-ary sequence which satisfies that open sentence. The problematical feature of Lewis-style definitions is, of course, that, as they require a *unique* realization, some pretty fancy maneuvering is needed to accommodate the possibility of psychological properties being multiply realized–i.e., realized in different creatures by different first-order physical properties–which problem, as I have remarked, is wholly irrelevant to the present concern (cf. [28] for an example of the fancy maneuvering).

Now, the L-theoretical definitions of B, D, and M with respect to T are:

(L-i) $B =_{df}$ the property which is the first member of the unique ordered triple of properties which satisfies $\ulcorner T(\Phi, \Psi, \Omega)\urcorner$,[12]

and thus refers to whatever property that property should turn out to be, if there is such a property.

(L-ii) $D =_{df}$ the property which is the second member of the unique ordered triple of properties which satisfies $\ulcorner T(\Phi, \Psi, \Omega)\urcorner$.

(L-iii) $M =_{df}$ the property which is the third member of the unique ordered triple of properties which satisfies $\ulcorner T(\Phi, \Psi, \Omega)\urcorner$.

Putting this aside momentarily, let us next observe that, whether or not (α), i.e., $T(B,D,M)$, and (β), i.e., $T^*(B,D)$, are logically equivalent, the Ramsey-sentence of (α),

(a) $(\exists\Phi)(\exists\Psi)(\exists\Omega)T(\Phi, \Psi, \Omega)$,

does not entail the Ramsey-sentence of (β),

(b) $(\exists\Phi)(\exists\Psi)T^*(\Phi, \Psi)$.

The reason that (a) does not entail (b) even if (α) entails (β) is, of course, that, if (α) does entail (β), the entailment obtains by virtue of the weak definability of M in terms of B and D, which gets lost when we existentially generalize on those terms. (Similarly, although 'Every eye doctor is rich and every oculist is happy' entails 'Every eye doctor is rich and every eye doctor is happy', it is not the case that '$(\exists\Phi)(\exists\Psi)$ (every Φ is rich and every Ψ is happy)' entails '$(\exists\Phi)$ (every Φ is rich and every Φ is happy)'.)

It is now easy to prove (2′). Assume that B, D, and M are defined by (L-i)-(L-iii), respectively. Then

$T(B,D,M)$ iff there is a triple $\langle P_1, P_2, P_3\rangle$ which uniquely satisfies $\ulcorner T(\Phi, \Psi, \Omega)\urcorner$,

and

$T^*(B,D)$ iff there is a triple $\langle P_1, P_2, P_3\rangle$ such that:
(a) $\langle P_1, P_2, P_3\rangle$ uniquely satisfies $\ulcorner T(\Phi, \Psi, \Omega)\urcorner$ and
(b) $\langle P_1, P_2\rangle$ satisfies $\ulcorner T^*(\Phi, \Psi)\urcorner$.

But, as we have already observed, the fact that some triple satisfies $\ulcorner T(\Phi, \Psi, \Omega)\urcorner$ does not entail that *any* couple satisfies $\ulcorner T^*(\Phi, \Psi)\urcorner$. Consequently, if B, D, and M are L-theoretically definable with respect to T–i.e., defined by (L-i)-(L-iii)–then (α) does not entail (β), and therefore (2′) is true.

The proof of (2), like that of (2′), is based on the evident nonequivalence of the Ramsey-sentences (a) and (b), and the strategy, similar again, is to show that, if B, D, and M are defined by (i)-(iii), then (α) does not entail (β), because the fact that $\langle P_1, P_2, P_3\rangle$ satisfies $\ulcorner T(\Phi,\Psi,\Omega)\urcorner$ does not entail that $\langle P_1, P_2\rangle$ satisfies $\ulcorner T^*(\Phi,\Psi)\urcorner$. The slight complication (no more than that), best relegated to a footnote, is in showing that, if B, D, and M are defined by (i)-(iii), then

$T^*(B,D)$ iff there is some triple $\langle P_1, P_2, P_3\rangle$ such that
(a) $\langle P_1, P_2, P_3\rangle$ satisfies $\ulcorner T(\Phi,\Psi,\Omega)\urcorner$ and
(b) $\langle P_1, P_2\rangle$ satisfies $\ulcorner T^*(\Phi,\Psi)\urcorner$.[13]

So I conclude that, if weak definability obtains, then belief is not theoretically co-definable with any semantical$_o$ construct. But the weak definability thesis is, I believe, quite plausible, and thus, given the (extended) functionalist assumption, so is the thesis of the semantic$_o$ independence of belief; and the plausibility of the conjunction of weak definability and the semantic$_o$ independence of belief just is the plausibility of intention-based semantics.

3 *Believing*, I shall continue to assume, is a relation, a relation between a person and a value of the quantifiable variable 'p' in the schema

x believes that p,

a relation to things having truth-values and standing in logical relations to one another (cf. [13]; [32], Ch. 2; and [42]).

There are but two possibilities regarding the values of 'p', the relata of belief. Either (a) they are things which *are* contents: propositions, in one sense or another, extralinguistic, abstract entities which intrinsically have truth-conditions; or else (b) they are things which *have* content: formulas or representations, linguistic or otherwise, which only contingently have, or ascribe, truth-conditions.

If the Fregean hypothesis (a) is correct, then (relative to the functionalist assumption) there is, in a manner of speaking, only one task for the theory of belief, and only one place where a semantical$_o$ concept might enter into that theory, namely, in the (extended) functionalist account of that relation which must obtain between a person x and a content p in order for x to believe that p.

On the other hand, if the anti-Fregean hypothesis (b) is correct, then the theory of belief will have, as it were, two components: (a) *a theory of content* for the relata of belief, and (2), as before, a theory of the relation between x and the sentence or representation p, content having, so to say, been provided for p, by virtue of which x believes that p. Consequently, on the anti-Fregean hypothesis there are, in principle, two places in the explication of belief where the need for recourse to semantical$_o$ properties may be felt. I shall return to the anti-Fregean issue in the next section.

Suppose that, together with our ongoing extended functionalist assumption, the Fregean hypothesis (a) is correct. Then there is some theory T–known or unknown, common sense or scientific–such that, for each proposition p in the domain of T, the theoretical definition of belief with respect to T determines an extended functional property with which the belief property *being a belief that p* is identical.[14] Is it plausible to suppose, ignoring now the argument from weak definability, that there is some semantical$_o$ construct which T will theoretically define in tandem with the belief relation?

At least three conditions must obtain in order for there to be a theory T and a semantical$_o$ construct M such that M and belief are theoretically co-defined with respect to T:

1. *M must occur in T*; i.e., it must be possible to state T in a way which nonsuperfluously uses M. This is clear enough: T can hardly determine the extension of M unless M is a theoretical primitive of T. An example of a theory which might with some plausibility be thought to define belief, but does not appear to satisfy this condition, is our common sense, belief-desire theory of behavior, where by this I primarily mean that system of generalizations involving the constructs belief, desire, and intention which, when conjoined with circumstantial facts, yields explanations of behavior, linguistic and otherwise, and is regarded as common knowledge among those who have acquired those psychological concepts. It is the common sense theory implicitly invoked when one explains Alma's falling to the ground as a flinging of herself to the ground, because she both believed that that was the only way to avoid the collision of a rock with her head, and desired not to suffer the consequences of such a collision. Such a theory, construed broadly, will comprise generalizations which pertain to the ways in which sensory stimulations determine perceptual beliefs, beliefs determine further beliefs, beliefs and desires determine further desires and intentions, and intentions determine behavior. Now it is of course true that semantical$_o$ concepts enter into our explanations of linguistic behavior, but they do so *only as the values of the theory's propositional variables*, and not, therefore, in the theory itself. To be sure, Janet uttered 'Il pleut' because she desired to tell Etienne that it was raining, and believed that those sounds meant that it was raining in Etienne's language. But, needless to say, this no more shows that the semantical$_o$ concepts utilized in this explanation enter into the belief-desire theory invoked than the fact that Janet put a quarter in the vending machine because she desired a cup of chicken soup shows that *chicken soup* is a construct of our common sense psychological theory. The only constructs *of the theory* to enter into such explanations of behavior, linguistic or otherwise, are belief, desire, and intention.[15]

2. *The role of M in T must be rich enough to determine the extension of M.* The mere fact that M co-occurs with belief in a theory which defines belief does not show that M itself is theoretically defined with respect to that theory. But whether or not M is theoretically defined with respect to T we can, by following the procedure outlined in Section 2, form the *presumptive* theoretical definition of M with respect to T, and in a given case it may be obvious that the presumptive definition of M with respect to T is altogether inadequate to determine the extension of M, and that therefore M is not theoretically definable with respect to T. For example, suppose that a theory T' contains the relational predicate 'x asserts that p', but that the only role this predicate has in T' occurs in the generalization 'If x asserts that p, then,

typically, x believes that p'. In this case it is obvious that the role of assertion in T' is not nearly rich enough for it to be defined by its presumptive theoretical definition with respect to T', for satisfaction of the definiens of that presumptive definition will clearly fail to provide a sufficient condition for the satisfaction of its definiendum. Or, to take another example, in discussing a certain sort of reliability theory–that is to say, a theory which would account for the ways in which our beliefs are reliable indices of the truth of the propositions believed–Hartry Field suggests that such a theory might need to "mention acquisition of a public word that refers to something as one of the mechanisms by which people can become causally related to an object in the way that is relevant to having beliefs about the object" ([13], p. 60). Even if Field is right about this (I doubt that he is), it is still *very far* from clear that the presumptive theoretical definition of *reference*$_o$ with respect to such a theory would suffice to determine its extension.

3. *There must not be any theory T^*, not containing M, such that the theoretical definition of belief with respect to T^* would suffice to determine its extension.* Here I advert to the intuitive point which preceded the argument from weak definability (Section 2). M and belief are theoretically co-defined with respect to T only if belief cannot be explicated apart from its interlocking role with M in T. But, for any such T, let T^* be that theory obtained from T by replacing all occurrences of M in T with its intention-theoretical definiens. It is, I should think, encumbent on one who would argue for the co-definability of M and belief with respect to T to show that belief could not be defined with respect to its role in T^*; and, I should further think, the only *prima facie* feasible line for one to take against this maneuver is to argue that semantical* constructs do not have psychological reality, and that, therefore, T^* will not be equivalent to T–which line, I have argued, is not really very feasible.

Now it happens that, roughly speaking, I can think of only three candidate theories for the functional definition of belief relative to the Fregean hypothesis: (i) a common sense belief-desire theory; (ii) a certain sort of reliability theory; and (iii) a certain kind of theory of radical interpretation suggested by various Davidsonian and neo-Davidsonian writings.[16] For whatever it is worth, I hazard the opinion that, quite apart from any problems from which these theories might suffer, a thorough examination of them would reveal that none of them satisfy the foregoing conditions (2) and (3), that (i), as I have already given reason to believe, does not satisfy condition (1), and that (ii) only a little less clearly fails to satisfy it.

4 Hopes for an incursion of the semantic$_o$ into the psychological might initially rise when we turn, as now, to the anti-Fregean hypothesis (Section 3), which eschews contents, and construes believing as a relation to something which has content, its having of content not in turn self-defeatingly to appeal to contents, but to be explicated in an extensionalist, somehow truth-theoretic vein.

At the most general level of abstraction we have an exhaustive division among the anti-Fregeans between

(A) theories which construe belief as a relation to sentences, or utterances, in the belief *ascriber's* language,[17] and

(B) theories which construe belief as a relation to sentences, or internal representations, of the believer's.

Thus, for the (A)-theorist the variable '*p*' in the schema '*x* believes that *p*' ranges over sentences in the language of one ascribing a belief, regardless of whether the one to whom the belief is ascribed speaks that language, or any other (for we do ascribe beliefs to very young children and other languageless animals), while the (B)-theorist will construe '*p*' as ranging over formulas in some "language", inner or outer, of the person to whom a belief is ascribed.

As regards my utterance of

(†) Raffaele believes that love is cruel,

the (A)-theorist will claim that this utterance is true just in case Raffaele stands in the belief relation to my sentence, or foregoing utterance of, 'love is cruel'. The (B)-theorist, if he does not refuse the task of accounting for the logical form of ordinary language belief ascriptions (cf. [42], pp. 208-209), will claim that (†) is true just in case Raffaele stands in the belief relation to some sentence, or mental representation, of his, referred to on the occasion of my utterance of (†) by virtue of its standing in a certain relation to my sentence 'love is cruel'.

Both (A) and (B) theories admit of a further division.

(A)-theorists divide into those who claim that:

(A-1) Belief is a *semantical*$_o$ relation to a sentence, or utterance, of the ascriber's,

and those who claim that:

(A-2) Belief, while a relation to a sentence of the ascriber's, is *not* a semantical$_o$ relation;

and it will be well to pause awhile here before taking up with the (B)-theorists.

It is not difficult to locate the motivation for (A-1).

Frege held that the verb 'believes' in sentences of the form '*x* believes that *p*' was a two-place relational predicate which related a person to the proposition expressed by the sentence in the '*p*' position. Subsequent philosophers have generally appreciated the need for a relational account of belief, but many have been reluctant to join Frege in seeing belief as relating a person to a proposition. But if belief is a relation, and not a relation to a proposition, then to what can it be a relation? The objects to which belief relates us must provide content for our beliefs, they must have truth-value, and they must stand in logical relations to one another. There is, it is easy at this point to feel, but one alternative to propositions: sentences, or at least, as on Davidson's paratactic account, utterances of them [5]. But if a relation to sentences, then no doubt to *our* sentences; after all, we ascribe beliefs to creatures without language, and to people who speak languages of which we are totally ignorant. But if, in uttering (†), I am referring to my sentence 'love is cruel' in such a way as to ascribe to Raffaele a mental state with a certain content—a state that is true if, and only if, love is cruel—then this must, it would seem, be effected by virtue of the content, the meaning$_o$, that 'love is cruel' has for me. In this way, not altogether unintuitive, we arrive at the idea that belief is a semantical$_o$ relation to a sentence (or utterance), that is, a relation to a sentence (or utterance) that obtains by virtue of the meaning$_o$ which that sentence (or utterance) has for the belief ascriber.

Evidently, this position precludes, on pain of vicious circularity, a reduction of meaning$_o$ to beliefs and intentions, while, at the same time, the sane (A-1)-theorist will not propose that meaning$_o$ is explicable independently of belief and intention. Rather, he will no doubt claim, as one of them has, that:

> neither language nor thinking can be fully explained in terms of the other, and neither has conceptual priority. The two are, indeed, linked, in the sense that each requires the other in order to be understood; but the linkage is not so complete that either suffices, even when reasonably reinforced, to explicate the other. ([8], p. 8)

Relative to our (extended) functionalist presupposition, the (A-1)-theorist must maintain that belief and some semantical$_o$ construct are to be functionally reduced only via their theoretical co-definitions with respect to some theory in which they occur as primitive, co-ordinate constructs.

But I can make no sense whatever of the hypothesis that it is the case both that: (a) belief is a semantical$_o$ relation to sentences or utterances in the belief ascriber's language and that (b) the semantic$_o$ is to be theoretically co-defined with belief in the way indicated.

Imagine that we have constructed a psychological theory of Martians in which Martian beliefs relate Martians to our sentences by virtue of their meanings$_o$ in our public language, as the (A-1)-theorist requires. (Perhaps the theory contains generalizations of the form '(σ) (if x believes σ and σ is true$_o$ in English iff ____, then . . .)', or '(σ) (if x believes σ and σ is used by us to say$_o$ that such and such, then . . .)'.) Now consider the claim that this theory is a functional theory, and that the semantical$_o$ predicates pertaining to our language are defined—i.e., have their extensions determined—by their roles in this Martian psychology.

The patent absurdity of this proposal does not reside in the thought that there might be a true psychological theory of Martians which had recourse to the semantical$_o$ properties of our words. For those properties might have their point by way of their entering into contingent descriptions which enable us to refer to internal Martian states involved in the causal nexus explanatory of Martian behavior. The absurdity lies rather in the thought that our semantical$_o$ constructs might be functionally defined by their roles in Martian psychology. For it makes sense to suppose that the semantical$_o$ constructs pertaining to our language are functionally defined by Martian psychology only if it makes sense to suppose that the physical properties which realize those constructs, and determine their application, play an explanatory role as regards Martian behavior; but it is absurd to suppose that the physical facts which determine the semantical$_o$ properties of my words have anything whatever to do with the behavior of Martians.

But the problem is in no way mitigated if, keeping all else the same, we now suppose that the psychological theory applies not to Martians, but to us and all other mature humans. For how can the physical facts which determine the semantical$_o$ properties of my words be in any way explanatory of *Raffaele's* behavior?

Essentially the same objection can be made from a different direction. Suppose that we have a psychological theory T of the behavior of all rational humans which treats belief as a semantical$_o$ relation to our words, and consider the *presumptive* theoretical definition of a given semantical$_o$ property M of our words with respect to T. Intuitively, any physical property which realizes M must be such as to realize in *us* certain psychological states, and to determine for *us* a certain use of the marks or sounds which bear M. Yet it is impossible to see how this could be the case if M were functionally defined by its presumptive definition with respect to T. For the functional property equated with M by that definition will not require a realization of M to play any special role in our production of marks or sounds, but (if this can even be made intelligible) will instead merely require the realization to play a certain role as regards the explanation of anyone's behavior, no matter what language, if any, he or she speaks.

So much for (A-1) theories.

The (A-2)-theorist agrees with the (A-1)-theorist that belief is a relation to a sentence in the belief ascriber's language, but, unlike the latter, denies that the relation is semantical$_o$. I know of only one theory of this type, the terrifically ingenious theory developed with painstaking care and great subtlety by Brian Loar in his book *Mind and Meaning* (see also [31]). To the best of my knowledge, Loar's theory is the only fully developed functionalist theory of belief, and, a fortiori, the only such theory according to which belief is completely semantic$_o$ independent.

All refinements aside, Loar maintains that a belief ascription performs two tasks: (a) it ascribes to a person a state of a type having a certain functional role, and (b) it ascribes to that state a certain truth-condition. And, to repeat, it performs both tasks in a way that is wholly independent of the semantical$_o$ properties of the sentence which indexes the functional role and, in its way, ascribes the truth-condition. The central idea is that, with respect to (a), one exploits the fact that there are certain formal, nonsemantical$_o$ relations holding among sentences, and that, with respect to (b), one exploits the fact that a homophonic Tarski-predicate is definable on English sentences in total independence of any semantical$_o$ properties those sentences might have.[18]

As regards (a), the claim is that 'believes' is partly defined by a certain function f–determined for the most part by the role of 'believes' in our common sense belief-desire theory–from agents, times, and indexing sentences to internal state-types such that, for example,

> f (Raffaele, t, 'love is cruel') = P just in case P has the functional role indexed by 'love is cruel' in our belief-desire theory, and Raffaele is at t in a token of P.

The sentence 'love is cruel' here serves as an *index* of a certain functional role, and this just by virtue of the fact that certain *formal*, syntactical relations which obtain between that sentence and others, relations specifiable without reference to any roles those sentences play in a public language, are, as it were, mirrored by certain causal, counterfactual, and transitional relations among internal, physical states. To say that a given neural state-token is a belief that love is cruel is, in a manner of speaking, to say that it is a token of a type whose causal (etc.) relations to sensory inputs, to other internal state-types, and to behavioral outputs is in such and such respects like the such and such formal relations which 'love is cruel' stands in to other sentences.

The idea as regards task (b), the ascription of truth-conditions, is that, where T is the homophonic Tarski-predicate for English,

a state-token s is *true* iff $(\exists\sigma)$ (s is a belief that σ and σ satisfies T),

where, as before, s's being a belief that σ is a matter of its being of a type which has that functional role indexed by σ in our psychological theory. As Tarski-predicates are definable without reference to any semantical$_o$ properties, this account of the truth-conditions of beliefs is also untainted by the semantic$_o$. Loar further succeeds in motivating this account of the truth-conditions of beliefs in terms of a theory of the reliability of beliefs, but, as his theory of belief is semantic$_o$ independent, I shall press on.

(B)-theories are also of two types, one of which maintains that:

(B-1) Belief is a semantical$_o$ relation between believers and *their* public language sentences.

One immediate problem (there are others) with this is that it withholds beliefs from very young children and all other nonspeaking animals, which seems neither enlightened nor correct. Our present concern, however, is whether we can make sense of the thought that there is some functionalist account of belief which conforms to (B-1), and whether, if there is such a theory, it shows belief to be semantic$_o$ dependent.

We have to imagine a theory T laid out so as clearly to exhibit belief as a semantical$_o$ relation to believers' sentences. Perhaps T would contain generalizations of the form

If x believes σ and σ means$_o$ ___ for x [or, is true$_o$ in x's language iff ___], then . . ..

I am fully convinced that a patient discussion of T would reveal that it could not possibly define its semantical$_o$ terms, in that the presumptive definitions of those terms with respect to T would be seen not to provide either necessary or sufficient conditions for belonging to the extensions of the terms they purport to define. However, I cannot take the suggestion that such a theory, functionally construed, would define its semantical$_o$ terms sufficiently seriously to be patient with it, and I very much doubt that any *functionalist* has ever been a (B-1)-theorist. I shall therefore content myself with the following already familiar reason for thinking that no semantical$_o$ construct could be co-defined with belief with respect to T. If T^* is formed by replacing T's semantical$_o$ terms with their intention-theoretical definientia, and if, as I have argued, semantical* properties are realized, then, it would seem, T^* is true if T is. In that case, what sense can be attached to the claim that belief can be defined, not with respect to T^*, but only with respect to T?

Finally, we arrive at the last of the anti-Fregean hypotheses, and second type of (B)-theory, the topical idea that:

(B-2) Belief is a relation between a person and a mental representation, a "formula" in his "language of thought."

Beliefs, having content, are representational states; a belief that Mitterand defeated Giscard is a state which, in its way, represents its being the case that Mitterand defeated Giscard. If, as it is reasonable to suppose, beliefs are neural states, then those neural states, having content, are mental representations. Mental representations, then, are easily purchased.

Nor is it unlikely that our mental representations belong to an internal *system* of mental representation, a language of thought; for this further claim is simply the claim that our mental representations, the neural states which realize our beliefs, can be viewed, like sentences of a natural language, as organized structures, the representational features of the whole derived, in ways describable by finitely specifiable recursive conditions, from the representational features of its constituent parts and structure.

And if we are to talk of neural states as being formulas in a language of thought, then we might as well talk of the meanings and references of these formulas, for, to forestall the naivest confusion, such talk is just talk of the representational features, or content, of mental states, and is in no way incompatible with intention-based semantics, nor with any other semantics for natural language that is consistent with our having beliefs. Letting the subscript 'i' index semantical properties in the system of internal representation, we may say that, given that we think in a language of thought, the intention-theorist's claim is that the semantic$_o$ reduces to the semantic$_i$, or, to revert to an older idiom, that the theory of content for public language reduces completely to the theory of content for thought.[19]

Now we have still not arrived at (B-2); for the Fregean, who maintains that belief is a relation to a proposition, is free to subscribe to the hypothesis that we think in a language of thought, which, *for him*, would mean that (where 'R' and 'R'' range over relations, 'σ' over formulas in x's language of thought, and 'p' over propositions):

$$(\exists R)(\exists R')(x \text{ believes that } p \text{ iff } (\exists \sigma)(xR\sigma \text{ and } \sigma R'p)).$$

Here R'–whatever it turns out to be–would (assuming a unique R') be for the Fregean the meaning$_i$ relation which holds between internal sentences and the propositions which are their meanings$_i$ (and, if he is also an intention-theorist, he would claim that the specification of R' need make no appeal whatever to anything semantic$_o$).

But, coming now to the specific motivation for (B-2), it is very tempting to take the internal formulas *themselves* as the relata of belief, and this for at least two reasons: (1) it would put the relata of belief in the head, where they can play causal roles in the production of one's behavior, and (2) it would free one to pursue an extensionalist, truth-theoretic semantics for the language of thought, and thus to account for the contents of thoughts without appeal to contents (for some other possible motivations, see [15]).

Much of the motivation for the hypothesis that we think in a language of thought derives from the fact that the best existing theories of cognitive psychology construe mental states and processes as computational states and processes defined on formulas in an agent's inner code [14]. If such a theory employs anything like a reconstruction of our common sense notion of belief, it will construe it as a relation between an agent and a formula of his internal system of representation; will, that is, contain quantifications of formulas containing open sentences of the form

x believes* σ

wherein 'σ' is explicitly treated as a variable ranging over formulas of the inner code, and 'believes*' represents the reconstructed belief relation. The claim that believing is a relation between a person and an internal sentence may then reasonably be taken to mean that the functionalist reduction of belief is determined by the theoretical definition of it, or some reconstruction of it, with respect to such a theory (cf. [42], p. 209).

To believe that Bertrand Russell was wise is, according to the hypothesis (B-2), to believe some sentence in one's language of thought which means_i that Bertrand Russell was wise. If this is correct, then the theory of belief will have two components: (1) a theory of the belief relation, as a relation to internal sentences; and (2) a theory of meaning_i for the language of thought. How plausible is it to suppose that something semantic_o would enter into either subtheory, in such a way as to entail the semantic_o dependence of belief, should (B-2) be correct?

Nothing semantic_o would enter into (1), the theory of the belief relation. For it is plausible to suppose that this theory would require appeal to the semantic_o only if it requires appeal to the semantic_i features of inner expressions, and it is not plausible to suppose that. For, to echo a now familiar refrain,[20] nothing semantic at all will enter into that information-processing model of cognition which will, if the (B-2)-theorist is right, define belief as a relation to internal sentences, and this because *the representational character of internal representations*—the semantical_i features of the inner formulas—*plays no role in such psychological theories*, that, as far as such information-processing theories are concerned, the internal system might as well be an uninterpreted calculus. The point is related to one made earlier, that the *narrow* functional role of a state does not determine its content, i.e., its truth-condition. I shall briefly explain.

The *conceptual role* of an internal sentence is, loosely speaking, the union of the *narrow* functional roles of all the propositional attitudes (but notably believing, desiring, and intending) which have that formula as their object; so, to know the conceptual role of an internal sentence σ is to know the narrow functional roles of the belief that σ, the desire that σ, etc. Thus, a theory of conceptual role would tell us how sensory stimulations influence what sentences we believe, how our beliefs influence one another, and how beliefs and desires lead to further desires, to intentions, and, eventually, to bodily movements.[21]

Now the representational content of an internal representation may for the most part be identified with its truth-condition: to know that σ represents snow's being white is just to know that that is σ's truth-condition; and, quite summarily stated, the reason that it is irrelevant to psychological theories of cognitive processes that internal representations are representations is that

(a) the *conceptual role* of an internal formula is its only property which is ultimately relevant to information-processing models of cognitive processes;

but

(b) the conceptual role of a formula does not determine, and can be specified without making any reference to, the formula's semantical$_i$ properties, that, as regards the specification of conceptual role, one need only mention the *syntactical* features of the formula.

The idea behind (b) is that conceptual role is an intra-cranial property; bounded by sensory stimulation on the input side and by mere bodily movement on the output side, it is essentially a matter of what is inside the head; whereas, on the other hand, what *truth-condition* a sentence has is essentially a matter of how it is related to things outside of the head (cf. [12], pp. 397-398).

This returns us to (2), the theory of meaning$_i$ for the inner code, with an old question and, inextricably related to it, a new one. The old question is:

(i) What form will the theory of meaning$_i$ for the language of thought take? Which, in its way, is to ask: What determines the representational character of internal representations?

The new question is:

(ii) If it is irrelevant to conceptual role psychology that internal representations *are representations*, then why do we *need* internal *representations*? Why should it not be enough to construe believing and desiring as relations to syntactically specified, but meaningless, formulas? Why, in a word, is it important that beliefs and desires be construed as having *content*?

The *content* of psychological states is superfluous to a psychological theory whose only concern is the explanation of bodily movements in terms of the functional roles of internal states; for narrow functional role is all the functional role one needs for that, and the narrow functional role of a belief does not determine its content, and is specifiable without reference to that content. But an interest in a person's bodily movements is not the only basis for an interest in his beliefs. We may want to know what a person believes because we regard his beliefs on certain matters as being reliable indices of how the world is. If, for example, a normal person believes that she has been to Greece, then the fact that she has that belief is extremely reliable evidence that she has in fact been to Greece. This interest in the reliability of beliefs is not to be minimized. If intention-based semantics is correct, linguistic behavior is founded on our mutually exploiting the reliability of one another's beliefs.[22]

It is arguable, and has in fact been argued, that the ascription of truth-conditions to those states which are beliefs is necessary for the systematic exploitation of the evidential reliability of those states as indices of how the world is, and, further, that that is the only place where the need for truth-conditions arises (cf. [12], [13], [32], and [42]). This provides the only answer to (ii) of which I am cognizant: the representational character of an internal sentence, and so of an internal state, is mostly, if not wholly, a matter of its truth-condition, and truth-conditions must be ascribed to belief-states, and so to internal sentences, if we are systematically to exploit the ways in which they can be informationally reliable about the world.

This, in turn, as we shall presently see, gives us leverage on question (i).

The theory of meaning$_i$ for the language of thought will have two components: a conceptual role component, provided by subtheory (1), and a theory of truth$_i$ for the internal system, the proper concern of subtheory (2) (cf. [12]). The theory of truth$_i$ for the language of thought will tell us what it is that determines the truth$_i$-conditions of any formula in anyone's system of mental representation, and there are two possibilities concerning the form that such a theory might take:

(a) One possibility is that there is some theory T such that semantical$_i$ constructs (*refers$_i$ to*, *is true$_i$ of*, *is true$_i$*, etc.) occur as primitives of, and are theoretically defined with respect to, T, in such a way as to determine the truth$_i$-conditions of sentences in those systems of mental representation within the domain of T, and thereby the truth-conditions of those beliefs whose objects are those sentences.[23]

(b) Another possibility is that there will be an *extra*theoretical explicit definition of the form:

> For any system of mental representation M, and any sentence σ of M, σ is true$_i$ in M iff . . ..

Here is where the reliability answer to (ii) comes into play. If (a) is how the semantical$_i$ properties of inner expressions are determined, then the theory which determines them will be a *reliability* theory, a theory which tells us, relative to certain parameters, the conditions under which, and the degree to which, believing a sentence is evidence that that sentence is true$_i$, 'true$_i$' defined by its role in this reliability theory. If, however, the truth$_i$-conditions of internal sentences, and thus the truth-conditions of beliefs, are determined by an extratheoretical definition of kind (b), then that definition would evidently be in some way in terms of reliability considerations. For example, the following might provide the crude outline of one sort of first shot (where 'M' ranges over systems of mental representation and 'σ' over internal sentences):

> $(M)(\sigma)$ (σ is true$_i$ in M iff there is a Tarski-predicate T definable on M such that: (1) σ satisfies T and (2) thinkers in M are maximally reliable under T)

where the claim that x is more reliable under T than under an alternative Tarski-predicate T' might, very roughly, be taken to mean that, in general, the probability of σ satisfying T given that x believes σ is greater than the probability of σ satisfying T' given that x believes σ.

The bearing of all this on the over-arching question of the semantic$_o$ dependence or independence of belief relative to (B-2) is as follows.

We have seen that nothing semantical$_o$ will enter into the (B-2)-theorist's account of the belief relation, and now we are in a position to remark that nothing of the semantic$_o$ will enter into the theory of meaning$_i$ should truth$_i$ for the language of thought be explicable in way (b), and this for two reasons. First, it is, I submit, impossible to see how such an explicit definition *could* make appeal to the semantic$_o$. Secondly, such a definition, on pain of vicious circularity, could be ineliminably in terms of some semantical$_o$ construct only if that construct were in turn belief, and thus semantic$_i$, independent, and it is absurd to suppose that any semantical$_o$ construct is belief independent.

Thus the issue, as regards (B-2), boils down to this. Assuming that (B-2) is correct, and that semantical$_i$ constructs are theoretically definable with respect to some reliability theory, will there be one or more semantical$_o$ constructs co-defined with those semantical$_i$ constructs with respect to that reliability theory?

It would no doubt be egregiously foolhardy, not to say rash, to pronounce on this question in the absence of a fully articulated reliability theory for systems of mental representation, but, should I be forced at gunpoint to do so, I should bet that: (1) no semantical$_o$ construct will so much as occur in the reliability theory; (2) if some semantical$_o$ construct were to play some role in the reliability theory, that role would not nearly be rich enough to determine that construct's extension, which further suggests that (3) the reliability theory that was actually defining the semantical$_i$ constructs was the logically equivalent one obtained from the first by replacing its semantical$_o$ terms with their intention-theoretical definientia. But on these matters I have labored long enough.

5 Intention-based semantics is the theory that semantical$_o$ properties are identical to psychological properties bearing semantical* labels, and, as such, it is tantamount to the thesis that both (a) the semantic$_o$ is weakly defined by the semantic*, and (b) belief is semantic$_o$ independent. Weak definability seemed nearly enough unavoidable, granted the psychological reality of speaker's meaning*, not itself implausible. But anxiety arose on the score of the semantic$_o$ independence of belief: Could belief really be explicated without at any point appealing to the meanings and references of words in our public language? A materialist prejudice led me to assume that belief properties were extended functionalist properties, and it was obvious that the semantic$_o$ could not be explicated without reference to belief, the conjunction entailing that belief was semantic$_o$ dependent only if it was co-definable with some semantical$_o$ construct with respect to some psycho-semantical$_o$ theory. I then tried to give reasons for being more than a little skeptical of the co-definability thesis. First, there was the argument to show that weak definability was incompatible with the theoretical co-definability of belief and meaning$_O$, and then there was a lengthy, but hardly exhaustive, discussion of the theories which might provide the wherewithal for a functionalist reduction of belief, the upshot of which was, I hope, that: (a) the more a theory looked as if it might require co-definability (e.g., (A-1) and (B-2)), the less plausible it seemed that it could define the semantic$_O$ in tandem with belief; (b) none of the more intuitively attractive theories entail co-definability (at least as described); and (c) among the theories consistent with co-definability, there were, in each case, two or more reasons for thinking that they would not co-define belief and meaning$_O$.

I have used weak definability to argue against co-definability. But the lack of any viable candidate for the co-definition of belief and meaning$_O$ can in turn be used to support the intention-theorist's version of weak definability, and thus will serve to emphasize a point I made early on, that the motive for the reduction of the semantic$_O$ to the psychological was that it permitted the only viable reduction of the semantic$_O$ and the psychological to the physical. For suppose that belief is functionally reduced via its role in some theory, but not in tandem with meaning$_O$. This realistically leaves but three possibilities: (i) meaning$_O$ is irreducibly semantic, not functionally nor otherwise reducible to the physical; or (ii) meaning$_O$ does admit of a functionalist reduction by some semantical$_O$ theory, but not in such a way as to yield the co-definability of meaning$_O$ and belief; or (iii) meaning$_O$ reduces to the psychological, to belief, desire, and intention. But (i) violates materialist sensibilities every bit as much as its psychological counterpart: irreducibly semantical facts can hardly be thought more palatable than irreducibly psychological facts. And (ii) contradicts the evident belief dependence of meaning$_O$. Which leaves (iii). But if (iii), then meaning$_O$ is definable in terms of the psychological, and the question arises as to what those definitions might be. I submit that to this question intention-based semantics provides a reasonable answer.

I want to close this paper by commenting all too briefly on two features of it, which are: (1) that my discussion has proceeded on the assumption that there are determinate belief properties; that, for example, there is some state of mine which is, quite literally and determinately, my belief that snow is white; (2) that I have not argued for any particular functionalist theory of belief.

If I had to rank articulated functionalist theories of belief, I should perhaps put Loar's first. But Loar's theory has consequences which I would prefer not to have to live with. The problem is that, on Loar's account, there is some *narrow* functional property the possession of which is sufficient for a state's being a belief that snow is white, and likewise for belief properties generally. Loar is not unaware of the problematical aspects of this feature of his theory, and has deft things to say about it. But I have yet to be entirely convinced.

Let *n* be that neural state-token which is my belief that snow is white, if anything is that belief, and let Φ be that narrow functional property of *n*'s which entails every narrow functional property which n has. I am inclined to suppose that n's being Φ is sufficient for its being a belief, but not for its being a belief that snow is white, and that, if n is a belief that snow is white, then that is entailed by n's being Ψ, for some extended, *non*narrow functional property Ψ. But, to the best of my knowledge, it has so far proved impossible to give any kind of plausible account of what Ψ might be, or of the theory which determines it. It is enough to make one flirt, as I did elsewhere (cf. [42], pp. 220-222), with the Quinean thought that, there being no such Ψ, the contents of our thoughts are radically indeterminate, that there simply is no fact of the matter as to what I believe. It is of course much too soon to despair, and the despair of the indeterminacist can be every bit as facile as its opposite. Among other things, if we do not have beliefs with determinate content, then we need a much better account than any that has yet been offered of the illusion that our thoughts have determinate content, of the source of the indeterminacy, and, lastly, of its cost.

In a recent paper Paul M. Churchland argues that there are no beliefs, desires, or intentions, and then considers an attempted reductio of his eliminativist thesis which proceeds

> by pointing out that the statement of eliminative materialism is just a meaningless string of marks or noises, unless that string is the expression of a certain *belief*, and a certain *intention* to communicate . . . and so forth. ([3], p. 89)

Churchland's response to this reductio is to reject the theory of meaning which it presupposes. I hope that this paper shows that Churchland may not have fully appreciated the cost of his eliminativism.

NOTES

1. See also Bennett [1]; Grice [18] and [19]; Lewis [22], [26], [27]; Loar [29] and [32]; Peacocke [34]; Schiffer [39], [42], and [43]; and Strawson [46] and [47].
2. See Schiffer [39], Chapters 2 and 3. A general argument for the definition of '*S* means something in uttering *x*' is given on pp. 80-87. The verb 'to utter' and its cognates are used in an artificially extended sense which applies to nonlinguistic acts and products, so that 'in uttering' is here equivalent to 'in doing or producing'. An *activated* belief is, roughly, one that one consciously has in mind.
3. Bennett [1], pp. 126-127; Loar [32], Ch. 10. See also Kemmerling [21].
4. For the account of illocutionary* acts, see Strawson [46], and Schiffer [39], Ch. 4. For intention-theoretical accounts of reference, see Loar [30], and Schiffer [40], [41], and especially [42].
5. This differs from the treatment in Schiffer [39], Ch. 5, in not explicitly requiring that the mutual knowledge* be based on precedents pertaining to the constituents and structure of σ.
6. See Loar [32], Ch. 10, for a similar defense of the psychological reality of speaker's meaning*.

7. The claim that σ's meaning p in a public language is in this way equivalent to σ's meaning* p in no way implies that σ's only proper or literal use is to perform acts of speaker's meaning*. It is true, as Chomsky ([2], Ch. 2) and others have emphasized (Davidson has even dignified the platitude with the title 'the autonomy of meaning' [8]), that a sentence can properly function in numerous noncommunicative ways: in idle conversation, in thought, in story telling, etc. The intention-theorist's claim is that the communicative use of a sentence is unique among its various uses in providing a necessary and sufficient condition for its having meaning in a public language. I will not pursue the plausible, but separate, idea that the various noncommunicative uses of a sentence are made possible or otherwise explained by its communicative function.

8. For example, it may be that a state-type Φ has the *extended* functional role F in x's psychology just in case Φ has a certain *narrow* functional role F' in x's psychology and x's acquisition of a state having F' is owed to x's standing in such and such causal relations to a certain substance, or to creatures of a certain species. Here I am indebted to White [49].

9. (a) Thus, the functionalist needs two kinds of properties: functional *roles* (narrow or extended), which are relational properties of state-*types* (which are first-order, no doubt physical, properties of a certain kind), and functional *properties* (narrow or extended), which are properties of state-*tokens*–viz., a property of being a token of a state-type which has a certain functional role. Belief properties, being properties of state-tokens, are thus to be identified with functional properties. (b) Relational state-types will have functional roles; so the notion of a functional property is really just the limiting case, when $n = 1$, of an n-ary functional relation. We may thus say that R is an *n-ary functional relation* just in case, for some functional role F,

   $Rx_1, \ldots, x_n$ iff for some first-order (no doubt physical) relation R', $R'x_1, \ldots, x_n$ and R' has F.

   In other words, a functional relation is a relation among things which obtains when they are related by some relation which has a certain functional role. (c) My claim that the identification of belief properties with (extended) functional properties is the only *viable* way of being a materialist should not be taken to suggest that it is the only way of being a functionalist. There is also Lewis's version of the identity theory, discussed below.

10. '$=_{df}$' should be taken, with a grain of salt, as meaning that the definiens determines the extension of the definiendum, and not that it somehow gives its meaning. For at this stage we should not want to preclude the possibility that the functional definition of belief is to be provided by a scientific theory, perhaps even one that is yet unknown, rather than by a common-sense theory. I am more than a little puzzled by the idea that the extension which a term has *for me* could be determined by its role in a theory of which I am altogether ignorant, but there are evidently functionalists who believe this, and I wish now to be as untendentious as possible.

11. Except when the distinction matters, I shall follow the practice of implicitly equating a theory with some privileged, canonical statement of it. It should also be noted that, consistent with the theoretical definition of belief by a theory T, there is more than one way in which T might effect the identification of belief properties with extended functional properties. See below, note 14, and Loar [32], Ch. 4.

[12] More exactly,

$$B =_{df} (\iota\Phi)(\exists\Psi)(\exists\Omega)(T(\Phi,\Psi,\Omega) \wedge (\Phi')(\Psi')(\Omega')(T(\Phi',\Psi',\Omega') \\ \rightarrow (\Phi' = \Phi \wedge \Psi' = \Psi \wedge \Omega' = \Omega))),$$

and likewise, *mutatis mutandis*, for the definitions of D and M.

[13] The problem is to see how this could be the truth-condition for $T^*(B,D)$, given that B and D are defined by (i) and (ii). For let B^* and D^* be the *functional* properties which B and D stand for, if defined by (i) and (ii). Then, on the face of it, $T^*(B,D)$ iff $\langle B^*,D^*\rangle$ satisfies $\ulcorner T^*(\Phi,\Psi)\urcorner$, which is not equivalent to the desired truth-condition, stated in terms of the satisfaction of $\ulcorner T^*(\Phi,\Psi)\urcorner$ by first-order physical properties which *realize* B^* and D^*. In effect, we want it to be the case that: (a) B and D are defined by (i) and (ii) (and thus stand for B^* and D^*), *while* (b) the variables 'Φ' and 'Ψ' in the Ramsey-sentence '$(\exists\Phi)(\exists\Psi)T^*(\Phi,\Psi)$' range not over functional properties, but over the physical properties which realize them. See Loar [32], Ch. 3, for the way to have one's cake and eat it, too.

[14] Consistent with this characterization, there are, in principle, three ways in which the extended functional property for a given belief–say, the belief that snow is white–might be determined. (1) T might define 'x believes that p' as a functional relation, realizable by first-order, physical relations between persons and propositions just in case those relations have a certain functional role. (See Lewis [25]; Field [13]; Loar [32], Ch. 4. It is, to put it mildly, difficult to see what these realizing relations to propositions might be.) (2) T might define belief as a certain function $f(x,p,t)$ from persons, propositions and times to physical state-types, such that $f(x$, the proposition that snow is white, $t) = P$ just in case P has the functional role T correlates with the proposition that snow is white, and x is at t in a token of P. (See Loar [32], Ch. 4, and the discussion in Section 4 of this paper of this idea divorced from propositions.) (3) T might, so to say, functionally define only the monadic predicate 'x is a belief', and then determine an extended functional property for each belief property via a nontheoretical, explicit definition of the form

x believes that p iff $(\exists s)$ (s is a belief; x is in s; and sRp),

for a given specified R. An approach like this seems to be suggested in Dennett [10], and in Stalnaker [44] it is suggested that a promising approach would be via a refinement of the idea that x believes that p just in case x is in a belief state which, under optimal conditions, x is in only if, and because, p. (This approach will need considerable refinement before it provides a sufficient condition: under optimal conditions one believes that one sees something red just in case the receptor cells in one's eyes are being stimulated in such and such ways; but one's belief that one sees something red is not also a belief that the receptor cells in one's eyes are being stimulated in a certain way.) A variant of this approach, relative to the anti-Fregean hypothesis, is discussed later in this paper.

The issues raised with respect to alternatives (1)-(3) pertain less to the kind of theory which functionally defines belief, than to the way in which a given theory–say, our common sense, belief-desire theory of rational action–is to be regarded as providing such a definition.

[15] A similar argument would show that semantical$_o$ concepts do not enter into the theory's input conditions.

[16] See especially Davidson [6], [7], [8], and [9]; McDowell [33]; Platts [35], Ch. II; and Wiggins [50].

[17] I here ignore the complication that certain *de re* beliefs would be treated as relations to sentential complexes, ordered *n*-tuples of expressions and objects.

[18] A *Tarski-predicate on a language L* is any predicate T definable on L in the style of Tarski [48] so that for each sentence of L there is some condition such that the definition entails that the sentence satisfies T if and only if that condition obtains. For any L, there are indefinitely many Tarski-predicates definable on L. Ignoring indexicality, we may say that T is homophonic for L if, for every sentence σ of L, the definition of T entails $\ulcorner T\bar{\sigma}$ iff $\sigma\urcorner$, where $\bar{\sigma}$ is a structural description of σ. The crucial feature of Tarski-predicates as regards the purpose at hand is that the application of a Tarski-predicate to a sentence entails nothing whatever about any role which that sentence might play in a public language.

[19] Nor is there any circularity problem for the intention-theorist if one thinks in one's public language; if, that is, speakers of English think in English, speakers of French think in French, and so on. (I believe, incidentally, that, while a quite acceptable sense can be given to this claim, it does have to be *given*: one cannot simply say that a neural sentence and a spoken sentence can be tokens of the same sentence-type in the *same way* that spoken and written sentences can be type identical; for spoken and written sentences of a natural language are type identical by virtue of certain *conventions* which prevail among those who both speak and write the language; but there can be no conventions governing our "use" of our neural sentences (cf. Harman [20]).) The intention-theorist's claim, relative to the hypothesis that one thinks in one's natural language, is that that language will require two theories of meaning: a theory of meaning$_o$ for the language qua *public* language, and a theory of meaning$_i$ for it qua *system of mental representation*. That one's language will require two theories of meaning, if one both speaks it and thinks in it, is plausible quite apart from intention-based semantics. For meaning$_o$ in a public language is a matter of convention, and must somehow be explicated, if only partly, in terms of the beliefs and intentions of those who speak the language; whereas none of this even so much as makes sense as regards the internal system of representation.

Still, there is room here for what might seem to be a subtler criticism. If one thinks in one's spoken language, then one's inner language will be acquired as one acquires one's spoken language, and thus the ability to have the thought expressed by a sentence will be dependent on one's having acquired that sentence as a sentence of one's public language. But how, it might be wondered, can this sort of dependence of meaning$_o$ on thought be consistent with intention-based semantics, which seeks to explain the meaning$_o$ of a sentence, its role in a public language, in terms of the thought expressed by it? Thus Field has suggested that

> *part of* what makes a symbol in my system of internal representation a symbol that stands for Caesar is that this symbol acquired its role in my system of representation as a result of my acquisition of a name that stands for Caesar in the public language. If something of this sort is true, it would appear to defeat [intention-based semantics]. ([13], p. 53)

In fact, there is no incompatibility and nothing is defeated. Intention-based semantics does not require that one have propositional attitudes *prior* to one's having acquired a public language; it is perfectly compatible with the hypothesis that one's ability to have beliefs and desires–i.e., one's acquisition of a language of thought–proceeds *pari passu* with one's acquisition of a public language. As regards meaning$_o$ in the public language, the intention-theorist claims that σ means$_o$ p for x only when σ means$_i$ p for x (and x has certain beliefs and intentions involving σ), and as regards σ's meaning$_i$ in the inner language his claim is that this can be explicated without reference to σ's meaning$_o$ in a public language. None of this requires that σ mean$_i$ p *before* σ means$_o$ p. Perhaps it will help to put the point thus. We need to distinguish two claims: (1) the meaning$_i$ of σ *consists in* its having a certain meaning$_o$, i.e., its having a use governed by convention; (2) the meaning$_i$ of σ is (causally or otherwise) *dependent* on its having a certain meaning$_o$, or of the existence of certain public language conventions. (Devitt in [11], Ch. 3, conflates (1) and (2).) The intention-theorist is committed to denying (1), but he need not deny (2); he must claim only that the meaning$_i$ of σ consists in facts—no doubt mostly causal facts—which are speciable without reference to anything semantical$_o$. Let C be that semantic$_o$-free condition satisfaction of which by σ is necessary and sufficient for its having whatever meaning$_i$ it happens to have. It is consistent with the semantic$_o$-free nature of C that part of the explanation of how σ came to satisfy C mentions semantical$_o$ properties of σ, or the existence of certain public language conventions. It is absurd to suppose that meaning$_i$, like meaning$_o$, might *be* a matter of convention; it is quite another thing to suppose that the acquisition of an internal system of representation is dependent on the acquisition of a public language.

[20] See Field [12] and [13]; Fodor [16]; Loar [32]; Putnam [36] and [37]; Schiffer [42]; and Stich [45].

[21] See Field [12] and Putnam [36] (where he speaks of a 'theory of understanding sentences' rather than a 'theory of conceptual role'). A broader construal of conceptual role is adumbrated in Harman [20]; and in [42], pp. 211-212, I give reasons for preferring the narrow construal.

[22] A further application of reliability considerations, one a propos of the psychological theory of behavior, is argued for in Schiffer [42], pp. 217-219.

[23] Field suggests this functionalist account of the semantic$_i$ in [13]. The relation between the truth-condition of a belief and the truth$_i$-condition of the sentence believed is: x is a true belief iff $(\exists\sigma)x$ is a belief that σ and σ is true$_i$.

REFERENCES

[1] Bennett, J., *Linguistic Behavior*, Cambridge University Press, Cambridge, 1976.

[2] Chomsky, N., *Reflections on Language*, Pantheon Books, New York, 1975.

[3] Churchland, P. M., "Eliminative materialism and propositional attitudes," *Journal of Philosophy*, vol. 78 (1981), pp. 67-89.

[4] Davidson, D., "Truth and meaning," *Synthese*, vol. 17 (1967), pp. 304-323.

[5] Davidson, D., "On saying that," in *Words and Objections*, eds., D. Davidson and J. Hintikka, D. Reidel, Dordrecht, 1969.

[6] Davidson, D., "Radical interpretation," *Dialectica*, vol. 27 (1973), pp. 313-328.

[7] Davidson, D., "Belief and the basis of meaning," *Synthese*, vol. 27 (1974), pp. 309-323.

[8] Davidson, D., "Thought and talk," in *Mind and Language: Wolfson College Lectures 1974*, ed., S. Guttenplan, Oxford University Press, Oxford, 1975.

[9] Davidson, D., "Reply to Foster," in *Truth and Meaning: Essays in Semantics*, eds., G. Evans and J. McDowell, Oxford University Press, Oxford, 1976.

[10] Dennett, D., "Beyond belief," forthcoming.

[11] Devitt, M., *Designation*, Columbia University Press, New York, 1981.

[12] Field, H., "Logic, meaning, and conceptual role," *Journal of Philosophy*, vol. 74 (1977), pp. 379-409.

[13] Field, H., "Mental representation," *Erkenntnis*, vol. 13 (1978), pp. 9-61.

[14] Fodor, J. A., *The Language of Thought*, Thomas Y. Crowell, New York, 1975.

[15] Fodor, J. A., "Propositional attitudes," *The Monist*, vol. 61 (1978), pp. 501-524.

[16] Fodor, J. A., "Methodological solipsism considered as a research strategy in cognitive psychology" (with commentary by B. Loar), *The Behavioral and Brain Sciences*, vol. 3 (1980), pp. 63-110.

[17] Grice, H. P., "Meaning," *Philosophical Review*, vol. 66 (1957), pp. 377-388.

[18] Grice, H. P., "Utterer's meaning, sentence-meaning, and word-meaning," *Foundations of Language*, vol. 4 (1968), pp. 225-242.

[19] Grice, H. P., "Utterer's meaning and intentions," *Philosophical Review*, vol. 78 (1969), pp. 147-177.

[20] Harman, G., *Thought*, Princeton University Press, Princeton, New Jersey, 1973.

[21] Kemmerling, A., "How many things must a speaker intend (before he is said to have meant)?" *Erkenntnis*, vol. 15 (1980), pp. 333-341.

[22] Lewis, D., *Convention*, Harvard University Press, Cambridge, Massachusetts, 1969.

[23] Lewis, D., "How to define theoretical terms," *Journal of Philosophy*, vol. 67 (1970), pp. 427-446.

[24] Lewis, D., "An argument for the identity theory," in *Materialism and the Mind-Body Problem*, ed., D. Rosenthal, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[25] Lewis, D., "Psychophysical and theoretical identifications," *Australasian Journal of Philosophy*, vol. 50 (1972), pp. 249-258.

[26] Lewis, D., "Radical interpretation," *Synthese*, vol. 27 (1974), pp. 331-344.

[27] Lewis, D., "Languages and language," in *Language, Mind and Knowledge* (Minnesota Studies in the Philosophy of Science, Vol. VII), ed., K. Gunderson, University of Minnesota Press, Minneapolis, 1975.

[28] Lewis, D., "Mad pain and Martian pain," in *Readings in the Philosophy of Psychology, Vol. I*, ed., N. Block, Harvard University Press, Cambridge, Massachusetts, 1980.

[29] Loar, B., "Two theories of meaning," in *Truth and Meaning: Essays in Semantics*, eds., G. Evans and J. McDowell, Oxford University Press, Oxford, 1976.

[30] Loar, B., "The semantics of singular terms," *Philosophical Studies*, vol. 30 (1976), pp. 353-377.

[31] Loar, B., "Ramsey's theory of belief and truth," in *Prospects for Pragmatism*, ed., D. H. Mellor, Cambridge University Press, Cambridge, 1980.

[32] Loar, B., *Mind and Meaning*, Cambridge University Press, Cambridge, 1981.

[33] McDowell, J., "Truth conditions, bivalence, and verificationism," in *Truth and Meaning: Essays in Semantics*, eds., G. Evans and J. McDowell, Oxford University Press, Oxford, 1976.

[34] Peacocke, C., "Truth definitions and actual languages," in *Truth and Meaning: Essays in Semantics*, eds., G. Evans and J. McDowell, Oxford University Press, Oxford, 1976.

[35] Platts, M., *The Ways of Meaning*, Routledge & Kegan Paul, London, 1979.

[36] Putnam, H., "Reference and understanding," in *Meaning and the Moral Sciences*, Routledge & Kegan Paul, London, 1978.

[37] Putnam, H., "Computational psychology and interpretation theory," forthcoming.

[38] Ramsey, F. P., "Theories," in *Foundations*, ed., D. H. Mellor, Routledge & Kegan Paul, London, 1978.

[39] Schiffer, S., *Meaning*, Oxford University Press, Oxford, 1972.

[40] Schiffer, S., "Naming and knowing," *Midwest Studies in Philosophy: Studies in the Philosophy of Language 2* (1977), pp. 28-41; reprinted in *Contemporary Perspectives in the Philosophy of Language*, eds., P. A. French, T. E. Uehling, Jr., and H. K. Wettstein, University of Minnesota Press, Minneapolis, 1979.

[41] Schiffer, S., "The basis of reference," *Erkenntnis*, vol. 13 (1978), pp. 171-206.

[42] Schiffer, S., "Truth and the theory of content," in *Meaning and Understanding*, eds., H. Parrot and J. Bouveresse, Walter de Gruyter, Berlin-New York, 1980.

[43] Schiffer, S., "Indexicals and the theory of reference," *Synthese*, vol. 49 (1981), pp. 43-100.

[44] Stalnaker, R., "Thought," unpublished manuscript.

[45] Stich, S., "Autonomous psychology and the belief-desire thesis," *The Monist*, vol. 61 (1978), pp. 573-591.

[46] Strawson, P. F., "Intention and convention in speech acts," *Philosophical Review*, vol. 73 (1964), pp. 439-460.

[47] Strawson, P. F., "Meaning and truth," in *Logico-Linguistic Papers*, Methuen, London, 1971.

[48] Tarski, A., "The concept of truth in formalized languages," in *Logic, Semantics, Metamathematics* (transl. J. H. Woodger), Oxford University Press, Oxford, 1956.

[49] White, S. L., "Partial character and the language of thought," unpublished manuscript.

[50] Wiggins, D., "What would be a substantial theory of truth?" in *Philosophical Subjects*, ed., Z. van Straaten, Oxford University Press, Oxford, 1980.

Department of Philosophy
University of Southern California
Los Angeles, California 90007