1982

# Aspects Of Intentional Explanation

Neil A. Farnsworth

# ASPECTS OF INTENTIONAL EXPLANATION

by

Neil A. Farnsworth
Department of Philosophy

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
July, 1982

# ABSTRACT

A complete cognitive science will include generalizations explanatory of human behavior which refer to certain internal states of human agents. We investigate various issues in the foundations of cognitive science arising from this observation. In particular, it is argued that the taxonomic descriptions of behavior which occur in generalizations over behavioral types are intentional, i.e. such descriptions of behavior must respect the semantic contents of the mental states which produce behavior. This principle provides the basis for an argument for the ineliminability of a semantic component from a completed psychological theory. The concept of intentional explanation is examined and it is argued that though behavioral explanation must be cast in the intentional format, intentional explanation ought not to be constrained by normative rationality assumptions. Since intentional explanation, in many cases, requires reference to the mental states of the agents of behavior, it is essential to understand how mental states are to be individuated. We argue that functional criteria are inadequate for the individuation of mental states such as propositional attitudes, e.g. the belief that _P_, the belief that _Q_, the desire that _R_, et cetera. Widely known criticisms of the possibility of a concept of the semantic equivalence of beliefs are examined and rejected and semantic criteria for the individuation of beliefs are offered.

# ACKNOWLEDGEMENTS

I would like to express my gratitude to The Centre for Cognitive Science at The University of Western Ontario, and especially to its director Professor Zenon Pylyshyn, for support during the completion of this work. Without the background provided by Professor Pylyshyn's inspired work in the foundations of cognitive science the present study would never have materialized. I would also like to sincerely thank my supervisor during these last two years, Professor William Demopoulos, for his lively interest in the issues treated herein and for the encouragement he provided. Thanks are also due to Professor Ausonio Marras for his guidance during the early stages of my study of cognitive science and for several helpful suggestions which I have gladly incorporated. Professor Bruce Freed read the entire work carefully and offered many encouraging observations while providing indispensible discussion of many of the fundamental issues raised in this study. I would also like to thank Dr. Edward Stabler for penetrating discussion of the argument constructed in Chapter I for the ineliminability of semantics.

## TABLE OF CONTENTS

# INTRODUCTION

The purpose of these introductory remarks is to provide background for and a brief précis of some of the major issues with which this study of psychological explanation is concerned. Regarding the question of background, let us acknowledge at the outset that the approach to psychological explanation adopted in this study lies squarely within a rediscovered tradition in the philosophy of mind. The representational theory of mind, as Fodor takes pains to point out, can be associated with philosophers as diverse as Descartes and Hume. Only recently, however, has the representational theory of mind found new and vigorous proponents [Fodor, 1975, 1981; Pylyshyn, 1980, forthcoming]. An important mark of the representational theory of mind is that it treats certain processes in cognitive systems as interpretive operations and, thus, construes certain mental states as non-extensional interpretations of the circumstances in which cognitive systems find themselves. To say that cognitive processes, or the mental states they produce, are interpretive is to say, inter alia, that the contents of mental states are not always strictly determined by the stimulus conditions with which they are associated. This point can be easily and strikingly illustrated by any one of many different examples. Borrowing an example from

Fodor and Pylyshyn [1981], which we will modify only slightly, consider the case of two seafarers lost at sea. After many days of overcast skies, the clouds clear one night and each seafarer spies a certain star which happens to be Polaris. The physical stimulus conditions are equivalent for each soul lost at sea, but the mental states which each forms subsequent to sighting the star may differ markedly. For example, one agent might reflect upon the benign indifference of the universe to his lot, remaining as lost as can be. The other agent might come to believe that "the star I spy is The North Star" and immediately make for landfall, happy to be saved from impending doom. Each agent's interpretation or representation of what he sees differs although in some sense, i.e. in the extensional sense, what each sees is the same thing. A theory of cognitive processes must provide some way to accommodate the fact that this example illustrates: Cognitive agents construct non-extensional or intensional interpretations of their environmental conditions. The mental representation thesis is invoked precisely for the purpose of accommodating this fact since it allows us to account for certain differences between mental states which are associated with the same stimulus conditions. Now some other construct might be tailored for this purpose, but representational theorists will argue that any adequate theory of mental states will share essential features of the mental representation thesis.

In Chapter I of this work, for example, we argue that any adequate psychological theory must provide a systematic basis for the semantic interpretation of mental states -- this is tantamount to the claim that at least some mental states must be viewed as semantically endowed representations.

In its present incarnation, the representational theory of mind is closely associated with the theoretical framework commonly referred to as computational psychology. We may think of computational psychology as a particular version of the representational theory of mind. According to the computational perspective, mental representations may be attributed two distinct sorts of properties. One sort of property attributable to mental representations has to do with the form of the realizations of mental representations in cognitive systems; the other sort of property has to do with the semantic contents which endow mental representations. Properties of the first sort, i.e. those involving aspects of the form of representations, are construed on analogy with the formal properties of sentences. The rough idea is that mental representations can be thought of as syntactically articulated items in a language of thought. The properties exemplified by a representation in virtue of its syntactic structure are referred to simply as the formal properties of the representation. An important aspect of computational psychology is the thesis that properties of the second sort mentioned, i.e. those involving the semantic contents of mental representations, are conveyed (represented) by the formal properties of representations. This principle is sometimes put by saying that all semantic distinctions between particular representations, relevant to psychological explanation, are captured by syntactic or formal distinctions between the representations. However, th~~is~~ construal of the relation between the semantic properties and the formal properties of representations is excessively austere. We might simply say, instead, that the semantic contents of a representation are encoded by or conveyed by its formal properties. One reason that this latter construal is preferred is that we want to be able to assign actual interpretations to token representations; it is not enough merely to be able to determine when an arbitrary pair of representations have different interpretations. The important point in all of this is that computational psychology, as we will see in much more detail, carves out a theoretical and explanatory role for reference both to the formal properties and to the semantic properties of mental representations.

## SEMANTICS AND INTENTIONAL EXPLANATION

These brief comments concerning the representational theory of mind and computational psychology indicate that both views allow for, or even demand, the attribution of semantic properties to mental states. But how does psychological explanation become implicated in the issue of the semantic contents of mental states? One reason that psychological explanation is implicated in reference to the semantic contents of mental states has to do with the problem of taxonomizing behavior. In Chapter I we argue that any psychologically adequate taxonomy of behavior must respect the intended interpretations under which tokens of behavior are issued. We refer to a taxonomy of this type as an *intentional taxonomy*. Further, we argue that the intended interpretation of a token behavior is fixed by the contents of the mental states which produce the token. Thus, the problem of providing an adequate taxonomy of behavior can be solved only by attending to the semantic contents of mental states.

The notion of an intentional taxonomy of behavior allows us to characterize *intentional explanation* in the following manner. The subsumption of a token of behavior under an explanatory generalization is an instance of intentional behavioral explanation just in case the generalization is defined over an intentionally taxonomized behavioral type. The idea is simply that intentional explanation invokes law-like generalizations over behavioral types specified under interpretations, i.e. over behavioral types specified intentionally. Thus, we argue in Chapter I that much, if not all, behavioral explanation should be construed as intentional explanation.

Against this view, it has been argued, principally by S. Stich [1980], that a version of computational psychology is available which makes no reference to the semantic contents of mental states. If this challenge could be sustained, then our argument for the construal of behavioral explanation as intentional explanation would be in jeopardy. For if the semantic properties of mental states could be wholly ignored on some adequate framework for the explanation of behavior, intentional explanation could be eliminated in favor of what we call a "formalist" approach to the explanation of behavior. This challenge to the idea of the intentional explanation of behavior is gladly accepted for it provides a forum for a detailed argument for the ineliminability of a semantic component from psychological theory. The argument presented in Chapter I for the ineliminability of reference to the semantic properties of mental states serves to secure the concept of intentional explanation against the suggestion that all behavioral types accommodated to intentional explanation have fully adequate non-intentional explanations.

INTENTIONAL EXPLANATION AND RATIONALITY

There is another important challenge to the idea of intentional explanation to which we turn our attention in Chapter II. Here, the challenge is to the empirical status of intentional explanation. This challenge has been launched, principally, by D. Dennett [1978] and concerns the role of normative principles in the foundations of intentional explanation. According to Dennett, the theorist engages in intentional explanation just in case he makes certain normative assumptions about the systems whose behavior is to be explained. In order to understand why Dennett believes that intentional explanation has little or no empirical relevance it is necessary to understand the concept of a "normative principle" that he employs. Notice first that there are "norms" of essentially two distinct types: There are descriptive norms and there are prescriptive norms. In the sense in which a norm is descriptive, it simply captures or represents the normal tendencies of a system or class of systems. A descriptive norm conveys a contingent fact about some system or class of systems. In the sense in which a norm is prescriptive, it characterizes a preferred state of affairs, the state of affairs that rationally or morally <u>ought</u> to obtain, for a system or class of systems. Traditionally, there is but a single constraint on the specification of prescriptive norms: Such norms must not prescribe what is impossible for a system --

"Everything that ought to be done can be done."

However, prescriptive norms can and often do set standards above what is descriptively normal, although not above what is possible. For example, in war such things as the injury of noncombatants, propaganda, and unrestrained jingoism are normal. But clearly, it is possible to prescribe standards according to which certain activities normally associated with war are impermissible. The contrast between prescriptive and descriptive norms has clear illustrations in the domain of rationality as well. For example, the fallacy of affirmation of the consequent might constitute a normal tendency of a given system, but this form of inference is among those prohibited by our prescriptive norms of rationality. Throughout our study of Dennett's construal of intentional explanation we will systematically use the term 'normative principle' in the sense associated with <u>prescriptive norms</u>. This is, as we will see, the way that Dennett uses this term.

Now, Dennett argues that normative principles are necessary constraints on the intentional explanation of behavior. In particular, for Dennett intentional explanation is a form of explanation conditioned by the assumption that the agent whose behavior is to be explained will do <u>what rationally ought to be done</u>. This understanding of intentional explanation threatens its empirical relevance since the normative rationality principles which putatively constrain intentional accounts of behavior set standards above what is descriptively normal. Thus, in Chapter II we are concerned to argue that Dennett's construal of intentional explanation is inappropriate.

One way to understand Dennett's notion of intentional explanation is the following. Intentional explanation, *a la* Dennett, accommodates only instances of behavior which are rationally *justified* by the mental states which produce those instances of behavior; intentional explanation is just explanation by justification and rationally unjustified instances of behavior can have no intentional explanation. Now it has been suggested by thoughtful readers of our study of Dennett that one might charge Dennett with a conflation of the concepts of justification and explanation. For, if the normative principles which putatively constrain intentional explanation set standards above the empirical, descriptive norm, then the cognitive processes cited in the *explanation* of tokens of behavior will not always provide *justifications* of those tokens of behavior. However, it is somewhat unfair to accuse Dennett of the conflation of the concepts of explanation and justification. Dennett's *theory* of intentional explanation is that instances of such explanations always offer justifications of behavior. Moreover, it is the fact that intentional explanations, construed as attempts at justification, do not always explain instances of behavior that allows Dennett to inveigh against the empirical adequacy of intentional explanation. This is the wrong way to view intentional explanation, or so we will argue, but it is not an artless conflation of the concepts of explanation and justification.

## SPECIFICATIONS OF MENTAL STATES

Our arguments in Chapters I and II attempt to demonstrate, among other things, (i) that intentional explanation is required for many important types of behavior, (ii) that, contrary to Stich, the explanation of many important types of behavior requires recourse to a framework for the semantic interpretation of mental states, and (iii) that, contrary to Dennett, intentional explanation is an empirically significant form of explanation. Against the background provided by these theses, in Chapter III we embark upon a study of the problem of the individuation mental states. Since mental state types are referred to in the explanans of generalizations explanatory of intentional behavioral types, the theorist must have access to a procedure for assigning token mental states to their relevant types. That is, an intentional explanation of a token of behavior, B, is given by recourse to a generalization or to generalizations which specify the mental state types which cause or produce tokens of the type to which B belongs.

Thus, employing such generalizations in the intentional explanation of behavior requires some mechanism for the specification of the types to which token mental states belong.

There are two main approaches to the problem of the individuation of mental states. One can be broadly construed as a functional approach while the other, and the one that we favor, can be construed as a semantic approach. Many recent works in the philosophy of psychology are typified by an uncritical trust in the applicability of the resources of functionalism to the problem of individuating mental state types. This is ironic since functionalism has been severely and rather convincingly criticized in a number of quarters precisely for its inability to provide adequate criteria for the individuation of certain types of mental states. The most widely accepted criticisms of functionalism in this area concern the problem of individuating qualitative states such as pain and the perception of color. In Chapter III we argue that functional criteria are inadequate for the individuation of propositional states, i.e. propositional attitudes, such as the belief that P and the belief that Q. It is rather surprising that the ability of functional criteria to individuate propositional attitudes has not been examined in detail by either functionalists or by non-functionalists, while the inability of functionalism to individuate qualitative states is admitted even by ardent functionalists like Fodor [see Fodor, 1981, pp. 18-19]. Since we won't have occasion to concern ourselves with the criticisms of functionalism in connection with the problem of qualitative states, we might usefully, if briefly, rehearse one of those criticisms here. In essence, functionalism is the doctrine that the type to which a token mental state belongs is determined by the causal role that the token plays in the life of the organism in which it occurs. Thus, mental state types are to be defined, according to this view, by specifying their causal-functional roles. For example, the causal-functional role of a certain mental state type, _M_, might be "to indicate the presence of red objects in the physical environment and to underwrite red-object-directed behavior". This construal of the causal-functional role of the state type _M_ is very rough, but it allows us to illustrate the well known "qualia inversion" argument [Fodor, 1981, p. 19]. Although mental states, no doubt, possess causal-functional roles, causal-functional specifications of mental states such as the one given for the state type _M_ appear to be intrinsically incomplete: All those token mental states which serve "to indicate the presence of red objects in the physical environment and to underwrite red-object-directed behavior" need not share all interesting psychological properties. In particular, the mental state type which fills just this causal-functional role in an agent $\underline{S}_1$ might have the qualitative properties associated with the perception of blue objects in a normal agent $\underline{S}_2$, i.e. the token states which signal the presence of red objects for $\underline{S}_1$ may have the qualia normally associated with the perception of blue objects. Some functionalists, like Fodor, are resigned to the idea that functional specifications of qualitative states are intrinsically incomplete. But at the same time, it is often assumed that the prospects for the functional individuation of beliefs are much brighter. In Chapter III we examine the prospects for the functional individuation of beliefs and argue that they are not as bright as they need to be if belief types are to be accurately individuated.

We are left with the problem of individuating mental states in some manner that does not rely exclusively upon functional criteria. In this connection, we suggest that the type-membership conditions for token beliefs, and other propositional attitude tokens, should be given in terms of the semantic contents of beliefs. Adopting this approach and proposing actual semantic criteria for the type-identity of beliefs requires that we offer some treatment of the problems which H. Putnam charges against any concept of the equivalence of the contents of mental states. Such a treatment of Putnam's objections is offered in Chapter III along with proposed criteria for the semantic equivalence of beliefs.

It may be helpful to reiterate some of the central principles for which we argue in this study of psychological explanation. Some of the following principles are not completely unknown, but where they are widely known we provide novel arguments for their support.

(1) An adequate psychological theory must offer explanations for tokens of behavior from intentionally taxonomized types.

(2) Reference to the semantic properties of mental states is ineliminable from psychological explanation.

(3) Normative rationality principles are not necessary constraints on intentional explanations of behavior; intentional explanation can be construed as an empirically significant form of explanation.

(4) Functional criteria are inadequate for the task of individuating propositional attitudes.

(5) The type-identity conditions for certain mental state types are to be given in terms of their semantic contents.

As we observed earlier, these themes lie squarely within the context provided by the representational theory of mind and are compatible, or so we would argue, with computational psychology. As one can readily appreciate, some of the views which we are concerned to establish constitute criticisms of widely known approaches to psychological explanation. But this is as it must be since much that has been said or assumed by recent commentators on the foundations of psychology stands in the way of an understanding of the role that a theory for the semantic interpretation of mental states and behavior plays in the formulation of explanatory psychological generalizations. It is our express hope that the themes and lines of argumentation pursued here are complementary to the rich and exciting program for psychological theory that has been launched by such theorists as Jerry Fodor and Zenon Pylyshyn.

Let me take this opportunity to enter a note concerning a convention for specifications of beliefs adopted in this work. I will often place quotation marks around expressions which specify the contents of beliefs in order to indicate that the attributed beliefs are assumed to possess determinate formal structures, as well as contents. For example, we might attribute to some agent the belief that "rain is wet". The idea behind the use of quotation marks in this role is simply to provide a device that indicates that the attributed beliefs have specific forms, although their forms are not assumed to be indistinguishable from those of the English sentences used -- hence single quotes are not recommended for the general case. But, we will assume throughout that beliefs, when internalized by agents, do have formal structures of some kind or other.

# CHAPTER 1

## TAXONOMIZING THE DOMAIN OF EXPLANATION

### OVERT BEHAVIORAL TYPES

How are we to taxonomize the overt, observable, dated instances of behavior issued by psychological agents? Ought we restrict ourselves to reference to the observable properties of the events with which individual tokens of behavior are identified? Or, must we refer to certain covert conditions under which individual tokens of behavior are issued? This pair of alternatives represents a fundamental tension in the theory construction process in psychology. The explanation of instances of behavior by subsumption under explanatory generalizations logically presupposes a taxonomy of behavior that assigns tokens of behavior to relevant behavioral types. But, importantly, different classes of phenomena are circumscribed by differently motivated taxonomies of behavior. On the one hand, there are physical properties of the occurrences with which tokens of behavior are identified that provide easily accessible, objective, and measurable properties over which taxonomic rules might be defined. On the other hand, there are properties that tokens of behavior possess in virtue of the role for which agents intend the behavior they issue and such properties are not identifiable with the physical properties of overt behavior. May we taxonomize behavioral phenomena by defining type-identity conditions over observable physical properties and thereby render the task of individuating psychological types relatively simple or must we, instead, taxonomize behavior in such a way so as to respect the interpretations that agents formulate of their own actions?

According to the view to be argued here, the choice we make between these alternatives crucially affects the completeness of psychological theory. For, there exist important behavioral types that cannot be specified by defining a similarity measure over physical parameters of tokens of overt behavior. That is, there are specifiable types of behavior that cannot be captured within the nomological net, if we restrict type-identity conditions for behavior to reference to the observable physical properties of behavior. Tokens of behavior, in a great many cases at least, are assigned to psychologically relevant equivalence classes on a basis other than the physical form that they take. Consider the explanations of two instances of behavior that closely resemble each other in all observable physical properties. Each instance is constituted, let us imagine, by a sequence of bodily movements that realize the events described in the following way. An individual gains entry to a bank and proceeds to drill out the tumblers of the bank's safe; he accidentally triggers an alarm and shortly thereafter the police arrive. One instance of such an occurrence might be assigned to the type "attempted bank robbery", while another instance of an overtly indentical occurrence is assigned to the type "service by bonded locksmith". Moreover, an adequate explanation of each sequence of events must reflect this difference--perhaps by appeal to internal conditions of the agents that produce these tokens of behavior, e.g. what goals do the respective agents seek to fulfill by their actions? An explanation of the attempted robbery might begin with a specification of the beliefs and desires out of which the agent acts. The would-be thief may believe, for example, that crime pays and may desire easy wealth. The locksmith, on the other hand, presumably acts not out of a desire for easy wealth but, rather, out of a desire to render service to a client. It appears, then, that in some cases we must define type-identity conditions for behavior over such things as the intentions under which behavior is issued. More broadly, type-identity conditions for behavior are to be defined over what will be termed the "intended interpretations" of behavior, i.e. those interpretations which are generated by the agents of behavior themselves.

In any case, behavioral types, like those in the examples and the majority of inchoate behavioral types of folk psychology, are not specifiable in a physicalistic idiom that refers only to properties of the events with which tokens of overt behavior are identified. The reason is, quite simply, that there are no finitary physical descriptions, so restricted, that will pick out all and only the tokens of certain types.[1] The crucial issue here does not concern the adequacy or inadequacy of the taxonomic categories of folk-psychology for conceptualizing behavior, but rather the relevance and adequacy of reference to observable physical properties in the individuation of interesting behavioral types--types for which explanatory generalizations are sought. Now, some commentators on the foundations of psychological theory are unimpressed by garden variety examples like those offered above and suggest that intuitively specified behavioral types do not constitute phenomena that deserve systematic explanation. One possible view is that if the taxonomic categories of folk psychology are not coextensive with types specifiable by reference to the physical properties of overt behavior, then the folk taxonomy is to be eschewed in the theory construction process.[2] But, as we will see, this approach would force theory to ignore fundamental distinctions between various psychological or cognitive functions. A more sophisticated view, though a view in the same vein, was once championed by D. Dennett. In a widely read work Dennett suggests that the behavioral phenomena amenable to systematic explanation should be, roughly, physicalistically construed. Remarking on a hypothetical theory of the behavior of poor chess players Dennett maintains:

> If one wants to get away from norms and predict and explain the 'actual, empirical' behavior of the poor chess-players, one stops talking of their *chess moves* and starts talking of their *proclivities* to move pieces of wood or ivory about on checkered boards; if one wants to predict and explain the 'actual, empirical' behavior of believers, one must similarly cease talking of belief and descend to the design stance or physical stance for one's account. [Dennett, 1978, p. 22]

One of Dennett's arguments for this assertion appears to rest on the claim that when we construe behavior *intentionally*, that is, when we specify the type to which a token belongs in a way that respects the intentions under which the behavior is issued, e.g. when we take a token of behavior to constitute a "chess move" rather than "a motion of an arm that transports a piece of ivory across a checkered board", we make an assumption of optimal rationality. But, Dennett argues, psychological agents are not optimally rational--certainly poor chess players are not optimally rational. Since the normative assumption that makes intentional construals of behavior possible is not descriptive of the agents of behavior, Dennett reasons that intentional construals of behavior cannot represent the "actual, empirical" behavior of agents.

A central tenet of the view to be defended here is that, contrary to certain notions of "actual, empirical" behavior, the behavior of human agents must be intentionally construed for no non-intentional construal is capable of capturing certain uncontroversial, paradigmatic cognitive functions. In particular, it can be shown that any theory that taxonomizes behavior in a way insensitive to the intended interpretations under which behavior is issued cannot correctly identify, and *a fortiori* cannot explain, certain tokens of behavior and the types to which they belong.

Three related issues will be treated during the course of the argument of this chapter. First, we will rehearse a rather familiar argument to the effect that physical properties, in general, are largely irrelevant to the assignment of tokens of behavior to psychologically relevant types. Second, we will consider the tenability of a hypothetical psychological theory that predicts "bodily movements" by appeal to the beliefs and desires that produce such movements. Third, and last in this chapter, we will argue that there is no purely formal alternative to an intentional taxonomy of behavior, i.e. that no formal alternative has the capacity to capture the types of behavioral phenomena which are captured by intentional construals of behavior.

## Physicalistic Taxonomies and Intentional Taxonomies

Why are the physical properties of tokens of behavior irrelevant, in many cases at least, to a determination of the kind or type to which those tokens belong? To adopt an example preferred for its simplicity consider the behavioral type constituted by correct answers to a certain arithmetic problem. To a first approximation, all those subjects that correctly give the sum of two numbers, 2 and 3 say, issue type-identical tokens of behavior. Now, the fact that a subject gives the correct answer does not show, by itself, that any particular cognitive procedure has been applied and, hence, our first approximation to the behavioral type may have a finer-grained resolution. Consider in this regard the possibility that one subject may simply recall the answer from memory while another subject undertakes the execution of some more or less rigorous procedures, e.g. counting on her fingers, to determine the sum. It is important to be able to distinguish tokens of behavior that are the products of different kinds of processes if we are to characterize the cognitive ability, or the knowledge of a subject matter, possessed by various subjects. We would not, for example, attribute full knowledge of the addition function to someone who had merely memorized a set of sums. But this is only to point out that the behavioral type "correct answers to (PLUS 2  3)" has interesting sub-types. The important point is that the types constituted by correct answers to particular addition problems are ones for which interesting generalizations might be formulated. In the case of very simple problems, the explanation of a subject's "arithmetic ability" might very well refer only to the subject's capacity to maintain a long term memory and to the memory access function. In the case of problems for which a subject has no memory of the correct answer, the relevant generalizations will no doubt refer to algorithmic procedures that the subject has internalized. All of this is relatively uncontroversial. But how are we to determine to which tokens of behavior a given generalization of one sort or another might apply? How are we to identify instances of the type "correct answers to (PLUS 2 3)" itself? In particular, is there a set of physical properties of overt behavior by reference to which we may pick out all and only the tokens of this type or of its various sub-types?

Although the behavioral type constituted by correct answers to a particular arithmetic problem is not a type that theorists usually think of under the banner of "folk psychology", notice that this type shares an important feature of most folk-psychological types: Tokens of the type may take infinitely many diverse physical forms. For example, a subject may respond to an arithmetic question verbally in any natural language, in written script of different forms, e.g. English words, French words, Roman numerals, Arabic numerals, et cetera. A subject's answer might be given in ASL, by tapping on the floor a number of times, or in smoke signals. Furthermore, if one considers any one of these possible forms, it should be clear that token occurrences of an answer in any one form may diverge from each other in a host of properties, e.g. token utterances of the same English word will differ along a continuum of frequency and amplitude.

Given the wide range of physical occurrences that can be identified with tokens of the type "correct answers to (PLUS 2 3)", our problem is to specify those properties of tokens of the type over which type-identity conditions are to be defined. When we restrict our attention to the physical properties of overt behavior only two tactics are available: Either (i) specify a single set of physical properties definitive of the type or (ii) specify a finite disjunction of sets of physical properties such that each disjunct is individually sufficient for determining a token's membership in the type. Neither tactic holds promise.

The tokens of a behavioral type, for many such types including arithmetic problem-solving behavior, need not share a common physical description by virtue of falling under which they are assigned to the type. Consider in this connection a physical comparison of the utterance "five" and the corresponding ASL hand signal. The two overt occurrences do not possess a common physical description at any level that is theoretically relevant to their assignment to the same type. Indeed, were someone to abstract a putatively relevant description of physical properties common to both occurrences, it would be a simple matter to devise a code to convey the correct answer that called for physical movements that do not instantiate the properties common to the utterance "five" and the corresponding sign in ASL. Since this manoeuvre could always be employed in principle, it is clear that what makes a token an instance of a particular type is not its possession of a definitive set of physical properties.[3]

The second alternative--the specification of a finite disjunction of properties--is ruled out by the following considerations. Tokens of behavior may be physically type-identical and yet not belong to the same behavioral type. Examples of this genus are easily generated once one observes that a specific set of bodily movements, of any arbitrarily high degree of similarity, may be initiated on different occasions in order to accomplish very different tasks. Consider the motion of one arm that traces out a certain trajectory. Just that motion may be done by an individual, or by different individuals, in order to signal to someone, to issue a vote, to ease a pain in the shoulder, or to move a chess piece across a vertical board. Or consider the utterance of the term 'five' on different occasions. That speech act may constitute an answer to an arithmetic problem, an answer to a request for the time of day, an answer to a request for a child's age, or an answer to any one of infinitely many different questions. Thus, any disjunction of physical properties that purports to specify all possible physical forms of tokens of a given type will have within its extension tokens of distinct types. A more general conclusion that can be derived from these considerations is that since tokens of behavior of different significances, hence of different types, may be overtly realized by physically type-identical movements, the psychological characterization of what an agent does cannot be given by a physical description of the bodily movements that the agent undergoes.

Notice also that no gain is made by holding that the type-identity of token outputs is contingent upon input type. The situation we face when attempting to taxonomize output by reference to physical properties obtains, also, when we attempt to specify the input types to cognitive systems. Just as there are infinitely many diverse physical forms in which equivalent answers to an arithmetic problem may be given, there are infinitely many physically diverse forms in which the same arithmetic question may be put. A question might be posed verbally in any language understood by a subject, in numerous different written forms, or in a tactile code arranged with a subject. Token inputs may have nothing physically specifiable in common that is relevant to their assignment to a certain equivalence class. Further, there is no reason to believe that there is a finite set of physical properties such that each element of the set is individually sufficient for the assignment of an input token to a psychologically relevant type.

We have arrived at the conclusion that the type to which a token of behavior belongs cannot, in the typical case, be determined by reference to the physical properties of the token's overt manifestation. This conclusion should be viewed as a preliminary stage in a larger argument for the inadequacy of all taxonomies that do not respect the contents of the mental states upon which behavior is dependent. Notice, incidentally, that there are exceptions but no counterexamples to the rule. The overt behavioral type "perfect swan dive", for example, presumably answers to physicalistic criteria. Such a behavioral type does not constitute a counterexample to our preliminary conclusion simply because it has not been claimed that the irrelevance of physical properties to specifications of behavioral types is <u>criterial</u> to the concept of a behavioral type. Notice, too, that the discussion has been restricted to the possibility of physicalistic taxonomies that advert to properties of <u>overt</u> behavior. But, this restriction is interesting since it reveals the need to allow reference to covert processes when individuating behavioral types. Though considerations will be introduced later which serve to undercut the apparent possibility, for all we have said to this point it is possible that tokens of behavior belong to the same type just in case they fall under some unique description of covert physical properties--a possibility that D. Davidson has entertained [Davidson, 1981, p. 252]. It may be useful to consider the requirements for a physicalistic taxonomy of behavior based upon reference to covert properties. Now, we know that there are, for example, infinitely many ways to compute sums, infinitely many different algorithms for addition, and indefinitely many different physical systems in which a given algorithm can be realized. Hence, if behavior of the type "solutions to addition problems" depends upon the employment of an algorithm for addition, then there are infinitely many different covert physical processes that can eventuate in behavior of that type. This circumstance places heavy demands on a physicalistic taxonomy. In particular, it requires the effective enumerability of the distinct physical states or processes which constitute realizations of addition algorithms--something which we have little reason to expect. Moreover, it requires that the physical processes which can realize an algorithm for addition never realize anything else--a requirement that appears difficult to satisfy short of fixing positively all physical properties of the system. In any case, considerations which will later be urged against "formal" taxonomies of behavior generalize in rather obvious ways to physicalistic taxonomies which advert to covert physical processes. Suffice it to say at this juncture that it is hard, from the methodological perspective, to take seriously the possibility of such a physicalistic taxonomy since we have virtually no idea of what the relevant covert physical properties might be for any particular behavioral type, yet we have substantive concepts of many behavioral types. In the remainder of this section, we will disregard the possibility of taxonomies which advert to covert (and unknown) physical properties in order that we may more sharply contrast the idea of an intentional taxonomy with the idea of a non-intentional taxonomy of overt behavior.

On what basis then is behavior to be taxonomized? In a sense, we must already know the answer to this question since we freely and easily recognize the types to which most tokens of overt behavior belong. It is as if one were to ask "On what basis is a sentence to be judged grammatical?" Native speakers tacitly know the basis relevant to judgements of grammaticality and regularly employ that basis when identifying utterances as sentences in the native language that they have acquired. Now, according to an intentional taxonomy, a necessary condition for the type-identity of tokens of behavior is that they are issued under equivalent intended interpretations. A subject that utters the term 'five' and a subject that inscribes '5' on a piece of paper issue type-identical behavior only if the contents that the tokens of behavior are intended to convey are equivalent. Each subject may intend to convey information concerning the time of day, to supply an answer to the question "What is the square root of 25?" or something else entirely. But we cannot specify the type to which a token belongs unless we can approximate the conceptualization of the behavior formed by its agent. In particular, since physically type-identical bodily movements may realize either type-identical or type-distinct tokens of behavior, depending upon the interpretations under which they are issued, it is mandatory that we refer to, infer, or otherwise approximate the intended interpretations under which the movements are carried out when taxonomizing behavior.[4]

Although each token of overt behavior is identified with some token physical occurrence, an intentional taxonomy of behavior does not rely upon the physical properties of the occurrences with which tokens of behavior are identified when defining type-identity conditions for behavior. An intentional taxonomy constructs interpretations of the overt occurrences that realize tokens of behavior and, hence, allows for the type-identification of tokens that may have little or no physical resemblance to one another.

Interpretation allows us to assign a wide range of physically diverse occurrences to a common type. Not surprisingly, this feature of interpretation parallels a feature of behavior: There are many physically diverse ways in which a behavior of a given type may be done.

A non-intentional taxonomy of behavior, on the other hand, does not refer to the intended interpretations of behavior or offer intentional interpretations for the occurrences with which tokens of behavior are identified. Such a taxonomy is charged with the task of classifying tokens by reference to various overt, observable physical properties and is bound to incorrectly classify certain tokens of types picked out by an intentional taxonomy: A non-intentional type, defined by any set of physical properties or any disjunction of physical properties, will have in its extension tokens of distinct intentional types. For, any overt behavior, under the control of an agent, which instantiates some particular set of physical properties can be utilized to convey any piece of information that the agent may wish to convey. All that is required is that there exist individuals who know how to interpret the agent's behavior -- individuals, that is, who know the "code" the agent employs. Thus, a specification of physical properties of overt behavior will pick out tokens of behavior that from the intentional perspective on behavior are type-distinct. But the intentional perspective is to be preferred since it has an important advantage. It has the capacity to capture behavioral types such as "answers to (PLUS 2 3)". These are types that non-intentional schemes fail to individuate, i.e. the non-intentional alternative cannot specify a class co-extensive with "answers to (PLUS 2 3)". Insofar as such cognitive abilities as arithmetic problem-solving represent phenomena for which explanatory generalizations are sought, we appear to be committed to the intentional interpretation of behavior.[5]

Intentionality and Intensionality

The intentionality of psychological states, roughly their contentfulness, is connected in complex ways with what we might call the intensionality of their forms. But, the concepts are not equivalent and it may be useful to post a warning against their conflation. The idea that what we have termed intentional construals of behavior are required if behavior is to be adequately taxonomized should not be confused with a doctrine about the logical properties of sentences that specify psychological phenomena, i.e. phenomena amenable to subsumption under psychological generalizations. We have suggested that only the intentional interpretation of behavior can provide for a taxonomy that captures certain interesting psychological types--that to be made amenable to subsumption under generalizations about such things as "arithmetic problem-solving" behavior must be intentionally construed. Intentional construals are simply characterizations or descriptions of behavior that respect the intended interpretations under which behavior is issued. Such intended interpretations are determined by the content of the mental states, e.g. beliefs and desires, upon which behavior is contingent. Now a mental state has a content only if it has a subject matter and, hence, our use of the term 'intentional' is meant to be continuous with Brentano's [1874]. Unlike Brentano, however, we leave open the possibility of a system that is physical and yet intentional.

It is important not to conflate the concept of intentionality used here with the concept of certain properties of the logical form of certain sentences glossed as intensional properties. Historically, Chisholm's attempt to specify the class of sentences about psychological phenomena by reference to various logical peculiarities initiated a useful literature. But, the thesis that sentences about intentional phenomena have a certain logical form, what we may call logical intensionality, can be a source of considerable confusion. A recent example is provided by M. Boden's claim that the "logical features of intensionality" characterize all statements descriptive of the behavior of a system guided by internal representations--desires and beliefs [Boden, 1970, 1972]. Boden goes so far as to claim that the logical intensionality putatively characteristic of descriptions of a machine's behavior is a sufficient basis for the attribution of intentionality to physical systems:

> [The] notion of "model" provides a basis for the ascription of intensionality [sic] to machines....Insofar as a machine's performance is guided by its internal, perhaps idiosyncratic model of the environment, the overall performance is describable in intensional terms. That is, the logical features of intensionality mentioned earlier [as analyzed by Chisholm] will characterize any statements made about the machines to describe or explain performance guided by the model. [Boden, 1972, p. 128]

Though Boden rightly requires that descriptions of the behavior of a device guided by an internal representational system or a "model" must respect the content of the representations that guide its behavior, this constraint on descriptions of behavior does not restrict such descriptions to logically intensional forms.

For our purposes, it is important to see that an assessment of certain logical properties of descriptions of behavior is independent of an assessment of the theoretical adequacy of descriptions of behavior. In particular, logical intensionality is not a necessary condition for an intentional construal of behavior. Although we require that type-identity conditions are defined, in part, over the intended interpretations of behavior, this does not imply that all proprietary descriptions of behavior are logically intensional. Of course, many descriptions of behavior are logically intensional. Paradigmatic examples employ verbs like 'hunt', 'search', 'ridicule', and 'worship'.

(a) Sam searched for a unicorn.
(b) Sam worshiped Belial.

The intensionality of (a) and (b) is shown by the failure of the substitution of co-referring terms and/or the failure of existential generalization, salva veritate. For example, suppose that 'unicorn = a mythical one-horned horse-like creature'; the substitution of the coreferential term for 'unicorn' will not preserve the truth of (a). And, though Sam worships Belial, it does not follow that '(∃x) (Sam worships x)'.

But suppose that the following descriptions of pieces of Sam's behavior are tendered:

(c) Sam stabbed Joe.
(d) Sam hit the floor.

If it is true that Sam stabbed Joe and that 'Joe = Dick's best friend', then it is true that,

(e) Sam stabbed Dick's best friend.

and that,

(f) (∃x) (Sam stabbed x).

Analogous inferences are licensed by the substitution of identity and by existential generalization for (d). Thus, (c) and (d) are not logically intensional according to the two most widely accepted tests for logical intensionality and not, incidentally, logically intensional according to other criteria that Chisholm has suggested.[6] Boden's view, curiously, would prevent us from employing descriptions like (c) and (d) in construals of Sam's behavior and, in turn, preclude the inclusion of extensional verbs like 'hit' and 'stab' in a cognitive system's model or representation of the world. For Boden holds that "the logical features of intensionality" will characterize all descriptions of a cognitive system's behavior. This restriction on admissible terms in theoretically relevant construals of behavior is without motivation. We can acknowledge that proprietary descriptions of behavior must respect the interpretations under which behavior is issued--Sam conceives of his action as "stabbing Joe" and not as "stabbing Dick's best friend"--without thereby requiring that such descriptions have a logically intensional form.

In an interesting paper, A. Marras suggests that the classes of intentional phenomena can be delimited by a criterion that respects only the logical intensionality of their descriptions [forthcoming]. Though Marras' view may be essentially correct, it is important to see that no restrictions on descriptions of behavior follow from the view.

Marras suggests that an expression is intensional if and only if it is a nontransformable expression that (i) is intensional according to the criteria of existential generalization and/or non-truth-functionality, or (ii) entails an expression that satisfies (i). (A nontransformable expression is simply one which has no non-intensional analysis.) The thesis, then, is that all and only intensional expressions describe intentional phenomena. But, consider the problem of determining whether an expression satisfies (ii). For example, is 'Robby stabbed the foreman' intensional? Determining what sentences are entailed by this description of "Robby's behavior" requires much more than access to a set of truth-preserving principles of inference. What is required is nothing less than a psychological theory of Robby. Which sentences are entailed when, for example, 'Robby' is the name of a servomechanism, a robot in an automated production line, a chimpanzee, or a normal human agent? Under certain conditions, 'Robby stabbed the foreman' can be said to entail 'Robby intended to stab the foreman', but such conditions are to be established by a psychological theory. Thus, it appears that only when we know a great deal about the psychology of a system, can we assess the intensionality of descriptions of its behavior: We will require an intentional theory to identify the nontransformable intensional expressions. In short, intensionally describable phenomena may be coextensive with intentional phenomena, but we cannot employ the concept of intensional describability to delimit the class of phenomena in need of psychological explanation.

Notwithstanding these considerations, we will reserve the term 'intensional' for expressions for which existential generalization and/or the substitution of identity fail, and the term 'intentional' to mean, roughly, contentful. In these terms, the central point ~~here~~ is that there are sentences that pick out phenomena for which psychological explanations are required that do not exhibit the logical properties that mark intensional sentences. Hence, we need not constrain our descriptions of behavior by criteria for logical intensionality: Intentional psychology can and will offer explanations of behavioral phenomena described both by logically intensional and logically extensional sentences.[7]

Theories of Bodily Motion

We have not relied heavily upon the assumption that our folk-psychological intuitions individuate important equivalence classes of behavior but merely upon the assumption that correct answers to a particular arithmetic question, or sub-types of this class, form equivalence classes for which psychological generalizations should be sought. The latter assumption, nevertheless, lends considerable credence to the former assumption. Since the behavioral type "correct answers to (PLUS 2 3)" exemplifies those features of the intuitively individuated behavioral types of folk psychology which tend to make folk-psychological types suspect from the perspective of some theorists, we are justified in eschewing many folk-psychological types only insofar as we are justified in eschewing the type "correct answers to (PLUS 2 3)". Stated in a positive form, the point is that if addition problem-solving behavior is the kind of thing for which psychological theory should seek explanatory generalizations, then other intentional behavioral types, or types individuated by type-identity conditions defined over the intended interpretations of tokens of behavior, are likely within the purview of psychological explanation as well.

In any case, whether or not we choose to integrate extensive portions of folk psychology into a systematic psychological theory, it remains the case that there are certain behavioral types that are not specifiable by reference to objective, measurable, properties of overtly manifest tokens of behavior. Our interest in such types of behavior, which likely constitute the bulk of behavior, leads naturally to an interest in questions concerning the interpretation and content of the mental states that produce behavior--the processes that fix the intended interpretations of tokens of behavior. Such questions, however, are premature to the extent that there are apparent alternatives to a psychological theory which taxonomizes behavior by appeal to such entities as contentful mental states.

In fact, one might acknowledge the points made here about an intentional taxonomy of behavior and yet refuse to view psychology as a theory of, inter alia, events taxonomized under an interpretation. Our argument against this prejudice is based on an appeal to the idea of a complete psychological theory: If generalizations about such things as arithmetic problem-solving behavior are to be included in an ideally completed theory, then it will not do to abjure intentional construals of behavioral types. However, it is sometimes suggested that some or all of the relevant generalizations can be captured without adverting to the interpretations of behavior. In this connection, H. Field envisages the possibility of a species of functionalism that takes as its subject matter not intentional types, but bodily movement types. The theory that Field entertains is a species of functionalism since it appeals to mental states, which are to be individuated in terms of their causes and effects, in the explanation of bodily movement types. The appeal to internal states is presumably intended to afford the theory a measure of explanatory power that would make it a genuine alternative to an intentional psychology. But, as we will see, if internal states such as beliefs and desires are cited, then the lawful regularities, required by Field's putative alternative, between internal states and their causes and effects will not be forthcoming.

The successful prediction of bodily movements would constitute prediction of token physical events which realize tokens of intentionally construed behavior. But, one would not be entitled to consider a theory capable of predicting bodily motion as explanatory of behavior on account of the widely acknowledged opacity of explanation: E may explain P and P = Q, but it does not follow that E explains Q, where 'P = Q' is not logically true. Nevertheless, a theorist might reject a scheme that calls for the formation of generalizations over interpreted types of behavior and hold that the appropriate domain for a systematic psychological theory is provided by bodily movement. The putative advantage of a theory of bodily movement over an intentional psychology is that it dispenses with the need for a semantic component in psychological theory. Field asks,

> Should the semantics of the system of internal representation also be stated as a part of the psychological theory? That depends upon what we want psychological theory for. If the task of psychology is to state
>
> (i) the laws by which an organisms's beliefs and desires evolve as he is subjected to sensory stimulation, and
> (ii) the laws by which those beliefs and desires affect his bodily movements,
>
> then I think that it is clear we do not need to use the semantics of the system of representation in stating the psychological laws. [Field, 1976, pp. 43-44]

The assumption, more or less unargued for, is that if our task is to explain the movements of an organism that are contingent upon the organism's beliefs and desires, then we will not need to formulate interpretations of its behavior or of its beliefs and desires.

It should be noted at the outset that Field actually advocates the inclusion of a semantic component in psychological theory. Field argues that there is an epistemic relation between beliefs and the world and between the beliefs formed by individuals: Beliefs may provide evidence for other beliefs. For example, Carl Sagan tells an audience that the sun is 93 million miles, on average, from the earth. There is a relation between what Sagan believes and what some members of his audience believe, such that what Sagan believes is evidence for the truth of the newly formed belief in members of his audience. The "evidence" relation, Field maintains, can only be captured by a semantic theory since evidence is always evidence for the truth of a proposition or belief, and truth is a semantic notion _par_ _excellence_.[8]

The theory of bodily motion that Field entertains as an alternative to a theory containing a semantic component appears somewhat idiosyncratic: It is, after all, a theory of _bodily_ _movement_ that appeals to _beliefs_ and _desires_. This appearance may be lessened to some degree by the fact that, for Field, the beliefs and desires cited in the explanation of bodily movements are to be individuated in a purely functional fashion, i.e. by reference to their causes and effects, rather than by appeal to their contents.[9] In any case, in the passage cited, Field attempts to adumbrate a style of psychological theory positioned approximately midway between a purely physiological theory of bodily movement and a fully intentional psychology.[10] Just how plausible is such a theory?

Field envisions a theory with two central components. One component of the theory would be constituted by a set of laws that specify the contingencies of belief formation upon sensory stimulation, with a similar account to be given for types of desires. In order that the discussion may be simplified we will consider only the case of belief on the assumption that our conclusions can be easily extended to the case of desire. The other component of the theory would specify laws that correlate bodily movements with pairs of beliefs and desires. Consider the former project: The formulation of the laws by which an organism's beliefs evolve under sensory stimulation. This task requires (a) the individuation of equivalence classes of sensory stimulation and (b) the specification of a mapping from equivalence classes of stimuli into belief types. A crucial presupposition here is that at least some belief types are lawfully related to certain equivalence classes of stimuli. Field makes essentially the same point when he observes that the psychological theory he entertains requires that,

> There is a privileged class of sentences in the system of representation, called the class of observation sentences, with the property that each sentence in the class has associated with it a particular type of sensory stimulation. Whenever a sensory stimulation of the appropriate type occurs, the organism believes the observation sentence. [Field, 1976, p. 46]

Certain beliefs, then, must be under the strict casual control of stimuli. Specifying these beliefs is only one part of the problem confronting the theory that Field has in mind since beliefs other than those identified with observation sentences may affect behavior. Hence, taking the hypothetical theory seriously would require that one specify the laws that characterize the formation of non-observational beliefs upon the formation of observational beliefs. Every belief with functional significance would thereby be either directly or indirectly related to classes of sensory stimulation by the theory.

Can these requirements be met by a theory of belief? Consider the formation of the belief 'there are rabbits nearby' which Field gives as an example of a possible observation sentence. It would not appear that the belief 'there are rabbits nearby' is a causal consequence of a particular sensory stimulation type. One may form the belief subsequent to exposure to rabbit tracks, burrows, droppings, fur, overgrazed vegetation, or rabbit flies; one may form the belief subsequent to auditory stimulation by rabbit squeals, a rustling in the brush, thumps on the ground, or the utterance "There are rabbits nearby". Furthermore, the class of stimuli that might give rise to the belief contains stimuli that do not reliably indicate the presence of rabbits, such as those stimuli provided by a display of thousands of spent cartridges from firearms: The belief might be the product of inferences based upon bad evidence.

Moreover, stimuli that reliably indicate the presence of rabbits need not uniformly produce the relevant belief. A normal perceiver a few feet distant from two large rabbits and whose attention had been directed towards the animals might be expected to believe that 'there are rabbits nearby', but what belief would be attributable to an individual who believes that the alleged rabbit keeper is in reality a clever animator of stuffed toys? The salient point is that tokens of a particular belief type may arise under physically diverse forms of sensory stimulation and that a sensory stimulation of a given type may give rise to nonequivalent beliefs.[11] Field's schema for a theory of bodily movement requires the specification of laws governing belief formation upon sensory stimulation and it is here that such a theory comes to grief. The problem can be stated succinctly: The category "causes the belief 'there are rabbits nearby'" does not form an equivalence class of sensory stimuli; it does not form a physical natural kind. The category includes an unlimited number of diverse physical occurrences which may have no theoretically relevant properties in common other than their tendency to cause the belief 'there are rabbits nearby' in certain subjects. Tokens of the category of sensory stimulation in question are picked out only by reference to their tendency to inculate belief tokens of a certain semantic type and this is not an intrinsic physical property of token stimuli.

Analogous difficulties beset the attempt to specify the laws demanded by the second component of the theory of bodily movement sketched by Field. Here, laws governing the production of bodily movement upon the formation of beliefs and desires are called for. The problem is, by now, a familiar one: For any particular type of bodily movement there are indefinitely many distinct belief-desire pairs that may be implicated in the production of tokens of that movement type. Returning to an example offered above, a movement of one arm that traces out a certain trajectory may be initiated because one believes that a friend is near and one has the desire to greet the individual, because one believes that a mosquito is buzzing around one's head and desires to discourage the pest, or because one feels a pain in the shoulder and believes that exercise will alleviate the pain: Physically individuated movement types are not related in a one-to-one fashion to belief-desire pair types.

Although the preceding point is relatively obvious, it is important to emphasize the fact that tokens of a movement type are often the product of psychologically distinguishable processes: Movements type-identified on the basis of their physical similarity are often representative of different cognitive achievements. This alone is enough to disqualify all schemes for the explanation of movement types that do not refer to internal cognitive properties of psychological agents as candidate frameworks for a psychological theory, for a psychological theory of brute movement would fail to discriminate between certain intentionally type-distinct tokens of behavior. What we may conclude is that neither a purely physiological theory of bodily movement--because it would conceal many interesting psychological phenomena--nor a theory of bodily movement that appeals to beliefs and desires--because it demands nomological relations where there are none to be found--provides a basis for an adequate psychological theory.

As indicated above, Field does not advocate the development of a theory of bodily movement. He merely suggests that the possibility of such a theory should be taken seriously since it seems to offer an alternative to a psychological theory that makes systematic use of semantic concepts. What we have argued is that this putative alternative cannot be sustained.

The Ineliminability of Semantics

Is a semantic theory an ineliminable component of a completed psychological theory? It is often acknowledged that one's answer to this question depends upon what one wants a psychological theory for. But suppose that we desire to construct a psychological theory that can distinguish functionally distinct cognitive accomplishments, e.g. a theory that marks the difference between giving an answer to an arithmetic problem and reporting the time of day. In that case, we will require a theory that sometimes assigns tokens of physically similar bodily movements, e.g. those that realize the utterance, "It's five", on different occasions, to different equivalence classes of behavioral output. Now I take it that no one denies the desirability of a theory capable of distinguishing different cognitive accomplishments and, hence, it follows that an adequate psychology must contain a principled basis for taxonomizing behavior that does not rely merely upon a metric of physical similarity defined over outputs. To a first approximation, the right kind of taxonomy of behavior is one continuous with a folk-psychological theory that type-identifies only content-equivalent behavioral tokens. For example, the utterance "It's five" conveys a certain content when it is intended as a response to the question "What is 2 plus 3?" and a distinct content when it is intended as an answer to the question "What is the time?" A taxonomy of behavior that distinguishes these tokens on the basis of their distinct intended interpretations constitutes what we have called an intentional taxonomy. Ideally, a semantic theory employed in the process of taxonomizing behavior can be thought of as a mechanism for deciphering the intended interpretations under which tokens of behavior are issued. Given that we require an intentional taxonomy of behavior, it appears that a completed psychology will include a semantic theory.

A similar view has been promulgated by contemporary proponents of the computational theory of mind such as J. Fodor and Z. Pylyshyn.[12] Both assert that the systematicity of behavior is captured only under intentional description. However, the role of a semantic theory in computational psychology has been obscure to many. To appreciate why this is so, it is only necessary to consider certain fundamental aspects of the theory. According to the computational theory of mind, internal mental processes are to be construed as sequences of computations. Computation can be viewed as the transformation of symbolic expressions by procedures sensitive only to the form or the syntax of the symbolic expressions upon which they operate. Conceiving of mental processes as computational processes means that mental processes are purely formal in scope, i.e. they do not have access to the semantic properties of the symbols over which they are defined. On this view, the semantic properties of mental representations are relevant only insofar as they are *mirrored* in the syntax or form of representations. The computational theory of mind, then, appears to declare that all internal psychological processes are in principle formally specifiable.

Given this understanding of some of the central tenets of the computational theory of mind, what theoretical role is filled by a semantic theory that could not be filled by a purely formal or non-semantic theory? As we will see, some philosophers have suggested that there is no such role for a semantic theory. If the semantic properties of internal representations are uniformly mirrored by syntactic properties, why should anything be lost when formal conditions for the type-identity of outputs are given? While semantic considerations appear to be essential for taxonomizing behavior in such a way as to make it amenable to subsumption under explanatory generalizations, perhaps we might develop a formal etiology of behavior and define the type-identity conditions for tokens of behavior over their formal etiologies. The possibility of a purely formal psychology requires careful consideration, but to anticipate the conclusion of this investigation, a psychological theory that makes no use of a semantic theory will fail to correctly specify the conditions for the membership of output tokens in paradigmatic cognitive types: There is no purely formal way to recover the taxonomy of behavior that respects the intended interpretations of tokens of behavior.

If we acknowledge with Fodor and Pylyshyn that what we have termed an intentional taxonomy provides a foundation for capturing the systematicity of behavior, then a semantic theory is an eliminable component of psychological theory only if the taxonomy given by reference to semantic considerations can be recovered from non-semantic considerations. Now an intentional taxonomy assigns tokens of behavior to content types. For example, a subject that intends to convey an answer to an addition problem, (PLUS 2 3), by uttering the expression "five" issues a behavior of a particular content type $\underline{P}$, where $\underline{P}$ = solution to arithmetic problem (PLUS 2 3). Eliminability of semantic theory turns upon the correct answer to the following question: Can formal conditions for membership of a token in content type $\underline{P}$ be specified? In general, can formal properties be specified for the individuation of each behavioral content type, $\underline{P}$, $\underline{P}_1$, $\underline{P}_2$..., without reference to semantic considerations? In the attempt to construct a formal taxonomy we may allow some revision of the intentional taxonomy and, hence, we need not require complete agreement between the two. But, a formal theory that cannot construe a token behavior as an answer to an addition problem is not an adequate substitute for an intentional theory that can so construe tokens of behavior. Throughout the following discussion, the simple arithmetic example will be used as a test case for the possibility of a purely formal psychology since it is, presumably, uncontroversial that individuals sometimes solve arithmetic problems.

The case for a purely formal psychological theory has not been made out in great detail, but S. Stich sketches one form that an argument for a formal theory might take. Stich asks us to consider the behavior of a young child who has just begun to learn the rules of addition, but cannot yet reliably determine sums greater than six or seven. Stich envisions a formal theory of the child's addition problem-solving behavior in the following terms:

> The theory would postulate various symbolic structures and rules for manipulating these symbols in a variety of ways. The theory would also detail the functional architecture that subserves these symbol manipulations. If all went well, the theory would be able to predict (the child's) answers to various arithmetic questions. [Stich, 1980, p. 152]

But Stich does not believe that a theory able to predict answers to arithmetic questions would need to construct interpretations of internal operations or interpretations of output:

> It makes no difference what we say the symbolic structures represent. No answer we give has the least bearing on the success or failure of the theory in predicting and explaining arithmetic problem-solving behavior....The claim that the formal symbolic structures that play a role in computational theory also represent something, be it a number, a gnu, or the proposition that Fort Lamy is in Chad, has no explanatory function within the theory. On the computational view of mind, talk of representation is simply excess baggage. [Stich, 1980, p. 152]

Stich claims that explanation and prediction of a cognitive accomplishment like arithmetic problem-solving is independent of all semantic considerations. This view presupposes that tokens of the behavioral type "solutions to arithmetic problem (PLUS 2 3)", for which explanatory generalizations are sought, can be picked out without reference to their intended interpretations or any other semantic features, i.e., that behavior can be taxonomized in a manner free from interpretation and yet in such a way that token solutions to arithmetic problems are identifiable. Failure to pick out all and only the tokens of behavior that represent solutions to arithmetic problems is equivalent to the failure to subsume the correct tokens of behavior under generalizations over problem-solving in arithmetic.

Stich does not supply an argument to illustrate the possibility of a formal taxonomy that picks out the right equivalence classes of behavior, but merely asserts the irrelevance of semantic interpretation. As we have observed, the role of semantic interpretation is, among other things, to provide a basis for the identification of behavioral tokens and for their assignment to the relevant types. The short answer to Stich is that since we cannot distinguish tokens that convey solutions to arithmetic problems, for example, from other kinds of behavior unless we provide interpretations for tokens of behavior and/or tokens of input, we require such interpretations if we are to state the generalizations relevant to arithmetic problem-solving. We cannot explain or predict what we cannot correctly identify and where behavior is concerned, successful identification requires interpretation.

The obvious response from someone advocating the possibility of a purely formal psychology will be to maintain that behavioral outputs can, in fact, be correctly taxonomized by employment of some purely formal criterion. The suggestion is that the behavioral type "solutions to arithmetic probl~~em~~ (PLUS 2 3)", and every other output type, can be formally individuated. In order to argue for this view the advocates of a formal taxonomy will no doubt rely upon certain principles of the computational theory of mind: viz. the principle that internal operations are formal in scopey the principle that all semantically relevant properties are mirrored by syntactic properties, and the hypothesis that there exists a type-to-type relation between the syntactic structure of internal representations and their semantic contents. Assuming that the theorist may be allowed recourse to these principles, it will be argued that a purely formal psychology is incapable of explaining and predicting paradigmatic cognitive accomplishments such as problem-solving in addition.

In order to isolate the issues at stake here it will be helpful to distinguish two positions, each consistent with certain principles of computational psychology, that might be taken toward the place of semantic considerations in psychological theory. One view reflects a kind of "reductionist" stance while the other represents a kind of "eliminativism". The reductionist view which we have in mind embraces the principle that all relevant semantic contents are mirrored by the formal features of representations, and holds that systematic correlations between the semantic properties and the formal properties of representations are, in principle, specifiable. The eliminativist view holds that the wholesale methodological rejection of semantic considerations is possible without adverse consequences for the generality and predictive power of psychological theory and, thus, represents what we might call the "formalist" approach. The formalist-eliminativist holds, with Stich, that reference to semantic properties is not required even by a theory capable of recovering generalizations originally stated over semantically characterized types of behavior.

Let us examine each view more closely. Given that one important task of psychology is to provide explanations and predictions of behavior, a comprehensive psychological theory will include law-like generalizations that specify the contingencies of behavior. The simplest possible skeleton for generalizations capable of being employed in predictions of behavior is the following:

$$C_1x \rightarrow B_1y$$

The formula can be read roughly as "every occurrence of x's being $\underline{C}_1$ produces or brings about an occurrence of y's being $\underline{B}_1$," where '$\underline{C}_1$' is a description of certain conditions upon which behavior of a particular type is contingent, and '$\underline{B}_1$' is a description of a particular type of behavior. (There must be some such psychological laws, but not every psychological law need specify the contingencies of behavior, e.g. certain interesting generalizations might specify the contingencies of belief revision conditional upon acceptance of new information.) Whether one takes behavior to be contingent upon environmental state-types, histories of reinforcement, or internal content bearing states, the predicate on the right-hand side of the formula, $\underline{B}_1$, must have as its extension an equivalence class of behavior. The predicates that fill the appropriate slot on the right-hand side of the formula may be simple or complex as long as they unambiguously specify genuine behavioral types.

Now, one necessary condition for recasting a law-like generalization in a new idiom, without loss of explanatory power, is that the intersubstitution of predicates on each side of the generalization be restricted to coextensive predicates. Thus, it must be possible to formulate a predicate in the new idiom that specifies the same equivalence class of phenomena specified by the original explanadum--occurring on the right-hand side of the generalization. The reductionist and the eliminativist both assert that it is possible to formulate psychological generalizations, originally formulated in an intentional idiom, in a non-intentional or non-semantic idiom. The notion of a non-semantic idiom is, of course, terribly loose and inaccurate insofar as any idiom in which extension-fixing expressions can be formulated possesses semantic properties. The intended contrast here is between "semantic characterizations of behavioral types" and "non-semantic characterizations of behavioral types" which can be sharpened in the following way. Semantic characterizations of behavioral types are given by *interpretations* of behavior while non-semantic characterizations purport to circumscribe behavioral types by reference to something like structural properties, i.e. formal or syntactic properties of behavior and/or of the internal states upon which behavior is contingent. In these terms, both the reductionist and the eliminativist advocate the possibility of substituting formal descriptions of behavioral types for interpretive, semantically laden , descriptions of behavioral types. The two views differ, however, in their respective diagnoses of the resources required for the formulation of formal descriptions of equivalence classes of behavior.

On the reductionist position, in order to specify formal descriptions coextensive with semantic characterizations of behavioral types, one attempts to systematically correlate formal properties with semantic properties. Ideally, the reductionist would specify a mapping from formal or syntactic types into content types--such a mapping, the reductionist observes, is suggested by the principle that all relevant semantic contents of internal representations are mirrored in aspects of their formal structure. Suppose, for example, that the reductionist is able to specify a one-to-one mapping from syntactic or formal types into content types. We regard tokens of behavior as type-identical if they possess equivalent intended interpretations and, hence, given a one-to-one mapping from formal types into semantic content types, behavioral types would be specifiable by expressions in which predicates of content types do not occur. Another way to put the same point is as follows. A one-to-one mapping from syntactic types into content types is more commonly referred to as a "type-to-type" relation between syntactic types and content types. If such a type-to-type relation can be specified, then it is plausible to suppose that there will exist formal characterizations of semantically characterized behavioral types.

But, nothing asserted by this kind of reductionism is incompatible with the view that psychological theory must include a semantic component or with the view that taxonomies of behavior must respect the interpretations under which tokens of behavior are issued. For the problem of specifying a mapping from formal types into content types is a semantic problem: A successful specification of such a mapping would constitute nothing less than an interpretation of, a decoding of, or a semantics for, the internal language of thought. A decoding of some kind is assumed to be possible, in principle, by computational theory, but notice that such a decoding is not properly speaking a <u>reduction</u> of semantic content at all. Decodings don't explain that in which content consists, but merely specify which formal structures convey which contents.

Incidentally, a decoding will not specify universal and lawful relations between aspects of syntax and aspects of content: We have no right to suppose that the relation between the contents conveyed by a particular language, even the language of thought, and the syntactic forms of expressions formalizable in that language is universal since different languages may convey the same contents in different ways. The task of decoding a language is an empirical task to be carried out for each distinct cognitive system, e.g. machine, monkey, man, and Martian. Moreover, in the case of machines, at least, we can be assured that the relation between form and content is thoroughly arbitrary. Hence, the possibility of machine thought itself appears to guarantee that there can be no reduction of content to formal structures in the classical sense of reduction which requires that "bridge laws" be laws.

What we have referred to as the reductionist view does not present a challenge to the view that psychological theory must include a semantic component since the envisaged "reductions" are nothing but semantic interpretations of the language of thought. The eliminativist view, on the other hand, takes a different view of the problem of specifying equivalence classes of behavior. According to the eliminativist absolutely no reference to content types is required in order to specify behavioral types. Thus, the formalist-eliminativist will not attempt to map syntactic types into content types, but will attempt to show that resources sufficient for recovering psychological generalizations are provided by formal considerations, i.e. by reference to properties of the structure of internal states and the formal relationships between such states. On this view--hereafter referred to as the formalist position--one does not attempt to correlate aspects of form with aspects of content, but attempts to specify all interesting behavioral types by investigation of only the formal properties of mental processes. The formalist is aided, however, by certain assumptions about the relation between formal properties and semantic properties. Though the formalist will not implicate himself in the semantic interpretation of the language of thought, he might make use of the idea that all relevant semantic properties are mirrored by formal properties. Given that the relation of form to content is in some way systematic, the formalist holds out the hope that the formal individuation of internal processes will circumscribe processes which are of uniform semantic content and, thus, that one can specify processes which causally underlie the production of intentionally type-identical tokens of behavior without explicitly relying upon semantic considerations.

The formalist position does present a challenge to the view that psychological theory must include a semantic component. For, if all behavioral types are specifiable in a way free of semantic considerations, then there is, perhaps, no good reason to take seriously the idea that the internal states upon which behavior is contingent are semantically endowed. To make a case for the formalist thesis it must be shown that semantically characterizable behavioral types can be individuated by some purely formal criterion. This, it will be argued in what follows, cannot be shown.

Before proceeding to a critical evaluation of the formalist thesis, note that there is yet another view that might be taken toward the role of semantics in psychology: One might acknowledge that psychological generalizations are sometimes given for semantically characterizable types and, yet, refuse to acknowledge the need to recover generalizations over semantically characterized behavioral types. Such view does not challenge the thesis argued for here--that psychology must include semantic component--since it simply refuses to engage it. Assuming that there are generalizations to be had about such types of behavior as "arithmetic problem-solving", or sub-types of that class, the problem is to determine whether or not the behavioral type and its sub-types are specifiable without recourse to semantic considerations. This problem is simply not addressed if one merely asserts that such generalizations need not be captured.

### Formalist Taxonomies

The formalist thesis--that formal specification of the conditions for the membership of an output token in an intentional type is possible--is not unattractive. What makes the thesis initially attractive is that it appears to promise an alternative to a taxonomy of behavior that relies upon the putatively troubled notion of content. What we have referred to as the content conveyed by a token of behavior, or equivalently as the intended interpretation of a token of behavior, has been glossed as that property of a behavior by virtue of which it is assigned to a psychologically relevant equivalence class, but no systematic theory of content has been supplied. As it turns out, it is quite easy to recognize the content conveyed by an overt behavior in practice yet quite difficult to give a satisfactory theory of content. Though there are important issues here, one might note that the situation is analogous to that in linguistics: It is quite easy to recognize a grammatical sentence yet quite hard to give a satisfactory theory of grammaticality. Roughly, grammaticality in language _L_ is that property of expressions by virtue of which they are assigned to the class _sentence of L_. The ease with which we interpret the content conveyed by overt behavior, in many circumstances, is at least _prima facie_ reason for believing that tokens of behavior are generated under proprietary interpretations--similarly the ease with which we recognize grammatical sentences leads us to believe that native speakers must internalize a theory of grammaticality, i.e. a grammar.

As indicated above, the view that a semantic theory is a dispensible component of psychological theory may rely freely upon certain aspects of the computational theory of mind. Of particular significance in this regard is the view that all relevant semantic properties of internal representations are mirrored by syntactic properties, and the hypothesis that the nature of the mirroring relation is that of a type-to-type relation. The existence of such a type-to-type relation would require that a content of a given type is always conveyed by token formal structures of the same type and that contents of distinct types are always conveyed by token formal structures of distinct types. On the assumption that the type-identity of formal structures is determined by their formal indistinguishability, the necessary and sufficient condition for distinctness of content is just formal distinguishability.[13] Recall that a type-to-type relation between form and content can be thought of as a one-to-one mapping from formal types into content types.

Although the hypothesis that there exists a type-to-type relation between the form and the content of representations is taken by some to suggest the eliminability of references to content, notice that considerably weaker assumptions may be made. Suppose that for some content type P there were, say, two formally distinct encodings, F and F*, of content type P. In this case, the disjunction of F and F* would, given one other condition to be mentioned, specify the class of internal formal structures that convey content of type P.

No a priori considerations prevent us from supposing that there are many formally distinct representations of content type P. A theory that attempts to specify formal conditions for the assignment of an output to its relevant equivalence class will not be troubled by this circumstance if sets of synonymous but formally distinct internal sentences can somehow be specified by reference to their formal properties. In particular, the mirroring relation between the syntax of internal representations and their content must satisfy the condition (a) that the distinct encodings of any content type are effectively enumerable. If condition (a) does not hold it will not be possible to formally individuate internal representations in a manner compatible with content-sensitive individuation. Additionally, the mirroring relation must satisfy the condition (b) that formally identical expressions never encode different contents: The language in which internal processes are carried out must be non-ambiguous. If both conditions are satisfied it may, for all we know at this point, be possible to construct a taxonomy of internal processes that type-identifies representations on the basis of their formal properties alone and yet happens to type-identify only representations of equivalent content. We will call a relation between content and form in an internal language that satisfies (a) and (b) a uniform relation. A uniform relation can be thought of as a many-to-one mapping from formal types into content types. Notice that a type-to-type relation between the form and the content of representations is simply that special case of the joint satisfaction of conditions (a) and (b), or that uniform relation, in which the number of distinct encodings of any content type = 1 (one).

A type-to-type relation between form and content may fail to obtain without irrevocably jeopardizing the project undertaken by the formalist. The formalist's task is made more difficult, but not impossible, by the non-existence of a type-to-type relation. This fact has not been generally acknowledged. In an argument for the ineliminability of appeals to content, Fodor attempts to make his case by questioning the assumption that a type-to-type relation obtains. The initial supposition is that if such a relation does obtain no reference to the contents of internal states is required:

> It is conceivable that there should be some formal property (call it 'U') that mental representations have if they express the property of being a unicorn; and some (different) formal property (call it 'W') that mental representations have if they express the property of being a witch. So, then, instead of explaining the differences between Seymor's witch hunting and his unicorn hunting by reference to the difference between the contents of the causally implicated mental representations, we could explain it by reference to the difference between U and W. [Fodor, forthcoming]

But Fodor holds that this line of thought is unwarranted since,

> There is, however, really no reason at all to suppose that there are formal doppelgangers of each feature of the contents of mental representations that we need to advert to in our accounts of the intentional properties of behavior. Positing such "type-to-type" correspondences between formal and semantic properties of mental representations involves a much stronger assumption than that each causally efficacious difference in content must correspond to some formal difference or other. [Fodor, forthcoming]

Fodor correctly observes that a computational theory of mental processes interested in offering an account of the intentional properties of behavior need not assume type-to-type correspondences between formal and semantic properties. But, this falls short of showing that appeals to content are ineliminable: Suppose that the various distinct formal properties that determine a mental state to represent unicorns are enumerable, U, $U_1$, $U_2$..., and likewise for witches, W, $W_1$, $W_2$.... If the relevant formal properties unambiguously mirror the subject matter of a mental representation, then the difference between Seymor's unicorn hunting and his witch hunting might be explained, the formalist will argue, by appeal to the difference between the properties U, $U_1$, $U_2$... and the properties W, $W_1$, $W_2$...--in particular, by reference to the difference between some element of the first set and some element of the second set. The formalist view requires that the relation between aspects of formal structure and aspects of content is in some manner a uniform relation, not that it is an ideal type-to-type relation.

The uniformity requirement does, however, restrict the scope of a formal theory. Dennett has suggested that we are unlikely to find a type-to-type relation between the contents of mental states and their formal structures if we attempt to encompass the psychological processes of all species of cognitive systems within a single theory.[14] The intuition here is simply that different cognitive systems in all likelihood employ different representational schemes. We are in a position to demonstrate the correctness of Dennett's intuition and to go one step further to show that no uniform relation between content and formal properties can hold across all species of cognitive systems. Suppose that we successfully specify: (i) a possible internal syntax L, and (ii) for each content type expressible in L, the set of sentences that convey content of that type. It would then be a simple matter to construct a possible internal syntax L* that conveys content of a particular type P by a token sentence s, that conveys content of a distinct type, Q, in L. For example, we might construct, L*, by the intersubstitution of every occurrence of the terms 'green' and 'yellow' in L. The construction of L* merely proves the existence of a language in the class of all possible languages of thought that uses a sentence token formally indistinguishable from a sentence of L to convey a different content. Hence, the non-ambiguity condition will hold only within particular systems of internal languages rather than across all such systems. One moral that might be drawn from this observation is that a universal functional psychology interested in inter-species comparisons of propositional attitudes would make ineliminable reference to semantic properties since there can be no purely formal description coextensive with a content type P where formally indistinguishable expressions convey different contents--i.e. where the non-ambiguity condition fails. Since no description of formal properties can have within its extension all and only the tokens of a certain content type; comparing propositional attitudes across different species of cognitive systems will require reference to properties common to the members of equivalence classes of propositional attitudes. What such propositional attitudes have in common is not a set of formal properties but specifiable contents.

For our purposes, the failure of the non-ambiguity condition when we generalize over all possible languages of thought shows that the attempt to construct a formal taxonomy project of behavior must be carried out on a species-by-species basis.[15] Given such a restriction on the scope of the formal taxonomy project, the problem is to formulate formal descriptions or characterizations of equivalence classes of behavior that are coextensive with semantic characterizations of equivalence classes of behavior. As noted, the assumption that the relation between the contents of mental states and the formal structure of mental states is uniform guarantees that, for a particular species of cognitive system, there exists a many-to-one mapping from syntactic types into content types. Thus, there will exist a list that specifies all the possible forms of mental states of each particular content type. Now, if the uniformity assumption is unwarranted, then only descriptions that specify semantic contents will circumscribe the equivalence classes of internal states relevant to the determination of the type to which tokens of output belong. But it is highly unlikely that the uniformity condition will fail, if only because no theory of the formal structure of internal processes will be taken to be finalized until a uniform relation between semantic content and syntactic form is exposed. Nevertheless, the plausibility of the uniformity condition does not provide a solution for our basic problem: Can we individuate the right equivalence classes of behavior without reference to the content of the mental states causally underlying the production of tokens of behavior? Our problem, as formalists, is not just to assure ourselves that there are, in principle, predicates of syntactic structures which have the same extension as predicates of content types, but to circumscribe or individuate behavioral types by specifying predicates of syntactic structures. Just which *formal properties* must an internal process exemplify in order to qualify as a process that produces, for example, arithmetic problem-solving behavior? *A priori* constraints are insufficient for answering such a question, since the formal properties of any particular representation are, in principle, consistent with diverse interpretations, i.e. a given syntactic item can encode a variety of contents.

Formal specification of equivalence classes of internal representations is only one part of the task that confronts the attempt to construct a formal taxonomy of behavior. The formalist must somehow individuate output types by reference to the formal properties of internal processes, for a formal taxonomy of behavior must either (i) define type-identity conditions for behavior in terms of properties of overt behavior and properties of stimulus conditions or (ii) define type-identity conditions for behavior in terms of properties of the internal processes responsible for the production of behavior. The former alternative is ruled out by the observations in the first section of this chapter. This circumstance forces the formalist to turn to the second alternative. Is the equivalence of outputs a function of formally specifiable internal processes? How might a taxonomy of behavior of this kind be constructed?

The simplest non-semantic type-identity criterion for internal representations is formal indistinguishability. According to this criterion, representations are tokens of the same type if and only if they are syntactically identical. Generalizing over sequences of operations in which representations are implicated, one might impose a similarly motivated criterion for the type-identity of mental processes and hold that mental processes are type-identical if and only if they are, everywhere, formally indistinguishable. The formalist might claim that token outputs are type-identical if and only if they are the products of formally indistinguishable internal processes. Indeed, on the assumption that internal sentences are unambiguous, tokens of behavior produced by formally identical processes must necessarily belong to the same intentionally specified equivalence classes. The formal indistinguishability of internal processes is, then, a sufficient condition for the type-identity of outputs.

Yet, behavior cannot be accurately taxonomized by making it a necessary (and sufficient) condition for type-identity that tokens have formally indistinguishable etiologies. The failure of this approach is easily exposed. Many equivalence classes of behavior contain tokens that are realized by very different bodily movements, e.g. one might utter the term 'five' or inscribe the symbol '5' in answer to a given problem. On the highly plausible assumption that physically distinguishable bodily movements are determined by formally distinguishable commands to the motor effector groups, it will turn out that processes eventuating in output tokens of the same type are often formally distinguishable themselves. The production of very different bodily movements requires the occurrence of formally distinguishable operations at some point in the production of such bodily movements. Hence, contrary to the criterion that poses complete indistinguishability of internal processes as a necessary and sufficient condition for equivalence of outputs, formally type-distinct internal processes will often eventuate in equivalent outputs.

The formalist might respond to this circumstance by attempting to specify the set of formally distinguishable processes which, for each behavioral type, produce equivalent outputs. Given a selection of intentionally specified input-output pairs, the formalist might specify the set of internal processes that mediate those input-output relations. But this tactic simply presupposes an intentional taxonomy of behavior; it does not demonstrate that we can recover the correct taxonomy by examining the formal properties of psychological processes. In any case, there will be infinitely many distinguishable processes that eventuate in tokens of the same type, for it is a simple matter to arrange a code to convey content of some type P that calls for any sequence of physically possible bodily motions under the control of an agent.

A significant improvement in the formalist's position is possible. If a piece of rational behavior conveys content of type P, then it is plausible to suppose that the behavior is mediated by mental states whose content is commensurate with content of type P. Drastically oversimplifying, let us allow the formalist the luxury of assuming that the production of a token output, conveying content of type P, requires the formulation of a mental state that unambiguously expresses content of type P. Furthermore, if there is a type-to-type relation between the formal structure and the semantic content of representations, then one and only one internal sentence type S will express content of type P. Hence, the formalist might hold that the processes that eventuate in tokens of the content type P, are those that implement tokens of the sentence type S.

This position is an improvement over any we have examined insofar as it might succeed in specifying a necessary condition for the type-identity of outputs; nevertheless it fails to provide a sufficient condition for the type-identity of outputs. The simple but compelling reason is that tokens of sentence type S may be implemented in many processes that eventuate in output tokens that express contents of some type Q where Q is type-distinct from P: Just as the same premise can be employed in infinitely many arguments to semantically nonequivalent conclusions, an internal sentence that expresses content of type *P* can enter into infinitely many mental processes that produce type-distinct outputs. Quite simply, the internal occurrence of a sentence token that expresses content of some type *P* does not by itself determine an output to convey content of type *P*. Hence, the formulation of tokens of some formal sentence type *S* will not determine the type-identity of tokens of output.

We are now in a position to more sharply conceive the requirements of a purely formal taxonomy of behavior. The formalist hopes to identify properties of internal processes that determine the type to which each token of behavior belongs. The hypothesis that output-type is a function of internal process type is tenable enough to merit careful examination, but it is far from clear that the conditions that determine output type can be articulated in a purely formal manner. Now, it might appear that the formal individuation of a behavioral type such as arithmetic problem-solving is aided by the assumption that behavior of this type is contingent upon, among other things, the implementation of formal algorithms for the computation of arithmetic functions. Although there is an important observation here that we will examine in some detail, the fact that agents employ formal algorithms when solving arithmetic problems does not improve the formalist's chances of specifying output types.

Since we require that proprietary descriptions of behavior capture the cognitive accomplishments of the agents of behavior, we will not describe the behavior of a subject as arithmetic problem-solving if the subject has, for example, simply copied an answer from someone else. One naturally assumes that a behavior realized by bodily movements that inscribe '5' on a chalk board is not correctly described, intentionally, as a solution to an arithmetic problem unless the agent has determined in some principled manner the solution to what he takes to be an arithmetic problem. This consideration suggests that one necessary condition for the membership of a token in the equivalence class "solution to (PLUS 2 3)" is that an addition algorithm has been implemented for represented input (PLUS 2 3). Under conditions of normalcy for the system, internal sentences that are generated by addition algorithms are about addition problems and behavior that intends to convey what a sentence generated by an addition algorithm expresses is behavior of the addition problem-solving kind. Thus, a core element in the specification of the conditions under which solutions to addition problems are issued is the identification of the addition algorithm or algorithms employed by subjects. Since algorithms can be viewed as mappings between formal symbols, it might appear that the formalist is in a position to identify instances of arithmetic problem-solving behavior. This appearance is largely dispelled, however, once we recognize that the formalist position requires that it is in principle possible to determine which sequences of internal operations are arithmetic problem-solving operations, e.g. which operations constitute addition algorithms, without reference to interpretations of input or output. That is, the formalist may not assume that a token of behavior is an answer to an arithmetic problem and proceed to specify the internal operations that produce the token since this tactic merely presupposes an intentional taxonomy of output.

A subject answering arithmetic problems is confronted with a series of tasks. These include, roughly, such things as (1) the assignment of proximal stimuli to the input class "arithmetic problem", i.e. the subject interprets something as an arithmetic problem, (2) the implementation of some algorithm to determine an answer to the problem, and (3) the initiation and control of bodily movements that the subject believes will convey the answer arrived at. Each of the three tasks represent a necessary condition for the production of a behavior intentionally characterized as an "answer to an arithmetic problem". Hence, the formalist must specify formal properties whose instantiation constitutes the satisfaction of each necessary condition, i.e., he must be able to tell when the necessary conditions for the production of behavior of some type P are satisfied. The question before us, then, is the following. Can the formalist determine, by reference to formal properties alone, when the necessary conditions for the membership of an output token in a particular behavioral type are satisfied? For example, can one determine by reference to the formal properties of internal operations when an algorithm for addition has been implemented?

Suppose that internalized algorithms for the computation of the addition function could be identified by inspection of the formal properties of internal processes. In that case, the formalist could specify internal occurrences that satisfy one necessary condition for the production of addition problem-solving behavior. On the other hand, if addition algorithms are not identifiable by inspection of their formal properties, then no formal theory will have the capacity to determine when the necessary conditions for the production of addition problem-solving behavior are satisfied, since such a theory cannot tell when an addition algorithm has been implemented.[3] A formal theory that cannot determine when the necessary conditions for the membership of a token in an output type are satisfied will not be able to specify the type to which the token output belongs. This is precisely the situation that the formalist finds himself in when attempting to specify the internal conditions upon which behavior of the addition problem-solving type is nomologically contingent: Reference to the formal properties of a symbolic process does not enable one to determine what privileged function the process computes, i.e. to give the process its <u>intended</u> abstract mathematical interpretation. Another way to make the same point is to observe that the formal properties of an algorithm cannot determine the computation of some unique function since many diverse interpretations are always completely consistent with all the formal properties of an algorithm.

Consider a Turing machine algorithm for example. An initial tape configuration might consist of '|'s and '□'s in some order such as the following:

(i)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
|  | \| | \| | \| |  | \| | \| | \| | \| |  |

An algorithm is determined by the set of quadruples <u>M</u>.

| | | | | |
|---|---|---|---|---|
| (M) | $q_o$ | \| | □ | $q_o$ |
| | $q_o$ | □ | R | $q_1$ |
| | $q_1$ | \| | R | $q_1$ |
| | $q_1$ | □ | \| | $q_2$ |
| | $q_2$ | \| | R | $q_2$ |
| | $q_2$ | □ | L | $q_3$ |
| | $q_3$ | \| | □ | $q_4$ |

The machine table specifies a sequence of operations that will take tape configuration (i) to tape configuration (h)

(h)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | \| | \| | \| | \| | \| | \| | | |

In general, with the reading hand initialized on the leftmost occurrence of '|' the implementation of M will take any two finite sequences of '|'s, of length n and m respectively, separated by a single '□' to a finite sequence of '|'s unseparated by any '□'. The resulting sequence of '|'s will have length = (n+m)-1.

The formal properties of the Turing machine algorithm are completely described by M together with the relevant conventions on the construction of Turing machine algorithms. But these properties do not uniquely determine "the function M computes", i.e., they do not determine an interpretation for M. In order to determine what privileged function is computed we must specify a semantics for the code of the tape. As Pylyshyn has observed, "if we wish to explain the computation that the device is carrying out, or the regularities exhibited by some particular programmed computer, we must refer to objects in a domain that is the intended interpretation or the subject matter of the computations" [1980, p. 113]. This is correct, it seems to me, if for no other reason than the fact that the formal properties of a symbolic process are always consistent with many different interpretations. For example, one interpretation that could be given for *M* is that it computes '(n+m)-1' for any two natural numbers *n* and *m*. This interpretation is based on the construction of a semantics for the tape symbols '□' and '|' that specifies their referents in the set of natural numbers as follows:

Tape symbol '□' refers to 0,<br>
'|' refers to 1,<br>
'||' refers to 2....

Other assessments of the relationship between the notation of the tape and the natural numbers, or some other domain entirely, will force other interpretations of the function that *M* computes.

Tape symbol '|' refers to 0,<br>
'||' refers to 1,<br>
'|||' refers to 2....

On this interpretation, '□' is simply a divider with no semantic properties and the tape contains numerals for the natural numbers in a more conventional unary notation. On this interpretation *M* computes the *addition* function.

The point of this illustration is simply that the formal properties of an algorithm, or any internal formal process, are always consistent with an infinite variety of interpretations. Often different interpretations of the function computed by an algorithm will be exceedingly natural: Under the first interpretation, M determines that the predecessor of the sum of 3 and 4 is 6, while under the latter interpretation, M determines that the sum of 2 and 3 is 5. Hence, if we are to maintain that some procedure internalized by cognitive agents constitutes <u>an algorithm for addition</u> we must have recourse to a semantic theory since the totality of formal properties of the procedure does not determine the computation of some unique function.

## Rejection of the Formalist Alternative

The lesson to be drawn here is that the formalist cannot merely inspect formal properties when attempting to determine if the necessary conditions for the production of an answer to an addition problem have been satisfied. For one necessary condition for such output is the implementation of an addition algorithm and formal properties themselves do not uniquely determine a procedure to constitute an algorithm for addition. In order to know that an internal procedure is an addition algorithm one must determine the intended interpretation of the procedure. That is, one must justify the application of a semantic theory that specifies an interpretation.

Is it possible that the formalist might avoid the conclusion drawn here while acknowledging that the formal properties of an internalized algorithm do not determine the privileged function that the algorithm computes? If the implementation of distinct algorithms is a necessary and sufficient condition for the distinctness of outputs, then a theory capable of distinguishing internal algorithms, whatever their intended interpretations might be, will provide a basis for a taxonomy of behavior. The formalist might claim, accordingly, that behavior can be accurately taxonomized without recourse to the intended interpretations of internalized algorithms, i.e. without the specification of the function that each algorithm computes. One might argue, for example, that it is not necessary to identify some internal process $\underline{A}_1$ as an *addition algorithm*, but merely necessary to distinguish algorithm $\underline{A}_1$ from other internal processes. Behavior would then be taxonomized by reference to distinguishable internal algorithms, e.g. the class of behavior "outputs of $\underline{A}_1$" forms an equivalence class distinct from all equivalence classes constituted by outputs of distinct algorithms.

The force of this response depends crucially upon an adequate criterion for the distinctness of internal processes or algorithms. But there are, in fact, only two ways in which algorithms can be distinguished: (i) by reference to their formal properties and, (ii) by reference to the function(s) that they compute. Quite clearly, the formalist cannot rely merely upon the formal distinguishability of algorithms, since formally distinct algorithms may compute the same function. Thus, the fact that two algorithms $\underline{A}_1$ and $\underline{A}_2$ are formally type-distinct is not a sufficient condition for the nonequivalence of their output. A subject solving addition problems, for example, may call upon formally distinct but functionally equivalent algorithms in the course of his computation of various sums.

On the other hand, if a subject calls upon algorithms that compute different privileged functions, then the output generated will belong to different equivalence classes. But how are we to determine whether or not formally distinct processes compute different functions? The only way that such a determination can be made requires recourse to semantic interpretations of each algorithm for, as we have repeatedly observed, the formal properties of an algorithm are always consistent with an infinite variety of interpretations, i.e. they do not determine the privileged function that each computes. Since the formal distinguishability of algorithms is not sufficient for the nonequivalence of their outputs, and functional distinguishability of algorithms presupposes semantic interpretation, there exists no nonsemantic concept of "distinctness of algorithms" to which the formalist may appeal in the attempt to avoid reference to the intended interpretations of internal processes.

It is important to be clear about what the present view does and does not assert. We allow that there are formal algorithms that subjects use when answering addition problems and that, in principle, the formal properties of an algorithm can always be specified--just as a Turing machine algorithm is formally specified by M. What we have claimed is that the mere formal description of an internal process will leave unanswered questions that must be answered if we are to correctly taxonomize behavior. Analogously, if we are interested in assigning the output of M to a relevant equivalence class we must decide upon an interpretation for the symbolic expressions upon which M operates: Does the output of M belong to the equivalence class of "answers to addition problems" or to the equivalence class of "answers to predecessor-of-sum problems"? The answer depends, one supposes, upon what M is used for. We can use M for either purpose, and when M is used for one purpose rather than the other its output belongs to one equivalence class rather than another. The same consideration applies even more naturally to the internal processes of cognitive agents: The output types of such a process depend upon the process' intended interpretation. This intuition is consistent with the principled point made here: One cannot specify the equivalence classes to which inputs and outputs belong merely by reference to the formal properties of the processes that mediate input-output relations, for the totality of formal properties is always consistent with the assignment of inputs and outputs to many distinct equivalence classes. It follows that a psychological theory that formulates generalizations over behavioral types such as arithmetic problem-solving cannot be a purely formal theory--any more than an explanation of a Turing machine's capacity to solve addition problems can be purely formal.

This conclusion may seem puzzling to some. After all, some philosophers have thought that the diversity of possible interpretations for any possible formal operation suggests the irrelevance of semantic interpretation. Admittedly, we have employed the same observation in such a way as to, if you will, turn that argument on its head. Semantic interpretation is an ineliminable component of psychological theory because behavior can be correctly taxonomized only under such interpretations--we have assumed that it is correct to take all tokens of behavior that convey correct answers to particular addition problems to form an equivalence class of behavior. Far from suggesting the irrelevance of interpretation to psychological theorizing, the multiplicity of possible interpretations makes the interpretation of internal processes and the interpretation of inputs and outputs essential. For in lieu of such interpretations the repertoire of functions computed by cognitive agents is unspecifiable. But behavior is, in the main, issued under an intended interpretation. Hence, the internal processes that produce behavior must possess intended interpretations. Justifying the assignment of interpretations of internal process and to behavior, then, is a task crucial to the success of psychological theory.

Notes

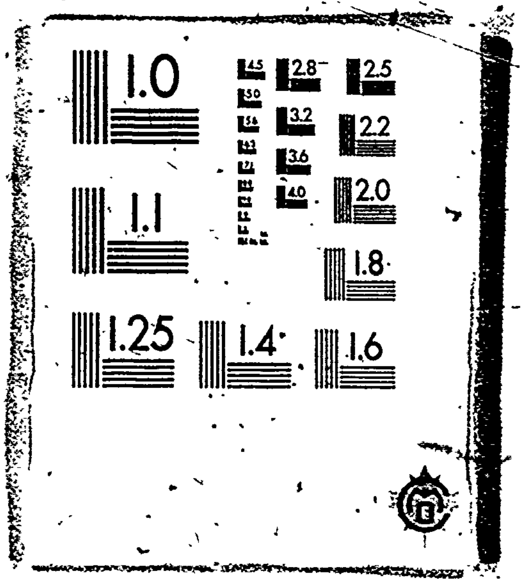
1. Later on we will allow the lifting of this restriction and consider the possibility of formulating complex predicates, defined over covert processes, that specify behavioral types. At the present time it is important to see that behavioral types cannot be individuated in a psychologically theory-relevant manner under the restriction imposed. That is, reference to the overt physical form of tokens of behavior is an insufficient basis for taxonomizing behavior. It is interesting that in actual folk-psychological practice we ordinarily have observational access only to the overt physical form of tokens of behavior and yet succeed in accurately taxonomizing behavior.

2. I have in mind here the so-called "bottom-up" theorists who suggest that psychology is to be constructed out of the kinds, and relations between kinds, individuated by such sciences as biology, evolutionary theory, and neurophysiology. This class of objections to intentional psychology is either easily dispensed with or impossible to engage: Insofar as a theorist admits that arithmetic problem-solving behavior, for example, is of interest, the bottom-up approach must provide a taxonomy that conforms at crucial points to the taxonomy given by appeal to the intended interpretations of behavior. Alternatively, a theorist might simply refuse to acknowledge that tokens of behavior that constitute the class "solutions to (PLUS 2, 3)" or one of its sub-types, merits explanation.

3. The relation between a code and the content that it is used to convey is completely arbitrary. Though tokens of behavior that intentionally convey a given content P are equivalent, a token behavior conveying content of type P may, in principle, take any physical form capable of instantation by motions of the human body under the control of an agent. Thus, assigning a token to its relevant type is a matter of properly decoding--interpreting--the token.

4. J. Haugeland [1978] requires that the inputs and outputs of an intentional system, what he calls "intentional black boxes", be taxonomized in a way independent of interpretation--he appears to assume that something like "syntactic" criteria are available for this purpose. This position seriously confuses the role of interpretation in psychological theory. For Haugeland, a device is construed as an intentional black box if and only if an "intentional interpretation that makes reasonable sense" can be given for an "articulated typology" of the device's inputs and outputs. An articulated typology, in turn, is a scheme for taxonomizing

# 2



input and output types. Haugeland would have us construct articulated typologies *ancillary* to any interpretation and this, I have argued, is not possible for human behavior--non-interpretive taxonomization is bound to misclassify certain tokens.

Notice also that Haugeland places unreasonably strong constraints on what can count as input or output tokens. He requires that "the influences of a device on its environment" can count as a token of behavior only if the syndrome of influences is never a token of more than one type. But, it is often the case that the physical events identified with tokens of behavior instantiate type-distinct tokens of behavior on different occasions of occurrence. A similar criticism of Haugeland--regarding the question of the non-ambiguity of uninterpreted tokens--has been urged by Dreyfus [1978]. The point that should be kept in mind is that tokens of input and output are only non-ambiguous *under interpretation*.

5. J. Fodor [1975] has argued persuasively that the *kind* predicates of physical theory cross-classify the phenomena (kinds) picked out by the terms of the special sciences, of which psychology is a paradigmatic example. This important observation is obviously related closely to the present point. But notice that we have not restricted the physicalist's resources for specifying behavioral types to the use of kind terms: Any disjunction of predicates of overt behavior, as long as the disjunction is finitary, is allowed. Now such disjunctions will not name individual physical kinds and will not, as we have seen, pick out behavioral types. Hence, our attack on the possibility of physicalistic individuation of behavioral types is not based on the failure of a type-to-type relation to hold between the kinds of physical theory and kinds of psychological theory. We have urged that interpretation is crucial for correct taxonomization of behavior since even a "heterogeneous and unsystematic disjunction of predicates" of overt behavior will not have the same extension as 'solutions to (PLUS 2 3)'. Not only is the behavioral type not coextensive with a physical *kind*, it is not coextensive with any disjunction of predicates of overt behavior.

6. Chisholm's attempt to specify the class of sentences about psychological phenomena was attacked, typically, by showing that there are sentences picked out by his criteria that are not "psychological". In particular, criteria for the non-extensionality of sentences pick out alethic modal sentences. In response, Chisholm [1967] retreated to the following criterion. A sentence is about a psychological phenomenon if (and only if) it is a substitution instance of *Mp* where '*M*' is an "intensional prefix" and *p* is a

closed sentential clause: (i) an expression is a prefix if and only if the result of appending closed sentences to it is a closed sentence (such terms are usually called modal terms), and (ii) a prefix is an intentional prefix if and only if the product of appending any closed sentence is a contingent sentence. According to these criteria, the expression 'Sam believes' is an intensional prefix, and

'Sam believes the sun is shining.'

is a sentence about a psychological state. But, 'John stabbed' is not an intensional prefix since it is not a prefix at all. Chisholm's criterion tends to pick out only sentences that report propositional attitudes. Chisholm has moderated his claim somewhat and suggests that the criterion is perhaps only a sufficient condition for psychologicalness --hence the parentheses in the criterion above.

7. The requirement for intentional construals of behavior does not mean that psychological properties of individuals can be specified only by sentences that are restricted to concepts employed in the production of behavior. In fact, there are psychological properties that cannot be referred to by descriptions that employ only those concepts employed in the mental states that produce certain tokens of behavior. For example, suppose that an individual happens to be extremely gullible when confronted with certain kinds of propaganda. An individual cannot issue a behavior under the intended interpretation "to gullibly accept what is said by putative authority figures". Gullibility might be feigned and an individual may even recognize from past experience his tendency to gullibility, but he cannot produce a particular behavior characterized by gullibility by intending to be duped on that occasion--one cannot know that he is being duped though on reflection he may realize that he has been duped. Importantly, however, in order to justify the attribution of gullibility to an individual we must have access to the intended interpretation under which he issues various tokens of behavior. We must know, for example, that he takes an instance on which he was easily duped to constitute an instance of "placing trust in the veracity of authority" or something roughly equivalent. For only when we know how the agent conceives his behavior can we characterize its nature against a broader background that may lead us to attribute a property like gullibility.

8. Field's line of argumentation for the inclusion of a semantic component in psychological theory is not as persuasive as we might like. Many theorists would hesitate to endorse the appeal to semantics if the only motivation for the appeal is a desire to reconstruct the "evidence" relation

between beliefs. Rather than characterize the evidence relation a theorist might propose instead to examine the statistics of belief proliferation: Does the probability that a subject will form a belief token of type P increase with the frequency with which the belief that P is expressed? We are not suggesting that such a study would deal with an analogue of the evidence relation but merely with a potentially non-semantic causal relation between beliefs which some theorists might hold to be a satisfactory alternative.

9. The functional individuation of beliefs and desires requires, of course, a specification of input and output types. If, as we have argued, a taxonomy of overt psychological phenomena that captures interesting equivalence classes of phenomena requires interpretation of inputs and outputs, then all "functional individuation" of internal psychological states is actually semantically based.

10. This interpretation of Field's motivation for entertaining such a theory of bodily motion was communicated to me by Field in discussion in the fall of 1981.

11. Notice that this means that the "stimulus meaning" of a term, a la Quine [1960], may fail to circumscribe a physical equivalence class of stimuli. An individual may assent to a term, for which there are dissent conditions defined over appropriate stimuli, under a diverse collection of stimuli. Further, stimuli included in the stimulus meaning of a term may be included in the stimulus meanings of non-synonymous terms.

12. One of Fodor's [forthcoming] clearest recent expressions of this view is put by saying that "we have no notion of behavioral systematicity at all except the one that makes behavior systematic under intentional description". Pylyshyn has repeatedly emphasized the role of intentional interpretation in computational psychology asserting, for example, that "if we wish to explain the regularities exhibited by some programmed computer, we must refer to objects in a domain that is the intended interpretation or the subject matter of the computation," [1980, p. 113; see also Pylyshyn's forthcoming book.]

13. Fodor [1980] is the originator of the "formal indistinguishability" criterion for the type-identity of mental states. It appears, however, that Fodor no longer considers the criterion plausible because of the very real possibility of "formally distinct but synonymous" representations, i.e. the possibility that a content P might have distinguishable syntactic encodings, see Fodor (forthcoming), and quotations in the main text.

14. Fodor reports Dennett's observation [forthcoming].

15. Under the restriction to specific species of cognitive systems we may still consider conditions for the "type-identity" of internal states and for behavioral output. But note that this is perhaps somewhat nonstandard: A central motivation for functionalism in the philosophy of mind is to provide a level of abstraction at which one can speak of the "equivalence" of psychological functions that are realized by different systems, e.g., man, machine, and Martian. For functionalism the relevant type-identities are not species-specific. What we have observed is that if one is a functionalist and cleaves to an extreme version of formalism, then type-identities between mental states are not definable across all possible cognitive systems. An internal representation with a given syntactic or formal structure S will realize different propositional attitudes in certain pairs of systems. Hence, the formalist-functionalist must adopt a strategy in which the first step is to specify type-identities relative to particular systems and only then to compare mental states across systems. It is unclear, however, whether or not this second step might be carried out under the constraints associated with formalism.

# CHAPTER II

# INTENTIONAL EXPLANATION AND THE RATIONALITY ASSUMPTION

## TWO CHARACTERIZATIONS OF INTENTIONAL EXPLANATION

Under the terminology adopted here an intentional theory of behavior is simply a theory that makes essential use of the idea that at least some of the internal states upon which behavior is contingent are endowed with contents. According to the present view, psychological theory will be or will contain an intentional theory since capturing certain behavioral types for which we wish to formulate explanatory generalizations requires an intentional taxonomy of behavior. If we may refer to certain contentful internal states as beliefs and desires, then intentional explanations of behavior are often instances of the following schema

(I) If an agent desires _G_ and believes that _M_ is a prudential means for securing _G_, then the agent will do _M_.

Explanation and prediction of behavior is accomplished through reference to the beliefs and desires out of which agents act. Such explanations and predictions are considered "intentional" since they advert to the contents of mental states. But, generalizations cast in this intentional format are to be thought of as empirical claims concerning causal relations between mental states-and behavior. This is a standard interpretation and has been strongly urged by Fodor [1981, pp. 100-123]. There are several aspects of intentional explanation as construed here that deserve comment. First, if tokens of behavior are to be predicted by reference to beliefs and desires, the theorist must have access to some technique for the attribution of beliefs and desires. Ideally, belief attribution might be founded upon an empirical theory of the belief formation process itself: Under what conditions are beliefs of particular content types likely to occur? This question conceals a considerable problem. Fodor has suggested that the problem of providing a theory of belief formation is essentially a problem of inductive logic: How are we to systematically constrain the information relevant to the confirmation of a hypothesis? If Fodor is essentially correct, experience in the philosophy of science shows that this is a question that we will remain unable to answer [forthcoming]. It is important to recognize, however, that the difficulties confronting a theory of belief formation do not vitiate intentional explanation. For though we may have no systematic theory of belief formation, there are a variety of methods that the theorist can employ to determine what beliefs an agent internalizes. These range from the interrogation of an agent to placing an agent, or oneself, in a specified set of circumstances and noting what beliefs occur to the agent. The idea behind the latter technique is to embed actual belief systems in various sets of circumstances and simply let them operate, i.e. generate whatever beliefs they will in those various circumstances. There is, no doubt, considerable latitude in belief formation procedures and this may result in the generation of different beliefs in similar sets of circumstances. But this technique for correlating beliefs with the circumstances in which they are likely to be formed holds promise for determining the belief formation capacities of various systems. For example, one might give subjects a problem to solve in which the resolution of the problem depends upon the acquisition of some, otherwise unavailable, item of information made available in a display presented to the subjects, and in this way determine if subjects are capable of forming a belief of a certain content type in a certain set of circumstances. The point illustrated by these considerations is simply that the theorist does not completely lack methods of belief attribution even if a systematic theory of belief formation is a utopian goal in psychology.

Notice, also, that according to the general format for intentional explanations of behavior, it is required that the actions an agent initiates are believed, by the agent, to be prudent in the circumstances in which he believes himself to be fixed. This qualification is merely one way of conveying several different constraints on predictions of behavior. For example, where the theorist predicts that an agent will do M, the agent must not only believe that method M is a way of obtaining goal G, he must believe that M is something that he could actually do. Further, the agent must believe that M will not viciously conflict with other of his goals of equal or higher priority. The value of these constraints is easily illustrated: We all believe that committing the perfect crime is a means for obtaining great wealth, but we don't all believe that we could commit such a crime and we don't all find the prospect of such a crime to be compatible with other personal aspirations.

To require that agents take their courses of action to be prudent is not to require that behavior accommodated to intentional explanation actually is prudent. Far from it, agents may take the wildest, most implausible schemes to be prudent while acting out of beliefs and desires. Agents may fail utterly to properly evaluate the impact of a certain method on other high priority goals and thus act imprudently. The prudentiality constraint is imposed simply to prevent the theorist from predicting that an agent will initiate every course of action that he believes to be a theoretically possible means for satisfying desires that he has internalized: Thus, though I may fully believe that "a successful two-man assault on Everest" is a means for achieving "international renown as a mountaineer", I'm not about to try it.

The prudence associated with a course of action is relative to the agent's own impressions and subjective judgement. What may seem prudent to one agent may seem foolhardy to the rest of us. Hence, by imposing this constraint on the beliefs out of which agents act, we have not required that agents have the capacity to determine the objective pragmatic value of their actions. We have not required, in particular, that all behavior subsumed by generalizations in the intentional format is rational. Now, it is time to observe that this construal of intentional explanation differs markedly from a widely known and influential view which does place a rationality constraint on the intentional explanation of behavior. The view of which I am speaking is, of course, the view that D.C. Dennett has constructed in a series of important papers.[1]

To better see what is at stake in the conflict between Dennett's view and that adopted here, recall that according to the present view intentional explanations of behavior are provided by empirical generalizations that cite content bearing states as essential elements of the circumstances upon which tokens of behavior are contingent. The appeal to the content of mental states has been founded upon our observations concerning the problem of taxonomizing behavior. Dennett offers a different evaluation of the basis for the appeal to the content of mental states. Initially, one might think that the more motivation that can be found for the appeal to the contents of mental states, the more secure the concept of intentional explanation becomes. But, on one interpretation, Dennett holds that the only foundation suitable for intentional explanation is, in reality, a foundation of sand: Dennett presents a characterization of intentional explanation that weds its credentials to the satisfaction of a *rationality condition* by behavioral systems. Roughly, for Dennett [circa 1970-1978] intentional explanation is founded upon the assumption that the agents whose behavior is to be explained are optimally rational. But, no actual agent is optimally rational and, hence, intentional explanation makes for vacuous psychology or, at best, a useful heuristic in the explanation of behavior.[2]

The interpretation of Dennett's view of intentional explanation is aided, ironically, by the fact that Dennett claims to have both (a) the right construal or analysis of the concept of intentional explanation, and (b) strong arguments against the empirical relevance of intentional explanation. An aspect of Dennett's motivation for abjuring intentional explanation, contrasted with his arguments against this form of explanation, is constituted by his assumption that there are, in principle, mechanistic or physicalistic approaches to the explanation of behavior and cognition of a fully adequate sort. But Dennett does not rail against intentional explanation simply because he believes that there are alternatives. Instead, Dennett constructs an argument against intentional explanation predicated upon its construal as an "inescapably normative" construct [Dennett, 1978, p. 285]. Intentional explanation, for Dennett, makes a normative, non-descriptive, assumption about the rationality of the agents of behavior. And, the falsity of the assumption upon which intentional interpretations are logically founded entails the empirical vacuity of such interpretations. Now, an argument like this positively demands a construal of "rationality" that sets a normative standard that people are unable to meet. Dennett's argument against intentional explanation, thus, provides evidence for the interpretation of the "rationality assumption" that he claims founds intentional interpretation.

Dennett's skepticism about intentional explanation extends to the concepts of belief and desire. Moreover, Dennett views these concepts, as he does that of intentional explanation, as "inescapably normative". Consider the following passage:

> Human beings or other entities can only aspire to being approximations of the ideal, and there can be no way to set a 'passing grade' that is not arbitrary....There is no objectively sufficient condition for an entity's really having beliefs, and as we uncover apparent irrationality under an intentional interpretation of an entity, our ground for ascribing any beliefs at all wanes, especially when we have (what we always can have in principle) a non-intentional, mechanistic account of the entity. [Dennett, 1978, p. 285]

As we can see from this reflection, idealization per se is not what vitiates intentional explanation. Empirical science is comfortable with idealization: We talk about physical forces under idealizations that serve as simplifying assumptions, e.g. "Suppose there were a frictionless plane, then...," but the concept of such forces has application even under conditions that are not ideal and simple. For Dennett, talk of beliefs and desires is coherent only under the idealization associated with the rationality assumption. Thus, on Dennett's view, the application of these, and associated concepts to actual, empirical systems is at best metaphorical or heuristic. Intentional concepts are never descriptive of actual systems; they represent normative standards which actual systems fail to satisfy.

On this interpretation, Dennett's view is an extreme one, but this interpretation is actually quite common in recent works by Dennett's critics and colleagues. Fodor, for example, claims that, for Dennett, intentional explanation logically presupposes rationality and is, thus, vitiated by the failure of people to satisfy the rationality condition:

> [Dennett's] point is not just that intentional theories rest on an assumption of rationality; it's that they rest on a counterfactual assumption of rationality. And, of course, a theory which entails (or presupposes) a falsehood can't be better than heuristic. [Fodor, 1981, p. 116]

Others share Fodor's assessment of the threat to intentional psychology, realistically construed, posed by placing a rationality condition at the foundation of intentional explanation. C. Cherniak, for example, offers a roughly equivalent interpretation of the situation:

> If the only possible rationality conditions on, e.g., beliefs are so idealized as to be inapplicable to humans, then any attributions of beliefs to humans cannot really be true; the attributed entities are at most useful myths. [Cherniak, 1981, p. 163]

Both Fodor and Cherniak resist the conclusions to which such considerations appear to lead. Fodor argues, roughly, that intentional theories can aspire to truth -- rather than merely to heuristic value -- on an interpretation of intentional generalizations as models of causal relations between states individuated by content. Cherniak suggests that the assumption of ideal or optimal rationality, putatively foundational to intentional explanation, can be attenuated. That is, that the ideal rationality condition can be exchanged for a "minimal rationality" condition which is literally applicable to human agents. Both strategies have merit, but both seem calculated to win support for a non-instrumentalistic interpretation of the posits of intentional explanation. Neither strategy engages what appears to be Dennett's central intuition as directly as one might like: "rationality is the mother of intention" [1978, p. 19].

In response to views like those of Fodor and Cherniak, Dennett may, and indeed does, continue to hold that the domain in which intentional explanation has proprietary application is fixed by the rationality condition. This means, among other things, that only those psychological processes which are rational, either fully rational or rational within some specified lower bound, are to be thought of as involving content bearing states. Thus, on this view, people believe, desire, strive, wish, worship, plot, conceive, and contemplate only insofar as they satisfy certain criteria for rationality. Now, Dennett has recently observed that there may exist rational processes in human agents [1979]. This observation, which may constitute something of a concession to Fodor's realism, sets the stage for acknowledging that beliefs and desires sometimes enter into the etiology of behavior. But Dennett remains adamant in his conviction that there is no justification for treating beliefs and desires as actual states in the etiology of behavior where the rationality condition fails to truly characterize internal processes.

Dennett's commitment to this view can be traced, it appears, to an underlying conception of intentional explanation as reason-giving explanation. That is, we often give an informal account of an individual's behavior by first attributing certain goals to the individual and then showing how the action performed could be construed to be a likely means toward the fulfillment of the attributed goal. Reason-giving explanations typically, it is said, attempt to justify the tokens of behavior for which they are offered. Naturally, there can be no justification for tokens of irrationally conceived behavior and, thus, reason-giving explanation appears to be tied to a rationality condition. For Dennett, intentional explanation is a form of reason-giving explanation and, thus, intentional explanation is essentially explanation by justification.

Contrary to Dennett, according to the view advocated here the extent to which people are rational does not fix the extent to which they are intentional systems. On this view, propositional attitudes -- beliefs and desires -- may enter into the etiology of a token of behavior even though the process whereby the token is produced departs in one way or another from canons of rationality.

Rationality is not, (i) definitive of the level of intentional interpretation, (ii) a necessary condition for the attribution of propositional attitudes, or (iii) constitutive of intentionality. If this view is correct, then the rationality assumption does not provide the foundation for intentional explanation. A fundamental point to be secured in the consideration of these issues is that our lamentable but typically human tendency to fail to live up to our full rational potential does not degrade our essential cognitive, intentional nature.

We have been attending to certain aspects of the view that rationality is in some way criterial to intentional explanation. There are, however, other important elements in Dennett's view of intentional psychology. In particular, Dennett apparently holds that intentional explanation is eliminable in favor of supposedly "more objective" frameworks for the prediction and explanation of behavior. Dennett has sketched three strategies for the "prediction of behavior", at least one of which must be considered non-intentional, and maintains that each provides, in principle, for a satisfactory account of human behavior. At one point Dennett claims that the employment of intentional and non-intentional forms of explanation for different classes of phenomena reflects "no difference in kind" lying between the phenomena accommodated by each respective form of explanation. This constitutes an important problem-area in Dennett's treatment of intentional explanation which will be examined in detail in succeeding sections. The various strategies for the prediction of behavior, sketched by Dennett, will be reviewed and it will be argued that the applicability of each strategy to human behavior cannot be demonstrated.

The other major problem-area for Dennett's construal of intentional explanation concerns, of course, the relation between the rationality condition and intentional explanation. In characterizing intentional explanation as reason-giving explanation, Dennett would constrain intentional interpretations of psychological processes in such a way as to make all such processes appear rational. For, only in this way, presumably, can a psychological process constitute a reason for behavior. Now, there is an important element of truth in this observation, and in a host of others for which we have Dennett to thank, but such considerations fail to show that a rationality assumption is a necessary condition for intentional interpretation. In this connection it will be argued that (i) the arguments for the rationality condition are inconclusive and that (ii) on the most natural analysis, intentionality is a presupposition of the attribution of rationality -- and of the attribution of irrationality for that matter.

By way of a final introductory remark, let me acknowledge that no examination of Dennett's work of this size can do justice to the many insights he has provided and to the scope of his examination of issues in the foundations of cognitive psychology. But no examination ought to set such a goal since it can be fulfilled best and in a most enjoyable way by reading Dennett's works themselves. Instead of attempting to do justice to the scope of Dennett's views, we will concentrate on fundamental aspects of his characterization of the form and domain of intentional explanation.

## THE THREE PREDICTIVE STANCES

As noted, a central theme in Dennett's account of psychological explanation concerns the alleged availability of a variety of stances or strategies for the prediction of behavior. Dennett specifies three such strategies for the prediction of behavior: the _physical stance_, the _design stance_, and the _intentional stance_. Each stance-type is defined by the assumption or set of assumptions that it makes about the systems whose behavior is to be predicted. Adopting the physical stance, for example, amounts to assuming that the system whose behavior is to be predicted is a physical system amenable to physical law. In a recent work Dennett elucidates the physical stance in the following terms.

> If you want to predict the behavior of a system, determine its physical constitution (perhaps all the way down to the micro-physical level) and the physical nature of the impingements upon it, and use your knowledge of the laws of physics to predict the outcome for any input. This is the grand and impractical strategy of LaPlace for predicting the entire future of everything in the universe....The strategy is not always practically available, but that it will always work is a dogma of the physical sciences. [Dennett, 1979]

Physical theory is obviously appropriate for predicting a wide range of events involving a wide range of systems, including sub-atomic particles, gasses, galaxies, and others, but it will not be conceded, in the argument to follow, that the universality of physical theory, properly understood, entails the epistemic adequacy of physical theory for behavioral predictions. In any case, Dennett apparently adopts the view that the physical stance is, in-principle, applicable to human behavior. Hence, the epistemic adequacy of the physical stance is a dogma that Dennett finds congenial.[3] Our reluctance to employ the physical stance in the domain of human behavior is, for Dennett, merely a sign of our pragmatic good sense: The sheer quantitative complexity of physical stance calculations of states of human agents rules out the practical utility of this strategy for predictions of behavior.

The physical stance, lying at one extreme, contrasts most sharply with the intentional stance, lying at the other extreme. Dennett offers the following concise outline of the predictive strategy that constitutes the intentional stance.

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. [Dennett, 1979]

Employing the intentional stance as Dennett directs will "yield a decision about what the agent ought to do [and] that is what you predict the agent will do" [1979]. The actions that are predicted from the intentional stance are those that careful consideration reveals to be the most appropriate or rational of the actions possible. Thus, intentional stance predictions are founded upon the assumption that the systems whose behavior is to be predicted are optimally rational. In fact, to make this single assumption concerning the rationality of a system is, for Dennett, to view the system as an <u>intentional system</u>.

Dennett characterizes the concept of rationality apposite to intentional systems asserting that "rationality here means nothing more than optimal design relative to a goal or optimally weighted hierarchy of goals" [1978, p. 5]. This conception of rationality appears to survive Dennett's most recent explication of the intentional stance but does require augmentation to provide an understanding of the rationality of goals themselves, since intentional stance prediction involves attributing to a system "the desires that it ought to have."[4] Dennett typically appeals to evolutionary or biological considerations when faced with the question of the rational appropriateness of goals. On this view, evidently, a system "ought to have" only those goals that promote its survival or other biological needs. Assuming that the goals and beliefs that a system ought to have can be specified in particular cases, adopting the intentional stance consists in adopting a "means-ends" conception of rationality and using the simplifying tactic of assuming that the system is designed so as to effect the most appropriate, i.e. the optimal, means at its disposal for securing its attributed goals.

Though there are no principled restrictions on the range of application of the intentional stance, Dennett tends to avoid recommending the gratuitous application of the intentional stance to simple inanimate objects, e.g. chairs, stones, and thermostats, by letting its preferred application range over only those systems for which the other stances are impractical. The intentional stance is not adopted with respect to the motion of a falling body, for example, because there exist fully general physical laws that explain the falling body's acceleration; the intentional stance is not applied to complex systems like alarm clocks because there exist fully adequate specifications of the design of such systems,by reference to which one may account for their operation. But when a system is simply too complex to succumb to either physical stance or design stance prediction, such as a computer programmed to play chess or a human agent pursuing a complex goal, we revert to the intentional stance: Assume that the system is fully rational, attribute to the system those beliefs and desires that it ought to have, and predict as its forthcoming action the most rational action that it could perform, in the circumstances in which it is embedded. Dennett conceives of the intentional stance as a sort of "short cut" strategy for predicting occurrences that are simply too complex to be amenable to prediction from lower-level stances, given practical considerations such as the length of time available for calculation. Thus, for Dennett, use of the intentional stance reflects a kind of epistemic frailty, or a lack of cognitive resources, that typifies human understanding of complex systems and, hence, of humans themselves.

The design stance is intended to lie at a level of abstraction somewhere between the physical stance and the intentional stance. We mention the design stance last because the criteria with which Dennett marks this stance are less definitive than those with which he marks the other stances. Dennett relies heavily upon one's intuitions in his elucidation of the design stance and often appeals to the concept of "design" in his characterization of this stance. But, it is the concept of "design" itself that requires elucidation. In a typical characterization of the design stance Dennett says,

> Sometimes...it is more effective to switch from the physical stance to what I call the design stance, where one ignores the actual (possibly messy) details of the physical constitution of an object, and, on the assumption that it has a certain design, predicts that it will behave as it is designed to behave under various circumstances. [Dennett, 1979]

Such a characterization of the design stance is useful only insofar as one already possesses a clear concept of "design". Perhaps the most helpful thing that Dennett says in this regard is the following:

> Different varieties of design stance predictions can be discerned, but all of them are alike in relying upon the notion of function, which is purpose relative or teleological. That is, a design of a system breaks it up into larger or smaller

> functional parts, and design stance predictions are generated by assuming that each part will function properly. [Dennett, 1978, p. 4]

The concept of function plays a crucial role in the design stance since the design attributed to a system is relative to the functions that the system is taken to serve. For example, the design of an alarm clock specifies only those aspects of its structure that are relevant to the fulfillment of certain goals or purposes, e.g. the accurate representation of local time and the ability to sound an alarm at any predetermined time within some period -- typically twelve hours. The point is that the functions that one takes the system to serve constrain interpretations of the structure of the system: Attributing different sets of functions, goals, or purposes may result in different interpretations of a device's design. Note that Dennett insists that "not all diagrams or pictures are designs in this sense, for a diagram may carry no information about the functions -- intended or observed -- of the elements it depicts" [1978, p. 4]. The crucial point is that a design must specify both (i) the overall functions or purposes of a device and (ii) the various subsystems of the device that subserve those purposes and any subpurposes.

If the design stance is to be employed in predictions of behavior, we must first assign purposes to the agents of behavior. This does not mean merely that each token of behavior is to be viewed as purposeful, but that the whole system -- the person -- is to be viewed as endowed with a proprietary set of functions or purposes. Given an assignment of purposes to an individual, one would proceed to decompose the various subsystems that have a role in subserving those functions. Once a "design" is arrived at, it is used to predict the behavior of an individual much in the same way that the design of an alarm clock could be used to predict the clock's activity: One simply assumes that "each part will function properly" and that the system will behave "as it is designed to behave."

This concludes our brief sketch of Dennett's three strategies for predicting the behavior of a system. It remains for us to evaluate the claim that each strategy constitutes a genuine methodology for predicting, and ultimately explaining, human behavior.

## PROBLEMS CONFRONTING DESIGN STANCE PREDICTIONS

The design stance, as deployed by Dennett, is beset with difficulties that severely degrade its usefulness as a strategy for predicting behavior. This is particularly interesting since nothing that Dennett says about the design stance sharply distinguishes it from the intentional stance. On the intentional stance, for example, one makes considerable use of teleological concepts of function and purpose when formulating a basis for the attribution of beliefs and goals to agents, e.g. the goals that an agent "ought to have" are those that are compatible with ultimate and basic purposes assigned to the agent. Reference to these teleological concepts can hardly be the distinctive mark of design stance predictions since they comport freely with the intentional stance as Dennett describes it. Further, a theorist adopting the intentional stance might find it useful to carve up internal systems according to their various roles in determining the behavior of an agent just as Dennett advocates carving up the subsystems of a device when attempting to specify the design of the device. The only salient difference between these two stances appears to be the assumption of rationality imposed by the intentional stance.

The rationality assumption, however, fails to draw a sharp line between the intentional stance and the design stance. For, as Haugeland has pointed out, Dennett's definition of rationality effectively collapses the two stances [Haugeland, 1980]. Noting that Dennett defines rationality as "optimal design relative to a goal," Haugeland observes that a well designed system, a mouse trap for example, more or less meets all criteria for an intentional stance treatment. Since "good overall design" is not effectively distinguishable from "optimal design relative to a goal" we can justifiably treat well designed systems as intentional systems. The intentional stance appears to be that special case of the design stance reserved for devices of optimal or near optimal design.

For Haugeland the deeper problem is that the intentional stance fails to provide a level of analysis capable of capturing any kind of organization or systematicity distinct from that accommodated by the design stance. Haugeland sees the situation in roughly the following way: Each of Dennett's stances appears to be designed to capture a distinct kind of order in observable phenomena: The physical stance captures the universal regularity in phenomena; the design stance captures the functional organization of certain systems, i.e. the kind of order apposite to answers to the question "How does it work?"; the intentional stance captures the rational organization of behavior. Haugeland's observation is that, on Dennett's criteria for rationality, "rational order" collapses back into a kind of "functional order." Haugeland's point is well taken, but it does not constitute an objection that Dennett need find disconcerting. After all, Dennett holds that any phenomenon predictable from the intentional stance can also be predicted from the design stance and that there is "no difference in nature" between the phenomena captured by each of the three stances. Hence, Dennett may acknowledge the accuracy of Haugeland's observations without the slightest hesitation for there is no reason to believe that the intentional stance should be sharply distinguishable from the design stance.

The design stance is, however, faced with other problems. In lieu of constraints that serve to unambiguously fix the proprietary functions or purposes of various systems; the designs of those systems cannot be objectively specified. And, it is implausible to believe that the requisite constraints can be made available for a wide range of systems -- in particular, for natural systems or non-artifacts. Thus, an unmitigated indeterminacy afflicts all specifications of the design of certain systems.

To see this consider the problem of specifying the design of any system for which the proprietary functions and purposes are unspecified. Suppose that we are asked to decipher the design of what looks like an alarm clock but that we are not told what purposes the device is supposed to fulfill, or the design goals which guided the device's creator. We might assume that the design goals of the system, i.e. the purposes to be served by the system, are reasonable accuracy as a time piece and reliability as an alarm. Given these design goals, we would proceed to trace out the various subsystems subserving the functions attributed to the system. In the course of this task certain aspects of the system's structure will be ignored. For example, are the assorted wheels, gears, and levers made of brass, tin, or steel? Further, certain features of the system will be taken to reflect bad design, e.g. a plastic wheel employed in a role which will cause it to suffer early wear, or a motor that is impossible to lubricate. But unless we have independent access to the design goals of the system, perhaps through interrogation of the engineer responsible for the botched job, we cannot lay claim to having specified the actual design of the system. What we take to be instances of bad design may actually reflect engineering genius if only we knew the actual design goals of the system and, thereby, the functions to be supported by the various subsystems. A case in point: possible design goals for alarm clocks include not only such things as reliability but such things as intolerability. An engineer might be charged with the task of designing a relatively reliable alarm clock that, nevertheless, ever so slightly hums, grinds, and squeaks while it operates. This is not a praiseworthy design goal to be sure, but it is one which may serve an important function, i.e. to irritate the clock's owners so much that they will be motivated to purchase another, more expensive, alarm clock.

This perverse case has all the features endemic to specifications of design in general. The design attributable to a system is relative to the functions that the system is assumed to serve, but for any system there are different sets of functions that it can be taken, with equal justification, to serve. This means that the design of a system can be consistently interpreted in a number of incompatible ways: optimally designed, sub-optimally designed, or poorly designed; designed to do *X* or designed to do *Y*, where '*X* → -*Y*' and '*Y* → -*X*' hold.

Design stance predictions of behavior are predicated upon the assumption that each element or subsystem will perform its proprietary function. Thus, different predictions will be made regarding the behavior of a system depending upon the functions or purposes which one supposes the system to serve. Biological and evolutionary considerations might lead us to attribute purposes like survival and reproduction to human agents and, thereby, to predict behavior related to such purposes. But, such a basis for the prediction of behavior is incomplete insofar as there are genuine instances of behavior that do not serve survival or reproductive functions. Moreover, *any* assignment of purposes or functions to an individual human agent is bound by an incomplete basis for capturing the behavior of the agent since an individual can always subvert the functions attributed to him by performing certain intentional actions: Though a putative purpose is survival, an individual can take his own life; and though a putative purpose is the creation of pleasure, an individual can cause himself pain. There is no reason to believe that we can somehow fix the grand and unalterable functions of human agents relative to which we are to determine the design of such agents, i.e. their purpose relative structure. The problem with the design stance lies squarely with the application of the purpose-relative concept of design to systems which are not constructed with particular purposes in mind. The indeterminacy which afflicts purpose-relative attributions of design to such systems renders such design specifications empirically vacuous.[5] Since this includes all natural systems or all non-artifacts, the design stance does not provide an objective basis for predictions of the behavior of human agents. This is an important observation to make regarding the design stance since Dennett considers this stance somehow "more objective" than the intentional stance. If it were to turn out that we were forced to forego the use of the intentional stance, the design stance would not provide an unobjectionable, objective alternative.

## PROBLEMS CONFRONTING PHYSICAL STANCE PREDICTION

Recall that Dennett says that "if you want to predict the behavior of a system, determine its physical constitution...and use your knowledge of the laws of physics to predict the outcome for any input" [1979, emphasis mine]. Now, we may grant the universality of physical theory in the sense that any physically describable occurrence in a system is subsumable under physical law, but this wholly innocuous acknowledgement does not entail the view that human behavior can be predicted, in any meaningful sense, through the application of physical law. We have already seen why this is so: A token of a physical output type, i.e. a type specified or individuated by a physical description, may instantiate type-distinct tokens of behavior. For example, a specific sequence of bodily motions may count as an instance of greeting a stranger or of signalling danger. Hence, if we simply predict the future occurrence of a token of a physical output type, the prediction will not specify the type of behavior to be produced. Such a prediction is, in a sense, multiply ambiguous between a variety of behavioral interpretations.

Since Dennett comes close to acknowledging this observation at times -- he suggests, for example, that reduction of psychology to physical theory is a mistaken goal [1978, p. xvii] -- it is curious that he continues to proclaim the adequacy of the physical stance for behavioral prediction. In "Intentional Systems," originally published in 1970, the emphasis is on the rejection of the intentional stance in favor of the design stance and the physical stance. Witness the following remarks:

> The decision to adopt the [intentional] strategy is pragmatic, and not intrinsically right or wrong. One can always refuse to adopt the intentional stance....One can switch stances at will without involving oneself in any inconsistencies or inhumanities. [Dennett, 1978, p. 7]
>
> If one wants to predict and explain the "actual, empirical" behavior of believers one must...cease talking of belief; and descend to the design stance or physical stance for one's account. [Dennett, 1978, p. 22]

Similar commitments are almost wholly retained in more recent writings. The point is worth belaboring since the thesis of the empirical adequacy of the physical stance conceals a primary motivation for appeal to a level of psychological explanation largely autonomous from physical theory. Dennett's view of the empirical adequacy of the physical stance for behavioral prediction is nowhere put more clearly than in the following passage.

> In [certain] cases the only strategy that is at all practical is the intentional strategy; it gives us predictive power that we can get by no other method. But, it will be urged, this is no difference in nature, but merely a difference that reflects

> upon our limited capacities as scientists. The LaPlacean omniscient physicist could predict the behavior of a computer -- or of a live human body, assuming it to be ultimately governed by the laws of physics -- without any need for the risky, short cut methods of either the design or intentional strategies. [Dennett, 1979]

Dennett appears to hold that since the super-physicist can predict every physical state of a physical device, the super-physicist can predict all the behavior of the device. But, the terribly loose sense of "behavior" that would make this thesis true would, at the same time, render it uninteresting, vacuous. This can be made clear by the following considerations.

If it turned out, contrary to fact, that all behavioral types were physically specifiable, then a theory in which only terms of the physicalistic idiom occur could be employed to predict instances of behavior. For, each token predicted would be assignable to a physically circumscribed type that happens to be co-extensive with a psychologically relevant behavioral type. But if there is no type-to-type relation between physically specifiable types and behavioral types, then predicting the occurrence of a physically described token will constitute only the prediction of an occurrence with which *some behavior or other* is to be identified. No determinate predictions of behaviour will be forthcoming from the physical theory. Now one may use the term 'behavior' to refer to all the physical states of a device and, thus, hold that the behavior of any physical system is predictable by a theory capable of specifying the sequence of physical states or physical transformations of the device. But, this says nothing about the relationship of physical theory to psychological theory and cannot be construed as an argument for the empirical adequacy of the physical stance.

What needs to be guarded against is the notorious slide from talk of intentional stance "predictions of behavior," through design stance "predictions of behavior," to physical stance "predictions of behavior." Dennett sometimes appears to suggest that psychologically relevant types of behavior are straightforwardly amenable to systematic treatment on each level. Once one abandons the attempt to specify predicates of physical phenomena coextensive with predicates of psychological phenomena, as Dennett has in eschewing psychological reduction, one must abandon the view that a theory in which only physicalistic terms occur is an adequate basis for predicting behavior.

Now it might be thought that one can consistently cleave to anti-reductionism and the thesis of the empirical adequacy of the physical stance by taking an eliminativist view toward behavioral phenomena. On this combination of views one would (a) admit that it is impossible to specify physicalistic predicates that are coextensive with the predicates that specify intentional behavioral types, (b) hold that the physical stance is an adequate basis for predicting behavior, and (c) attempt to reconcile (a) and (b) by taking the view that putative behavioral types mentioned in (a) do not represent genuine equivalence classes of real phenomena. It is not clear that the kind of eliminativism in question here is completely coherent or how one might argue for it. In any case, it is not a view that Dennett has espoused. At one time Dennett identified himself as an eliminativist with respect to certain mental states, such as beliefs and desires [1978, p. xx], but this is not the kind of eliminativism that would reconcile anti-reductionism with the physical stance -- incidentally, Dennett's eliminativist attitude may have changed for he is now prepared to say such things as "belief is a perfectly objective phenomenon" [1979]. Importantly, Dennett accedes to the view that intentional behavioral types, i.e. those types which resist physicalistic reduction, enter into systematic and, hence, genuine patterns of activity in the world. Dennett asks us to consider a group of Martian super-physicists predicting the activity of a Wall Street stockbroker:

> Take a particular instance in which the Martians observe a stockbroker deciding to place an order for 500 shares of General

> Motors. They predict the exact motions of his fingers as he dials the phone, and the exact vibrations of his vocal cords as he intones his order. If the Martians do not see that indefinitely many different patterns of finger motions, vocal cord vibrations -- even the motions of indefinitely many different individuals -- could have been substituted for the actual particulars without perturbing the subsequent operation of the market, then they have failed to see a real pattern in the world they are observing. [Dennett, 1979]

The crucial acknowledgement here is that a prediction of a physically specified sequence of events, such as the exact finger motions of someone dialing a phone and the subsequent motions of the individual's vocalization system, cannot fully capture the behavioral phenomenon realized by those physical events. For a given type of behavior, ordering 500 shares of General Motors, can be realized by an open, disjunctive class of physical events. Thus to capture the concept of the behavioral type, it is not enough to merely possess the capacity to predict the occurrence of token physical events that constitute possible realizations of behavioral tokens. As Dennett observes, having a concept of a behavioral type involves seeing that many different physical events, which otherwise may have nothing interesting in common, all fall under a common intentional description.[6] And, although Dennett does not say so, having a concept of a behavioral type involves seeing that extremely similar physical events, physically type-identical events, may fall under different intentional descriptions.

Insofar as Dennett recognizes that intentional behavioral types capture a genuine kind of systematicity in observable phenomena, which cannot be captured by non-intentional types, the eliminativist stance toward irreducible behavioral types is not open to him. The thesis of the empirical adequacy of the physical stance, is, then, best foresworn.

NORMATIVE ASPECTS OF THE INTENTIONAL STANCE

In this section we will pose a question for Dennett's characterization of the intentional stance. The crucial issue concerns the sensitivity, if any, of the intentional stance to empirical features of those systems whose behavior is predictable from the stance. We will suggest that although the intentional stance makes use of a normative model of the agents of behavior, no normative model is predictively useful unless constrained by empirical considerations. This circumstance raises certain vexing questions concerning the putative need for a normative component in intentional explanation.

On the picture that Dennett provides, a mark of the intentional stance is that it utilizes a fully normative theory of rationality as a calculus of behavior. The normative character of the intentional stance makes it largely unconstrained by empirical fact. On the intentional stance one assumes that <u>what ought to be the case, will be the</u>

case -- inverting the is-ought question familiar from metaethics. On this simplified picture of intentional explanation, the freedom from empirical constraints enjoyed by the intentional stance may be regarded as an advantage, of sorts, of the theory over design stance and physical stance theories, which are highly constrained by empirical considerations. The intentional stance can be used to predict the behavior of animals and of machines in such a way that its degree of success is determined by the extent to which animals are well ensconced in their environmental niches and to the extent that machines are well designed. Yet, the theorist applying the intentional strategy in predictions of behavior need not know the details of an organism's ecological embedding or the details of an artificial system's design. But, if the predictive successes of the intentional strategy are independent of empirical knowledge, then so are the failures of the strategy. When a system's empirical behavior is determined by a less than fully rational process, unbeknownst to the theorist, the hypothesis that what ought to be the case will be the case, is falsified and the intentional stance will not yield good results.

In circumstances in which the intentional stance fails to provide accurate predictions of behavior, Dennett's recommendation is that the theorist change stances. Consider the following remark:

> No one is perfectly rational, perfectly unforgetful, or invulnerable to fatigue, malfunction or design imperfection. This leads inevitably to circumstances beyond the power of the intentional stance to describe....In cases of even the mildest and most familiar cognitive pathology...the canons of interpretation of the intentional strategy fail to yield clear, stable verdicts about which beliefs and desires to attribute to a person. [Dennett, 1979]

Where the normative techniques of belief and desire attribution are out of step with the manifest behavior of a system, the system is not to be treated as if it internalizes beliefs and desires. For, "intentional explanation and prediction cannot be accommodated either to breakdown or to less than optimal design" [Dennett, 1978, p. 20].

What is slightly curious about this view is that any adequate set of rules for the attribution of beliefs and desires will in fact make essential contact with empirical considerations: Belief and desire attribution, even where the intentional stance generates accurate predictions, is rarely if ever a wholly normative matter. Given that empirical considerations, i.e. non-normative considerations, are requisite for intentional stance prediction, why must one change stances when the hypothesis of full rationality is falsified? Why not, rather, augment the empirical content of the intentional stance itself by making attribution rules sensitive to empirical aspects of belief-desire systems?

To see that normative constraints are not, by themselves, sufficient for the attribution of beliefs and desires one need only observe that such attributions cannot take place in an empirical vacuum. At the very least, the theorist requires information about the range of physical phenomena to which a system is sensitive, e.g. a description of the sensory modalities that provide input, including information regarding, among other things, the thresholds of sensitivity characteristic of the various input transducers. The theorist won't attribute the belief that 'there is a mouse in the grass' to a human 500 feet aloft in a balloon, but he might attribute such a belief to an eagle 500 feet aloft.

In general, intentional stance belief attribution operates on the following pattern: Given a catalogue of a system's desires and goals, one attributes those beliefs that constitute representations of information, available to the system, which is crucial to the fulfillment of those desires and goals. Dennett puts this in a roughly equivalent form, "attribute as beliefs all the truths relevant to the system's interests (or desires), which the system's experience to date has made available" [1979]. Such a belief attribution strategy relativizes a system's belief-set to its desires and, thereby, makes the attribution of desires a fundamental or antecedent step in the attribution of beliefs. But, the attribution of desires is at base an empirical project. Though Dennett proclaims that the fundamental rule for desire attribution is "attribute the desires a system _ought to have_," it turns out that he recognizes that there is no purely normative basis for the attribution of basic desires. Dennett characterizes the techniques of desire attribution as follows:

> We attribute the desires that the system _ought to have_. That is the fundamental rule. It dictates, on a first pass, that we attribute the familiar list of highest, or most basic, desires to people: survival, absence of pain, food, comfort, procreation, entertainment. Citing any one of these desires typically terminates the "why?" game of reason giving. One isn't supposed to need an ulterior motive for desiring comfort or pleasure or the prolongation of one's existence. [Dennett, 1979]

The observation here is that there can be no rationalistic justification for a basic desire: No _reasons_ can be given for possessing a basic desire; one possesses or fails to possess such a desire as a consequence of empirical factors. Obviously, many of our basic desires may be a consequence of the operation of evolution; systems that desire pain tend to disappear where the production of pain states is highly correlated with bodily damage. Dennett is, of course, fully cognizant of such facts and unabashedly makes use of them whenever they seem appropriate.

But, the crucial point here is that the intentional stance cannot operate as a purely normative strategy for the attribution of beliefs and desires. Though we are told to attribute to a system those beliefs and those desires that it "ought to have," we are told that this means that we should attribute those beliefs and desires that a system "would have if it were ideally ensconced in its environmental niche" [Dennett, 1981a, p. 43]. The intentional stance begins to look more like a severe empirical idealization than it does like a normative, rationalistic calculus of action. The empirical considerations essential in belief and desire attribution are hidden from view, to a certain extent, in folk-psychology attribution, calculation, and prediction. But, they are hidden only because they are so familiar and available. Actual experience provides us with information concerning the typical capacities of the various sensory modalities and, thereby, information about what a normal observer in a given set of circumstances might reasonably be expected to come to believe. Folk-psychological attribution is guided by a fabric of rules, some of them normative perhaps, laid over a foundation of empirical hypotheses about the belief formation capacities of individuals and about their characteristic basic desires. Non-empirical, normative rules may figure in attribution techniques by directing the theorist to assume that an agent's faculties are functioning properly -- and thus to attribute the beliefs that a given empirical system "ought to have" -- but such normative constraints are applicable only within empirically specified bounds.

If it is clear that the intentional stance is not completely empirically unconstrained, then an advocate of the putatively normative strategy for behavioral prediction must answer the following question: Why can't the intentional stance be rendered fully empirical -- i.e. as empirical as the paradigmatic "empirical sciences"? For, surely, mere contact with empirical considerations cannot vitiate the intentional stance since it relies upon empirical considerations for constraints upon belief and desire attribution in the original instance. In the same spirit, we might ask, what theoretical role is filled by normative constraints that cannot be filled by empirical constraints or by empirical idealization? Dennett holds that when strategies for the prediction of behavior are made to respect empirical, non-normative aspects of a system, the intentional stance is abandoned. In support of this view Dennett might either (i) stipulate the intentional stance's commitment to the rationality constraint as a matter of definition or (ii) show that the rationality constraint serves some theoretical role unfulfillable by non-normative constraints. Obviously the former, definitional, alternative is unattractive simply because it fails to engage the crucial issue: Does the theorist necessarily revert to a non-intentional form of explanation when the normative rationality constraint is dropped?

Dennett views intentional explanation in general as inextricably connected with the rationality assumption because he sees rationality as in some way criterial to our notions of belief and desire. At one point, for example, Dennett defines 'belief' as a disposition to act rationally:

> what it means to say that someone believes that P is that that person is disposed to behave in certain ways under certain conditions. What ways under what conditions? The ways it would be rational to behave given the person's other beliefs and desires. [Dennett, 1981a, p. 44]

Evidently, an individual has a belief if and only if he is in a state that "disposes" him to behave rationally; a state that disposes a person to behave in a way other than is rational in his present circumstance is, thereby, not a belief. Consider a wide-ranging collection of other remarks Dennett has made about rationality and intentional explanation: "The assumption that something is an intentional system is the assumption that it is rational" [1978, p. 11]; the rationality assumption "generates both an intentional interpretation of us as believers and desirers and actual predictions in great profusion" [1981a, p. 44]; "without the myth of rationality the concepts of belief and desire would be uprooted" [1981a, p. 45]; "we must descend from the level of beliefs and desires to describe [an individual's] mistake [in reasoning], since no account in terms of his beliefs and desires will make sense completely" [1981b, p. 66]. While these claims are lifted from a variety of contexts, they speak to the same general point: Rationality is a precondition for intentional interpretation of a system, for the attribution of beliefs and desires, and for the explanation of behavior in terms of beliefs and desires. Though we may agree that content bearing states, beliefs and desires, are causally involved in the production of much human behavior -- or dispose us to behave in Dennett's terms -- and are therefore to be cited in the explanation of behavior, why should we suppose that we possess beliefs and desires only to the degree that we are rational? Presumably, the rationality or irrationality of belief-driven actions depends upon such things as the reliability of the deductive inferences made from stock beliefs and the reliability of the inductive practices involved in the extraction of beliefs from available evidence. Both processes might depart from canons of rationality in certain instances and produce tokens of behavior that don't fully conform to normative principles. Why, then, view the rationality condition as a necessary constraint on all intentional explanation or on all explanation of behavior that cites beliefs and desires?

Dennett offers several, largely independent, arguments for the rationality condition. Rather than proceed directly to those arguments, however, it will be instructive to take a slightly different course and examine several of the more salient objections that have been made to the intentional stance with its accompanying rationality assumption. Taking this tack will allow us to see how well Dennett's basic characterization of intentional explanation is able to withstand certain criticisms while, at the same time, suggesting areas in which that characterization is vulnerable. It will emerge, I think, that Dennett's views may remain largely intact under the onslaught of certain rather natural lines of criticism that have been offered in the literature, but not under the criticism that will be constructed in succeeding sections.

## CRITICISMS OF THE RATIONALITY CONDITION

The advisability of imposing a rationality constraint on intentional theory has been questioned by Fodor and by Stich. Though their respective criticisms are similar in some respects, Fodor's primary target, unlike Stich's, is Dennett's claim that theories which advert to beliefs and desires can aspire to no more than mere heuristic significance. For this reason let us turn first to Stich's complaints: Stich's primary concern is that the adoption of the intentional stance, constrained by the rationality assumption , "would leave us unable to say a great deal that we wish to say about ourselves and our fellows" [1981, p. 47]. Consider the following characterization of this difficulty for the intentional stance:

> An intentional system, recall, is an ideally rational system; it believes, wants and does just what it ought to, as stipulated by a normative theory of rationality. People, by contrast, are not ideally rational, and therein lies a devastating problem for Dennett. If we were to adopt his suggestion and trade up to the intentional-system notions of belief and desire...then we simply would not be able to say all those things we need to say about ourselves and our fellows when we deal with each other's idiosyncrasies, shortcomings, and cognitive growth. [Stich, 1981, pp. 47-48]

This objection might be interpreted as an attack on the completeness of the intentional stance. That is, a theory that assumes the systems which it describes are optimally rational will have no capacity to model the cognitive shortcomings, the departures from full rationality, of those systems. On the highly plausible assumption that human agents suffer certain cognitive shortcomings, the intentional stance offers, at best, an incomplete picture of the cognitive apparatus with which humans are endowed. Stich seems to suggest this reading when he observes that folk-psychological explanation commonly makes reference to such cognitive failings as forgetfulness and calculational error -- it is as if we need a theory of error to augment a theory of rational, error-free practice.

Now, Dennett readily admits that departures from rationality are recalcitrant to an intentional stance account. Indeed, Stich cites a remark of Dennett's that says as much, "the presuppositions of intentional explanation...put prediction of lapses in principle beyond its scope...." [Dennett, 1978, p. 246; cited in Stich, 1981, p. 55]. The intentional stance is simply not designed to capture instances of cognitive failure. Dennett uniformly suggests that the design stance and the physical stance are to be employed for such purposes. Thus, Stich's complaints can hardly devastate Dennett's position. To construct a damning objection to the intentional stance along these lines, it needs to be shown that the design stance and the physical stance cannot do what is asked of them -- which has, in fact, been suggested above. Alternatively, one might attempt to show that the cognitive failures which the intentional stance fails to accommodate are capturable only by an intentional, i.e. content respecting, theory of cognitive processes. Stich does not, as far as I can determine, directly pursue either line of criticism. Instead, he suggests that our folk-psychological conception of human agents includes an acknowledgement and a treatment of their fallibility. While this may very well be true, it goes little distance toward refuting Dennett's claim that intentional psychology must impose a rationality constraint on descriptions of cognitive processes.

On the other hand, Stich may have in mind an argument to the effect that if (a) folk-theory is an intentional theory, and (b) folk-theory can accommodate lapses from rationality, then all intentional theories need not be constrained by a rationality condition. The trouble, or one trouble as the case may be, with this argument is that premise (b) is notoriously difficult to defend as Dennett points out in a reply to Stich. Dennett sometimes suggests that the intentional stance is just a codification of folk psychology [1981a, p. 45]. If this is correct then folk-psychology can no more accommodate instances of irrationality than can the intentional stance. In any case, shifting the focus of the argument in this direction -- What is folk-psychology in reality? -- is to select a battleground in which the determination of victory and defeat is all too elusive. I think that we must grant Dennett's observation that "the putative mental activities of folk theory are hardly a neutral field of events and processes to which we can resort for explanations when the normative demands of intentional system theory run afoul of a bit of irrationality" [1981b, pp. 70-71]. In summary, Stich correctly notes that cognitive lapses do not make sense from the perspective of the intentional stance, i.e. there can be no intentional stance characterization of cognitive failure or error. But, this circumstance is a difficulty to be charged against Dennett's characterization of intentional explanation only if it can be shown that only an intentional theory can properly accommodate certain cognitive failings.

Fodor's critical treatment of the rationality constraint is embedded in an argument against Dennett's instrumentalistic interpretation of the posits of intentional psychology. Fodor is not so much concerned with the rationality condition itself, but with the role of the rationality condition in Dennett's argument for instrumentalism. Yet, Fodor suggests that there is a fundamental difficulty facing the rationality condition: A characteristic feature of our propositional attitudes is incompatible with full rationality. Essentially, the point is that beliefs attributed to individuals must be opaquely construed while the rationality assumption suggests a transparent construal of beliefs. What it means, for Fodor, to say that beliefs are opaque is at least that certain truth-preserving forms of inference cannot be used to calculate what other beliefs an agent possesses on the basis of the (justified) attribution of a particular belief. For example, from 'a believes Cicero denounced

Catiline' and 'Cicero = Tully' we cannot infer that '*a* believes Tully denounced Catiline'. The issue of the opacity of propositional attitudes is complex and includes, for Fodor, not just the failure of the substitution of identity and the failure of existential generalization in belief contexts, but the lack of deductive closure for belief sets: Fodor maintains that "It is part and parcel of the opacity of *our* propositional attitudes that the inference from 'we believe *P*' and '*P* entails *Q*' to 'we believe *Q*' is not, in general, valid" [1981, p. 107]. This circumstance forms the basis of an objection to the rationality condition, presumably, because a fully rational agent's beliefs would be closed under the entailment relation. Dennett admits to having once suggested such a conception of rationality [see Dennett, 1981, p. 74]. Fodor's observation is that "the more rational the system, the less opaque its belief contexts," but we require an opaque construal of belief contexts.[7]

It is not clear, however, that rationality *per se* is incompatible with the opacity of propositional attitudes. It would seem to be a safe assumption that rational systems can have various degrees of knowledge, ranging from near ignorance to omniscence. Now one might argue that a system which internalizes the belief '*P*' and the belief '*P* entails *Q*', where '*Q*' is consistent with every other belief internalized, is not a rational system unless it internalizes the belief '_Q_'. Certainly a system that disavowed '_Q_' under these conditions would be less than rational. On the other hand, a system that believes '_P_', but does not have the information, does not know and hence does not believe, that '_P_ entails _Q_' may well be considered a rational system. For, it is one thing to criticize a system's rationality and quite another to criticize the scope of its knowledge. Consider the collection of famous mathematical conjectures: Perhaps one or more of them is _entailed by_ the rest of mathematics, but no one _knows_. It would be odd, to say the least, to charge mathematicians with irrationality simply because they have not realized that a certain complex entailment relations hold between a conjunction of propositions that they believe and some conjecture they would like to prove. The moral here is one that Dennett has recently embraced: Rationality is not to be defined by closure of beliefs under the entailment relation [1981, p. 74]. Though we may grant Fodor's observation that "it is part and parcel of the opacity of our propositional attitudes" that they are not closed under the entailment relation, perhaps there are criteria for rationality to be formulated that do not require the deductive closure of belief sets and which, thus, are compatible with the opacity of propositional attitudes.

The problem that faces the attempt to formulate criteria for rationality compatible with the opacity of propositional attitudes is aggravated by the obscurity in which the doctrine of opacity is shrouded. Fodor sometimes suggests that there are a number of independently sufficient conditions for opacity: The failure of existential generalization; the failure of the substitution of identity; and, given the citation above, the failure of deductive closure for belief sets [see Fodor, 1980, pp. 66, 72]. Viewing each as an independently sufficient condition, the opacity of propositional attitudes would be preserved by any criterion for rationality which does not license either rule of inference or the deductive closure for belief sets. This would seem easy enough to provide, in principle, since the substitution of identity, for example, is only licensed by imposing wildly unreasonable criteria for rationality: It would require not only the internalization of a substitution rule for the subject terms of representations of the form '<u>a</u> is <u>B</u>', but the internalization of a catalogue that specifies, for each term that may occur in the subject position of a representation, the complete set of extensionally equivalent terms. But, there can be no <u>a priori</u> specification of such a catalogue and, hence, the requirement for such a catalogue is a requirement on the scope of a system's knowledge, not a requirement on the degree of its rationality. We will have more to say about the issue of the opacity of mental states in another chapter; suffice it to say for now that the incompatibility of opacity with a tenable conception of rationality has not been demonstrated.

At the same time, to give Fodor his due, his remarks are directed toward a conception of rationality which Dennett once explicitly enunciated. Moreover, Fodor suggests that Dennett might evade the "opacity objection" to the rationality condition by attenuating this conception of rationality. Fodor entertains the idea that the relevant construal might just be non-irrationality, i.e. "that the system whose behavior is at issue won't act at random or go nuts" [1981, p. 109]. Dennett has tended to allow for the attenuation of the optimal rationality constraint on intentional explanation in recent works -- whether at Fodor's behest or not. Forfeiting the assumption of optimal rationality for intentional systems, Dennett says:

> I want to use 'rational' as a general-purpose term of cognitive approval -- which requires maintaining only conditional and revisable allegiances between rationality, so considered, and the proposed (or even universally acclaimed) methods of getting ahead, cognitively, in the world. [Dennett, 1981, p. 76]

By adopting this characterization of rationality, Dennett may successfully obviate Fodor's, and for that matter

Stich's, objections while retaining the spirit of the intentional stance commitment to a rationality constraint. Nevertheless, Fodor and Stich are, I think, essentially correct: Any substantive rationality condition is bound to be incompatible with certain properties of our belief systems, and any theory constrained by a substantive rationality condition is bound to make for an incomplete psychological theory -- the two points are obviously closely related.

The difficulties encountered in urging Fodor's and Stich's objections against Dennett's characterization of intentional explanation reflects the considerable solvency of Dennett's position -- at least a solvency possessed under moderate reformulation of the criteria for rationality. Yet, there are powerful and natural considerations which, it seems to me, suggest that the connection between intentional explanation and the assumption of rationality is not the connection that Dennett proclaims. In order that we may better understand these considerations, let us turn to an examination of certain arguments that have been made in favor of imposing a rationality constraint on intentional explanation.

## ARGUMENTS FOR THE RATIONALITY CONSTRAINT

By definition, if one is to adopt the intentional stance one must assume that the system whose behavior is to be predicted is rational. Now, one might take it to be an argument for the rationality constraint that the intentional stance actually works to a degree, i.e. it yields accurate predictions in a significant number of instances. Dennett has repeatedly claimed that the intentional stance is often used and that it often works [see 1978, pp. 8-11; 1979; 1981a, pp. 44-45]. Let us suppose, for the moment, that this observation is correct. Do we have, then, an argument for imposing a rationality constraint on intentional explanation in general? Consider an altogether different domain of explanation: Suppose one were to argue that physical theories, of medium-size, slow-moving bodies, that impose a Euclidean geometry on space often work; i.e. that the accuracy of the predictions provided by such theories is statistically significant. No one would take such a consideration to demonstrate the necessity of a Euclidean constraint on physical explanation. Similarly, the putative fact that the intentional stance works cannot demonstrate that its defining assumptions represent necessary constraints on intentional explanation.

Although Dennett goes out of his way to argue that the intentional stance can actually work -- even in places in which he contends that intentional stance explanation is vacuous [compare: 1978, pp. 8-11 and pp. 15-16] -- he does not suggest that the pragmatic usefulness of the rationality constraint, alone, makes it ineliminable in intentional explanation. For that purpose, Dennett brings to bear a wide variety of other considerations. There are, however, two themes that appear most prominent in Dennett's advocacy of the rationality constraint. One involves the problem of securing predictive power in an empirically insensitive theory. The other theme is constituted by an analysis of folk psychology as a species of reason-giving explanation. These themes will be treated in turn.

Securing predictive power is no small task. Recall that, for Dennett, the intentional stance is, putatively, empirically unconstrained, though it has been suggested here that this view is at least slightly misleading. The intentional stance must supply predictions of actual, empirical behavior -- which it supposedly does in voluminous quantities -- without empirical information about the systems whose behavior it predicts. Given that the intentional stance is sundered free from empirical constraints, non-empirical constraints of some kind must be provided to guide its predictions. Where else but to normative constraints is the theorist to turn? Thus, if we were convinced that intentional explanation is intrinsically non-empirical, we might have an argument by exhaustion of a simple pair of possibilities, i.e. empirical constraints or non-empirical constraints, to the conclusion that intentional explanation is essentially normative. But, this way of looking at the issue is not informative since one important question at stake is whether or not intentional explanation should be construed in such a way as to render it empirically insensitive in the first place. The task, then, is to show that intentional explanation is essentially normative.

Intentional explanation, Dennett will agree, adverts to such things as beliefs and desires. If the theorist is to predict behavior on the basis of ascription of beliefs and desires, configurations of beliefs and desires must be correlated with behavioral types by formulation of intentional, psychological generalizations. Alternatively, such generalizations might be thought of as characterizing the ways that beliefs and desires figure into the etiologies of behavior. How do beliefs and desires interact to produce tokens of behavior? Dennett holds, roughly, that the theorist engaging in intentional explanation must necessarily suppose that beliefs and desires interact in nothing less than a (fully) *rational* way for,

> one gets nowhere with the assumption that entity *x* has beliefs *p*, *q*, *r*... unless one

> also supposes that x believes what follows from p, q, r...; otherwise there is no way of ruling out the prediction that x will, in the face of its beliefs p, q, r...do something utterly stupid, and, if we cannot rule out that prediction, we will have acquired no predictive power at all. [Dennett, 1978, p. 11]

The point that Dennett endeavours to illustrate is that any intentional theory that foregoes the imposition of a rationality constraint will be a powerless predictive tool. Dennett strongly suggests that the rationality constraint serves as an idealized guarantor that all consequences of beliefs will be believed. But, what is essential in Dennett's argument is not that belief sets are actually closed under the consequence relation -- which we have observed they are not -- but that beliefs have a systematic role in the production of other beliefs and of behavior: Generating behavioral predictions by calculating the effects of configurations of beliefs and desires requires that beliefs interact with beliefs and with desires in regular or systematic ways. The rationality constraint simply directs the theorist to assume that <u>logic</u>, e.g. propositional logic, quantificational logic, decision theory, models the relations among propositional attitudes. Such an assumption imposes a kind of systematicity on relations among propositional attitudes describable by the theory. For Dennett, logical or rational systematicity in relations among beliefs is essential for intentional prediction since, presumably, without such a systematicity a system would act randomly or it would often do "something utterly stupid."

Now, theories of behavior that take behavior to be contingent upon configurations of propositional attitudes are worthless predictive tools if relations between propositional attitudes, and between *configurations of* propositional attitudes and types of behavior, are arbitrary. But, what has not been demonstrated is that conformity to normative principles, the rules of logic, constitutes the unique sufficient condition under which intentional predictions are possible. Such a demonstration would require a proof that there are no *a posteriori* constraints adequate for that purpose. It is here that a major lacuna in Dennett's argument is found. Empirically motivated models of relations among beliefs are not considered as possible sources of constraints on intentional predictions since, apparently, on Dennett's view to formulate such models would be to change stances. But, if the systematicity in relations among beliefs can be empirically characterized, then the normative -- and contrary to fact -- model provided by the rationality assumption might be replaced.

It is worth noting that Dennett's "argument from predictive power" has been adopted by others. C. Cherniak, for example, has recently argued that any theory of belief which imposes no rationality constraint on belief systems "is without predictive power: using it we can have virtually no expectations regarding a believer's behavior" [1981, p. 164]. The argument for this conclusion is that according to a belief theory lacking a rationality constraint,

> no inferences, however obvious and useful, need be made from the beliefs, and the belief-desire set is not required to guide at all the choice of appropriate actions.... Consequently, the theory...makes a mystery of our everyday successes in predicting behavior on the basis of belief-desire attributions. [Cherniak, 1981, p. 164]

Such an argument assumes that if beliefs are not related in a rational way, or for Cherniak a "minimally rational" way, then they are related in unprincipled ways. This is an unwarranted assumption. The de facto relations among beliefs need not be rational to be non-arbitrary or systematic. If we view beliefs and desires as progenitors of other beliefs and desires and of tokens of behavior, then we naturally take beliefs and desires to be embedded in a causal nexus of some sort -- a computational system according to one theory. Further, where there are causal connections there will be, in general, empirical regularities. Certain causal regularities between semantically interpreted beliefs may conform to truth-preserving rules of inference and, thereby, constitute rational connections between beliefs. But there is no particular reason to believe that regularity is exhibited by a belief system only if all relations among beliefs conform to normative principles, yet this is precisely what we are asked to believe by Cherniak and by Dennett.

The observation that an empirical theory of relations among beliefs may support intentional predictions is due, originally, to Fodor: "it is not postulates of rationality that license intentional stance predictions: it is (mini- or maxi-, formal or informal) theories about who is likely to have which propositional attitude when, and what behavioral consequences are likely to ensue" [1981, p. 108]. Now, one plausible belief attribution rule is "if _P_ · (_P_ → _Q_) is attributed to an agent, then attribute _Q_." Though the "logic" of the rule is unassailable, for Fodor the reliability, if any, of a belief ascription practice is a matter of _fact_, not a matter of _logic_:

> If an organism has the belief that _P_ and the belief that if _P_ then _Q_, then it has the belief that _Q_. This is, of course, _not_ a logical truth; it's an empirical generalization which specifies (presumably causal) relations among mental states by reference to the content of those states. As such, it will be true only if the mental states, together with their causal relations do, in the intended sense, constitute a model of the logic. [Fodor, 1981, p. 114]

According to this causal-empirical analysis, the conformity or non-conformity of relations among beliefs to principles of logic is determined by causal mechanisms that structure belief sets. It is difficult to see how Dennett and Cherniak could avoid this analysis. Dennett, for example, has suggested that the intentional stance works to the degree that nature has designed us to be rational: "treating each other as intentional systems works...because we really are well designed by evolution" [1981a, p. 44]. Dennett would appear to hold that there are relations among beliefs determined by some natural, evolved system which happen to constitute a model of the normative constraints imposed by the rationality assumption. In short, the appeal to evolutionary considerations seems tantamount to the admission that there are empirically specifiable relations among certain mental states.

Whether or not Dennett will allow that there may exist causally determined regularities among beliefs, the tactical point is that Dennett's advocacy of normative constraints on intentional explanation is consonant with the existence of causal, or otherwise naturally determined, regularities among beliefs. And, where there are causally determined relations among mental states, a theory acquires predictive power and reliability by constraining its calculations by a characterization of empirically possible relations among mental states. So if the theorist wants to provide for predictive power, one way that he might do so is by formulating an empirical theory of belief systems. As long as this is a possible option, one cannot demonstrate that strategies for the prediction of behavior that advert to beliefs and desires must constrain belief attribution by normative criteria *for rationality*. For, even if the rationality constraint affords some measure of predictive power and reliability, it may do so only because there are underlying causal regularities, between semantically endowed mental states, that can be interpreted as instantiations of certain principles of logic.

Another argument for the rationality condition prominent in Dennett's work is founded upon observations of folk-psychological practice. As we have observed, at times it appears that the intentional stance is offered as an analysis of folk theory [1981a, p. 42]. In any case, Dennett's central observation is that folk-psychological accounts of behavior are typically reason-giving accounts. The folk theorist explains why an agent did a particular action by attributing a goal to the agent and then showing how the action performed can be construed to be a rational means for securing the attributed goal. Thus, reason-giving accounts of behavior offer *rational* justifications for behavior. Dennett conceives of intentional explanation as a form of reason-giving explanation: The pattern of explanation found in folk psychology is deployed as a strategy for predicting behavior simply by imposing the rationality constraint on attributions of beliefs and desires and by assuming that agents will do what they ought to do. Folk psychology often offers retrospective rational justifications for tokens of behavior; the intentional stance predicts only rationally justifiable tokens of behavior.

Now, reason-giving explanations of behavior appear to presuppose the rationality of the agents whose actions they subsume. After all, it is useless to offer a rational justification for the behavior of an arational or of an irrational agent. Dennett says,

> a feature of folk psychology that sets it apart from both folk physics and the academic physical sciences is the fact that explanations of actions citing beliefs and desires normally not only describe the provenance of actions, but at the same time define them as reasonable under the circumstances. They are reason-giving explanations, which make an ineliminable allusion to the rationality of the agent. [1981a, p. 42]

The argument for the ineliminability of the rationality constraint implicit here appears to rely upon the conceptual connection between the ideas of reason-giving explanations and rationality. There is, no doubt, an important connection between the two concepts for, presumably, an agent that is not "rational" in some appropriately broad sense cannot act out of reasons: So, to a first approximation, a reliable principle might be that "to the extent to which it is supposed that agents act out of reasons, agents are rational." Dennett, it would seem, defends some version of this principle.

These considerations, however, do not establish the need for substantive rationality constraints on intentional explanation. The "rationality" needed to underwrite reference to an agent's reasons for behaving the way he does is indifferent between ideal rationality, minimal rationality, and irrationality. All that is required is that the agent act from factors interpretable as reasons. It is not required that an agent's reasons are uniformly good or that they epistemically justify his actions. Nothing would seem to be clearer than the fact that agents act from bad reasons on occasion. Thus, a theorist might deploy the actual reasons from which an agent acts without thereby offering reasons that rationally justify the agent's actions. If this is conceptually possible, then intentional explanation must not be uniformly construed as <u>explanation by justification</u>. A simple example will serve to illustrate the point: Consider an inferential error in which a relatively primitive vegetarian with the true belief that 'if something is edible, then it is vegetable matter and it is green' infers that 'this is edible' upon forming the belief that 'this is green vegetable matter.' If the agent's world is forgiving, his ensuing ingestion of the green vegetable matter may have little adverse consequence. But, in such a case, the theorist does not show the agent's action to be justified by citing the manifest reasons out of which the agent acts. In this instance the agent's actual reasons are subject to rational criticism: Acting on beliefs derived by "affirming the consequent" is dangerous for one's health.

If the theorist heeds Dennett's recommendations and attributes only those beliefs and desires that justify actions, then there can be no basis for the rational criticism of an agent's beliefs, inferential practices, and decision strategies. Moreover, provision of a basis for the criticism of "reasons for actions" would seem to be an important desideratum for intentional theory. But, we cannot describe the reasons for action formulated by agents in such a way as to make them amenable to rational criticism if we constrain all attributions of reasons by the rationality assumption. We are confronted with a clear choice: either cleave to the rationality constraint and forfeit the possibility of criticism or preserve the possibility of criticism and reject the rationality constraint.

Current research strategies in the field of human decision making have shown a preference for the latter alternative. The assumption of such strategies is that it is possible to describe the actual reasons out of which agents act without imposing a rationality constraint. Part and parcel of this approach is an acknowledgement that an agent's motivations needn't be rationally unassailable in order to constitute reasons for action. A. Tversky and D. Kahneman, for example, have become widely known for offering empirical accounts of decision procedures, utilized by actual agents, that do not conform to those of idealized, normative decision theory. Of course, Dennett is fully cognizant of the work of Tversky and Kahneman. What is interesting, in the present context, is not that such work shows that people are less than optimally rational, but that it reveals that the reasons for which individuals act can be described even though those reasons are not predictable by a theory that assumes optimal rationality. Consider some very basic observations concerning human decision making. One normative rationality constraint is the following. 'When confronted with a choice among alternatives, always select the alternative with the greatest monetary expectation.' It has been widely recognized, apparently since work of Bernoulli in

1738, that this principle does not uniformly guide human behavior. Individuals do not uniformly select the outcome with greatest monetary expectation. Sure gains are often preferred over possibilities of even greater gains: Given a choice between (a) $80 and (b) an 85 percent chance of winning $100, people choose the $80 even though the monetary expectation (product of the amount of a possible gain and its probability) of (b) is $85. The lesson seems to be that people tend to show "risk aversion." But, when people are forced to choose between alternatives each of which represents a loss, they tend to show "risk seeking." Sure losses are often rejected in favor of possibilities of even greater losses [see Kahneman and Tversky, 1982, pp. 160-173]. Neither of these results would be predicted by a theory that includes the normative rule that directs the agent to choose the alternative with the greatest monetary expectation. More importantly, the empirically demonstrable tendency of individuals to overvalue certainty when confronted with a choice between gains illustrates the claim made above. The reasons out of which people act do not coincide with the reasons out of which it would be rational to act according to the normative theory. One chooses $80 over an 85 percent chance of winning $100, but one does so for reasons that are amenable to criticism. Thus, the argument for the rationality constraint which appeals to the conceptual connection between the concepts of "acting for reasons" and "rationality" fails to show that reasons are attributable only where a normative model of rationality constrains the attribution process.

Dennett, in a closely related argument for the rationality constraint, has suggested that the theorist has no access to less than (fully) rational beliefs. Dennett proclaims:

> This is not to say that we are always rational, but that when we are not, the cases defy description in ordinary terms of belief and desire....An intentional interpretation of an agent is an exercise that attempts to make sense of the agent's acts, and when acts occur that make no sense, they cannot be straightforwardly interpreted in sense-making terms. Something must give: we allow that the agent either only "sort of" believes this or that, or believes this or that "for all practical purposes," or believes some falsehood which creates a context in which what had appeared to be irrational turns out to be rational after all. [Dennett, 1981b, p. 67]

In one sense, this is perfectly correct. Reason-giving explanations must pass a test of intelligibility, i.e. putative reasons must make sense to someone or they will not constitute genuine reasons for performing an action.

However, the requirement that configurations of beliefs and desires which lead to actions "make sense" does not show that we must assume that believers satisfy idealized, normative canons of rationality. Given a specification of an agent's semantically interpreted beliefs and desires we can ask whether they make sense, but to whom must they make sense? To a Bayesian decision theorist? To the agent's social peers? To the agent himself? Once such questions are taken to heart, the problem of belief attribution can be transformed from the problem of determining the most rational beliefs that could be formed in a given set of circumstances to that of determining what beliefs will make subjective sense to an agent in that set of circumstances. There are, roughly, two available measures of what "makes sense": subjective, almost phenomenological tests; and objective, normative standards. The two measures will not deliver the same evaluation of each case. When our hungry vegetarian in the example above confuses a necessary condition for edibility with a sufficient condition for edibility he is endeavouring to evaluate his circumstances by reference to principles available to him. The fact that he acts upon an illicitly derived conclusion shows that he is willing to bet heavily upon the cogency of his cogitations. His evaluation of the situation makes sense, or so it seems to him for the moment. Yet, his subjective appraisal of the situation is easily faulted, and might on another occasion be criticized by the agent himself.

Dennett suggests that instances of error in judgement fall outside the domain of intentional explanation since any account "will have to cope with the sheer senselessness of the transition in any error" [1981b, p. 66]. But this way of looking at the situation gives us an unreasonably conservative access to an agent's beliefs. For, following Dennett, when an agent avows the antecedent of a conditional upon ascertaining the truth of its consequent he has made a senseless transition and cannot rightfully be ascribed a belief with the content of the antecedent. A tenable theory of error, contrary to Dennett's suggestions, would appear to require an acknowledgement of the possibility of belief formation through vicissitudes of less than wholly reliable inferential practices.

Access to unjustifiable or fallaciously derived beliefs is not possible, of course, if one demands that all belief attribution be conditioned by normative assumptions. An empirical theory of belief systems is ruled out by the view that we must preserve the rationality of agents in interpretations of their beliefs and desires. But what argument could show that we must so preserve the rationality of agents? If the argument is supposed to be that rationality preserving interpretations are always in principle available, then it falls short of showing that rationality preserving interpretations are to be preferred.

The weight of evidence would seem to indicate that an agent's subjective sense of what makes sense does not always conform to canons of idealized rationality. How else could it be that unintentional errors are committed in most inopportune circumstances? As previously noted, Dennett sometimes agrees that human agents are not fully rational [1979]. Thus, he cannot avail himself of the view that only characterizations of agents under which rationality is uniformly preserved are allowable -- even though he holds that only rational beliefs are attributable. In the final analysis, Dennett's observation that intentional interpretations must "make sense" falls short of establishing that attributions of beliefs and desires must be constrained by normative criteria for rationality since beliefs may satisfy subjective conceptions of what makes sense without conforming to any normative standard.

## RATIONAL COMPETENCE

Any treatment of arguments for the imposition of a rationality constraint on intentional explanation would be seriously incomplete without a consideration of the notion of rational competence. In this regard, an argument has recently been constructed by L.J. Cohen that might be adduced in support of Dennett's conception of intentional explanation. Dennett has, in fact, made reference to Cohen's argument [see Dennett, 1981b, p. 66].

Cohen suggests that we may introduce a competence-performance distinction in the domain of human rationality on analogy to the competence-performance distinction so crucial in linguistics.

Consider the role of the distinction in linguistic theory. The problem of specifying the "knowledge of language" possessed by native speakers is hardly interesting if all utterances, all linguistic performances, are taken to be well-formed elements of the speakers' language. Since performance may contain virtually any sequence of expressions, a "grammar" for performance might just be a system that generates all random sequences of morphemes. The notion of an internalized grammar -- a competence model for linguistic knowledge -- becomes interesting where there exist principled techniques for filtering from a corpus of performance those utterances that do not constitute well-formed elements of the language. Similarly, a theory of human rational competence must rely upon techniques for filtering errors in judgement from the class of all human reasoning performances, in order to define its subject matter. In linguistic theory this task is performed by the linguistic intuitions of native speakers and, thus, "linguistic intuition is the ultimate standard that determines the accuracy of any proposed grammar" [Chomsky, 1965, p. 21]. It is assumed that such intuitions are themselves generated by the speaker's linguistic knowledge or underlying competence.

Cohen argues that since norms of rationality are subjected to the test of intuition, we must similarly suppose that individuals possess an underlying rational competence. Moreover, we must credit individuals with a rational competence fully equivalent to any and all idealized, normative criteria of rationality, i.e. any principle of rationality that is favoured by intuition must be credited to the competence system that generates the intuition. (Compare: "Any set of grammatical rules definitive of the structure of a sentence acceptable to intuition are to be credited to linguistic competence.") Cohen offers several succinct summaries of this argument. In a typical passage Cohen maintains that,

> where you accept that a normative theory has to be based ultimately on the data of human intuition, you are committed to the acceptance of human rationality as a matter of fact in that area, in the sense that it must be correct to ascribe to normal humans [sic] beings a cognitive competence -- however faulted in performance -- that corresponds point by point with the normative theory. [Cohen, 1981, p. 321]

Notice that the argument here appears to support only a conditional conclusion: If normative theories are founded on human intuition, then humans must be credited with a competence on par with those normative theories. But, it should be noted that Cohen carefully argues that there can be no other basis for normative theories than that provided by intuition [pp. 318-319].

There are several questions concerning Cohen's position that might usefully be posed at this juncture. For example, how could one establish that the rational competence of all adult humans is equivalent? Cohen's argument, technically, can establish at best that full rational competence is possessed by those individuals whose intuitions are actually invoked in the evaluation of proposed normative theories. Typically, the relevant intuitions are those of the theorists themselves, e.g. logicians, economists, decision theorists, mathematicians, et cetera. Upon the discovery of conflicting intuitions, a normative theorist would attempt to dissuade those with such intuitions from their view. Failing reindoctrination, the normative theorist's only alternative is to declare that those with recalcitrant intuitions are members of a distinct "rationality group." The existence of recalcitrant intuitions would spell the demise of the idea of genuine normative theories for, as Stich observes, there would exist no criteria for the evaluation of the merits of the respective rationality groups [Stich, 1981b, pp. 353-354]. Here, the analogy between rational competence and linguistic competence is strained. For, in the case of linguistic theories the existence of different linguistic groups is no threat to the descriptive claim made by any theory of linguistic competence. The existence of different rationality groups, however, would present a threat to the normative claim made by a theory of rational competence. If there exist different rationality groups, the concept of a normative theory of rationality must be rejected in favor of a set of descriptive theories that characterize the competences of different groups. Although there is reason to believe, with Quine, that the idea of discovering a different logic in an alien group is unintelligible, when formulations of normative theories for new domains are at stake it is not unusual to observe theorists eschewing their peers' "intuitions" [see Quine, 1960, pp. 57-60]. This is no demonstrative refutation of Cohen, but a request for clarification. What empirical content is contained in the view that you must "accept the inherent rationality of your fellow adults" [p. 321], whereby Cohen means all normal adults?

Shifting our attention to another issue, Cohen's utilization of the competence-performance distinction enables him to characterize agents as possessing optimal systems for deductive inference and for decision making while allowing for suboptimalization in performance. Any departures from the canons of rationality are to be charged to performance factors, e.g. shifts of attention, restrictions on memory size, and restrictions on processing time. This issue is complicated, however, by the fact that under the influence of such performance factors suboptimalization or "satisficing" is often the pragmatically necessary procedure as H. Simon has insightfully shown [Simon, 1969, pp. 64-65]. Given a technique that will yield an optimal decision but consume so much time that the decision is no longer valuable when produced, a procedure that will yield a less than optimal decision within the allotted time is to be preferred. Thus, performance that departs from idealized normative principles should not always be viewed as merely a case of malfunction. This is, however, the way in which Cohen tends to view performance:

> accounts of actual performance under different conditions are to be obtained by experiment and observation; and hypotheses about the structure and operation of human information-processing mechanisms must then be tested against the facts of competence and performance that it is their task to explain. The structure or design of such a mechanism must account for the relevant competence, but its operation must be subject to various causes of malfunction that will account for flaws found in actual performance. [Cohen, 1981, p. 322]

Though Cohen is able to allow for systematic departures from rationality by citing the mechanisms responsible for performance error, this seems a slightly jaundiced way to view typical patterns of human reasoning -- as mistakes that an otherwise faultless system is doomed to make because of limitations built into the machinery in its servitude. Dennett, too, appears to adopt this view of processes that depart from idealized norms of rationality [1981b, p. 66].

The issue we are confronted with here is quite fundamental. Where we find patterns of reasoning that suboptimalize, do we take our discovery to illuminate typical errors or the definitive patterns of human thought? Given that people have the capacity to acquire rules for optimalizing their decisions, should we take it that all acquirable rules actually structure native competence, but are frustrated in their attempts to guide behaviour through the intervention of performance factors? Or, do we attribute to *de facto* rational competence rules for suboptimalizing?

Cohen would likely take such an issue to be irrelevant to his central claim. He would likely maintain that if we are forced to consider important aspects of the structure of typical patterns of reasoning as determined largely by performance errors, then so be it: any revision of our intuitions in this matter is a small price to pay for the discovery that the rationality of humans, in general, is completely unassailable. Given that Cohen's position will, no doubt, be retained in the face of such questions, how might Dennett's analysis of intentional explanation find support in Cohen's view of human rationality? In one approving reference to Cohen's line of argumentation, Dennett suggests that there is a case to be made for the attribution of full rationality to people: "It is at least not obvious that there are any cases of systematically irrational behavior or thinking" [Dennett, 1981b, p. 66]. Further, Cohen's insistence on a competence-performance distinction in the domain of rationality may have a much more direct relation to Dennett's view of intentional explanation. Dennett holds that the assumption that the agents of behavior are rational is the definitive constraint on intentional explanation. Liberally paraphrasing, explanations of behavior generated by the intentional stance postulate a system that instantiates the canons of rationality. If this interpretation captures an aspect of the intentional stance, then it might appear that Cohen's view combines nicely with Dennett's view. For, we might well be inclined to agree that the system of rational competence plays a crucial role in producing behavioral output and, thus, that there exists a system which instantiates the canons of rationality. Under these circumstances, explanations of behavior would always, ideally, make reference to the system which embodies human rationality. All (intentional) explanation of behavior would then involve reference to a fully rational subsystem. So much, at the present time, for the manner in which Dennett might utilize Cohen's view.

Notice, first, that if Dennett embraces an argument for the existence of a system of rational competence underlying the production of behavior, he would likely need to abjure the view that the appeal to rationality is of merely heuristic or instrumental significance. If such a genuinely rational system crucially figures in the etiology of behavior, then the rationality assumption would appear to be literally descriptive of a subsystem implicated in the production of behavior. But, more importantly, the appeal to a competence theory of rationality cannot buttress the construal of intentional explanation that Dennett offers. Recall that, for Dennett, intentional explanations are given by justifications, i.e. by reference to reasons relative to which the behavior to be explained is rational. We might accept the view that rational competence underlies all performance, but then an agent's competence is not fully realized in his performance. Given this situation, tokens of behavior will be produced by processes that, fully described in intentional terms, do not provide justifications for those tokens.

That is, tokens of behavior will sometimes be produced by processes in which semantically interpretable mental states occur, where -- in turn -- those semantically interpretable mental states constitute the agent's reasons for behaving as he does, but where the agent's reasons do not justify his behavior. If it were the case that the agent's performance matched his competence, then an agent's reasons would always justify the agent's behavior. But, since we know that in many cases performance falls short of competence, we know that an agent's reasons will in those cases fall short of a justification for his behavior. The conclusion must seem slightly ironic from Dennett's perspective: The appeal to a competence-performance distinction may serve to guarantee the rationality of human agents, but it guarantees at the same time that some actual tokens of behavior are the product of less than rational processes. For this set of behavioral tokens, there are no justifying explanations, no intentional stance accounts, to be given. For Dennett, where there is no intentional stance account to be given there can be no explanation in terms of beliefs and desires. Yet, a great many tokens of behavior done in accordance with beliefs and desires will likely not receive complete justification since performance factors are so very pervasive in human reasoning and judgement.

## REJECTION OF THE RATIONALITY CONSTRAINT

As we have noted, intentional explanation under the rationality constraint can be viewed as explanation by justification. This construal of intentional explanation has the adverse consequence of unduly restricting the scope of intentional explanation. For, while the theorist can treat positively any system as if it were rational, to do so is often to engage in the manufacture of unfalsifiable, empirically empty, predictions. A case in point: Dennett considers the attribution of propositional attitudes to a lowly lectern, guided only by an assumption of rationality: "it seems the lectern here can be construed as an intentional system, fully rational, and believing that it is currently located at the center of the civilized world...and desiring above all else to remain at that center" [1979]. The obvious prediction forthcoming from these attributions is that the lectern will "stay put." Though generating such predictions for lectern behavior is not a practice that Dennett endorses, it is not a practice that he is in a position to rule out in any principled way. Thus, intentional explanation under the rationality constraint might initially appear to enjoy an unrestricted range of application. This is an appearance, however, that is easily dispelled.

The intentional stance cannot accommodate any cognitive process that falls short of providing a justification of the behavior in which it eventuates. There will exist such non-justifying processes if, as most would agree, human behavior and human decision making is sometimes less than fully rational. One might attempt to expand the scope of intentional explanation under the rationality constraint by attenuating excessively severe and demanding criteria for the rational justification of behavior. This tactic, however, is inadequate for it not only obscures the role of normative constraints on intentional stance accounts of behavior, it fails to alleviate the fundamental difficulty. There will always exist a class of tokens of behavior that falls outside the class of rationally justifiable tokens of behavior. Otherwise, so-called "rationality constraints" will impose absolutely no effective constraints on predictions and explanations of behavior. That is, if criteria for rationality are attenuated sufficiently so that all behavior is considered rational, the attenuated criteria will not serve to constrain prediction or explanation. Thus, any genuine constraints on explanation imposed by criteria for rationality will rule some tokens of behavior unamenable to an intentional stance account.

Dennett's account of intentional explanation is confronted with the following problem: The rationality condition limits the application of the intentional stance to rationally justifiable tokens of behavior, but some unjustifiable tokens of behavior require intentional explanation. Hence, intentional explanation is not to be thought of as conditioned by a rationality assumption and not to be construed on the intentional stance model. Demonstrating the damage that this argument does to Dennett's construal of intentional explanation requires that we exhibit instances of less than rational behavior which require, nevertheless, intentional explanation. Recall that such cases have been presented above: Risk-aversive and risk-seeking behavior is less than rational according to a certain normative theory, but such tendencies constitute the "norm" in human behavior nonetheless.

At this juncture the prospect of a fruitless and inconclusive series of arguments arises. Against the view that, for example, risk-seeking behavior is less than rational, the advocates of rationality constraints might offer attenuated criteria for rationality according to which risk-seeking behavior is to be considered rational. With the attenuated criteria in hand, the opponents of rationality constraints might, in turn, point to other tendencies in human judgement that depart from those criteria, e.g. the tendency of individuals to evaluate objectively equivalent options as constituting non-equivalent alternatives [see Kahneman and Tversky, 1982, p. 166]. Thereupon, the response from the advocates of rationality constraints will likely be to attenuate the criteria for rationality even further. Responding to this manoeuver, the opponents of rationality constraints will point to still other tendencies exhibited by human judgement that depart from the newly attenuated criteria. The results obtained from such a continuing series of arguments and counterarguments are bound to be less than incisive, less than illuminating. But, this is one battle which we are in a position to forestall.

One way to avoid the need to grossly attenuate criteria for rationality is by appeal to a competence-performance distinction. Utilizing the distinction allows the theorist to attribute a substantial rationality to human subjects while acknowledging the degenerate character of performance, i.e. the actual processes responsible for the production of behavior. Though the competence-performance distinction is extremely attractive in application to the domain of human rationality, the appeal to the distinction incurs heavy costs for advocates of rationality constraints. As we have seen, once the distinction is made the theorist must recognize that the processes which eventuate in tokens of behavior are sometimes less than rational, less than perfect executions of competence. Hence, those processes which depart from competence in some way, and the behavior they produce, will not be amenable to an intentional stance account. Nevertheless, the intervention of typical performance factors, e.g. limitations on memory, limitations on processing time, and shifts of attention, does not militate against an intentional interpretation of cognitive processes. That is, an interpretation that assigns beliefs and desires to an agent is not ruled out simply because the formation of beliefs and desires in the typical agent is constrained by performance factors. Indeed, part of the explanation of why an agent possesses the beliefs and desires that he does will be provided by appeal to a description of performance factors. But, this means that an intentional interpretation of an agent's mental states will be, in part, governed by the principle that belief and desire formation procedures depart from the requirements of rationality.

It looks, then, as if it is a serious error to view intentional interpretation and explanation as conditioned by a rationality assumption. Moreover, the inability of intentional explanation under the rationality condition to accommodate a wide range of behavioral and judgemental phenomena is symptomatic of a flawed analysis of the concept of intentionality. How is it that this problem for the intentional stance arises? Consider a fundamental question concerning the rationality condition. What are the preconditions for viewing a system as rational? Some might interpret Dennett as holding that there are no such preconditions since anything can be treated as if it were rational. Such an interpretation would be disingenuous. Though nothing in particular need be true of a system which one merely assumes to be rational, making a rationality assumption carries certain principled commitments. There are, of course, many possible interpretations of the principled commitments entailed by the rationality assumption. Here is an interpretation of three more or less autonomous but compatible commitments: First, one may attribute rationality to the course of action adopted by an agent on the view that the rationality of overt behavior is a function of its perceived instrumental value for securing fixed goals: To do so is to commit oneself to the view that rational behavior is goal-directed or purposeful. Second, one may attribute rationality to beliefs derived from other beliefs on the view that the rationality of derived beliefs is a function of the deductive validity of the inferential procedures via which they are derived: To do so is to commit oneself to the view that there are specifiable logical relations between semantically endowed mental states. Third, one may attribute rationality to perceptual beliefs on the view that the rationality of perceptual beliefs is a function of the epistemic reliability of an inductive practice for the selection and testing of hypotheses. To do so is to commit oneself to the view that there are relations between environmental circumstances and the contents of certain mental states of an epistemically justifiable kind.

The point is that to assume that rationality can be legitimately attributed to a system is to assume certain commitments. What, after all, is attributed when "rationality" is predicated of a system? Given that there are rational processes in systems that we wish to understand, the three views sketched above generate salutary commitments for a theory of rationality. The three views are connected by a single overriding commitment to the intentionality of rational systems, i.e. to the capacity of some systems to internalize semantically endowed representations. This commitment is unavoidable. When we assume that a human agent, or some other system, is rational we make a concurrent presupposition to the effect that the agent formulates intentional attitudes or content bearing mental states. If attributions of rationality presuppose the intentionality of those systems considered rational, then explanations of behavior that appeal to rationality are founded upon the intentionality of behavior emitting systems. Since one cannot account for the intentionality of a system by reference to a concept that itself presupposes intentionality, there can be no hope of founding intentional explanation upon a rationality assumption.

To see that intentionality is a presupposition of rationality, and of irrationality as well, consider what is perhaps the least controversial of the three domains of rationality attribution mentioned above, i.e. that in which rationality is predicated of beliefs generated by operations upon stock beliefs. Here, our concern is to determine if the relations between elements of an agent's belief set are rational. The normative theory that sets the standards for rational belief systems will impose certain constraints. Such a normative theory might, for example, require the deductive closure of belief sets. Since this constraint is highly implausible, consider a much less stringent constraint: A belief system is rational only if the procedures whereby new beliefs are derived from old beliefs are <u>truth-preserving</u>. In order to assess the rationality of beliefs in accordance with this constraint, we must treat beliefs as bearers of truth values -- or as it is sometimes said as truth-valuable. Now, if beliefs have truth values, they have truth conditions. If beliefs have truth conditions, they have semantic content. Indeed, the content of a belief, to a first approximation, is given by its truth conditions. The point to be drawn from these familiar observations is that evaluating the validity of an inferential practice, and hence the rationality of a derived belief, requires viewing beliefs as endowed with content: no content, no truth conditions; no truth conditions, no truth value; no truth value, no basis for the assessment of the logical validity of the derivations performed.

Rationality is exemplified by a belief system when the relations between beliefs constitute a model of an appropriate normative theory. We may think of a theory of rationality in this domain as a set of restrictions on relations between the semantic contents of beliefs. A system must be attributed _irrationality_, in turn, when the _de facto_ relations between its beliefs violate, to some significant degree, the restrictions imposed by the normative theory. On this approach to questions of rationality, the extent to which an individual, or an artifact of artificial intelligence research, is rational is an empirical matter and a matter of degree, i.e. how well does the actual, empirical system fit the normative theory? An actual system will more or less conform to the normative theory at any given time and may conform to the theory better at one time than at another. Notice that there are, roughly, two distinct ways in which a system can fail to conform to a normative model. One form of nonconformity is arationality. That is, a system might simply fail to provide a natural domain in which a normative theory could be satisfied. The other form of nonconformity is irrationality. That is, a system might implement procedures dependent upon the content of internal states, but produce transformations of content not allowed by a normative theory. In order to meaningfully attribute irrationality to a system, then, it is necessary that certain of its states be viewed as semantically endowed. For, only violations of principles that restrict relations between the contents of mental states, or violations of principles that restrict relations between mental states and instances of behavior, can result in irrationality.

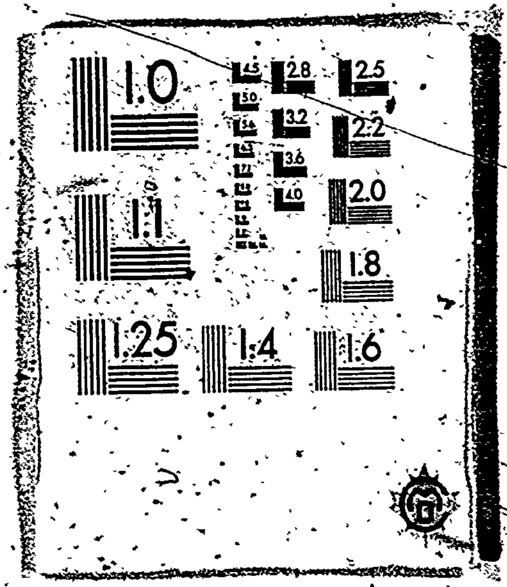
The idea that principles of rationality which restrict relations between mental states and instances of behavior might be violated merits a comment. According to the view of intentional explanation urged against Dennett here, the type to which a token of behavior belongs is a function of the belief - desire context of the agent who produces the behavior. This view does not presuppose that the behavior an agent issues need be rational relative to the agent's belief-desire context. In fact, there are instances of behavior whose role or significance in the life of an agent can only be captured by acknowledging that such behaviors are irrational. For example, consider an agent who desires G and believes that M is a prudential means for securing G. It is by no means necessary, either logically or nomologically, that the agent's ensuing behavior need, in every case, be characterized as "an attempt to do M". A particular instance of behavior produced in this belief-desire context may be best characterized as a failure to do what is rational by, if you will, the agent's own lights. In this connection, there is a long standing tradition in philosophy, beginning roughly with Aristotle, that acknowledges Akrasia, incontinence, or "weakness of the will" as an intentional phenomenon. The inaction of an agent in such a case, his forebearance of "an attempt to do M", is a rational failing on his part. For, the agent's behavior is not determined by what he sincerely avows to be the rational sources for behavior in the context in which he believes himself to be fixed. On the other hand, if we assume, as Dennett insists we must, that intentionally interpretable behavior is always rational relative to the belief-desire context of its agent, then we must deny that it is logically or nomologically possible for an agent to believe that M is a prudential means for securing G and, yet, fail to attempt to do M. Indeed, this is just the tack that Dennett takes on this question. Dennett invents a propositional attitude called an "opinion" which is like a belief with one important difference: Opinions don't, or needn't, have a significant role in the determination of behavior [see Dennett, 1978, p. 307]. Dennett relegates all causes of Akrasia to the class of behaviors incompatible with one's opinions but compatible with one's behavior determining beliefs in order to save the hypothesis that an agent's beliefs and desires are always rationally related to an agent's behavior. Not only does this appear to be an *ad hoc* adjustment, it is also irrelevant to the issue of the conceptual relation between intentional concepts and the rationality condition. For, to maintain that opinions, unlike beliefs, need not be rationally related to instances of behavior, or to an agent's behavior determining beliefs, is to allow that there are intentional concepts, e.g. the concept of an opinion, that do not presuppose rationality constraints -- which is, of course, the view that we have been concerned to establish. Thus, on the most natural analysis, it is logically possible that certain instances of behavior may be irrational in relation to an agent's beliefs and desires, but in order to characterize such instances of behavior fully we must advert to the contents of the agent's beliefs and desires.

The point here is that the very same conditions that provide the foundation for attributions of rationality also provide the foundation for attributions of irrationality. The relevant conditions are those constitutive of a system's intentionality. Rationality and irrationality alike are properties exemplified only by systems which internalize content bearing states. Now, since intentionality is a necessary presupposition for the attribution of rationality, when we have determined something to be a rational system, we can be assured that it is, a fortiori, an intentional system. That is, if something is rational, then all the preconditions for rationality will necessarily be satisfied. Viewing an agent as rational thus forces us to view the agent as internalizing semantically endowed mental states. In a restricted sense, Dennett is correct when he observes that the rationality assumption "generates...an intentional interpretation of us as believers and desirers" [1981a; p. 44]. But, the observation is correct only because the theorist is committed to the intentionality of a system if attributions of rationality or of irrationality are meant to be taken literally. This is simply a case in which identifying a superstructure commits us to the foundation upon which it is erected. Rationality is one such superstructure, irrationality is another; both are erected upon a foundation provided by relations between semantically endowed mental states.

Viewing rationality constraints as necessary restrictions on intentional explanation prejudices our theoretical assessments of human rationality, both in competence and performance. Rationality constraints limit the application of intentional explanation to logically proprietary transitions among mental states and to rationally justifiable tokens of behavior. Imposition of rationality constraints simply presupposes the semantic contentfulness of mental states while excessively constraining semantic interpretation. Thus, intentional explanation under the rationality assumption is rendered incapable of accounting for certain essentially intentional processes, e.g. those that realize irrationality as well as those that simply depart in one way or another from competence but which involve beliefs and desires.

To close on a more positive theme, notice that our analysis of rationality provides an argument for the inclusion of a semantic component in psychological theory in addition to the argument constructed in the first chapter. If a task of psychology is the evaluation of the nature and extent of human rationality, then psychological theory must include a semantic component. For rationality and irrationality are attributable only if a semantic interpretation of mental states is provided.

# 3



## FOOTNOTES

[1] Dennett discusses his view of intentional psychology in many papers. It may be useful to note that the dates assigned to Dennett's works in the text are not indicative of the conceptual development of Dennett's ideas. For example, "Intentional Systems" originally appeared in 1970, but I have used the pagination of Dennett's Brainstorms [1978]. The work "Three Kinds of Intentional Psychology" dates from May 1979, but has only just been made available in print [1981a]. Citations from Dennett's "True Believers: The Intentional Stance and Why It Works" are taken from the text of a lecture given by Dennett November 30, 1979, which, we are promised, is to appear in a volume of the 1979 Herbert Spencer Lectures. Thus [1981a] is an earlier paper than [1979].

[2] There is another argument that Dennett has employed in an attempt to illustrate the vacuity of intentional explanation. He claims that (a) such explanations presuppose rationality and (b) rationality is a property of cognitive systems that psychology must explain, but this means that (c) intentional explanations simply presuppose what psychology must explain [1978, pp. 12, 16]. Hence, intentional explanation makes for vacuous psychology. This argument has not recurred in more recent of Dennett's works and is, in any case, slightly curious. Not the least of its problem is that premise (b) seems to contain an implicit denial of the proposition that 'humans fail to satisfy the rationality condition'. For, surely, no theory is obligated to explain a property, e.g. rationality, that is not exemplified by the subject matter of the theory. Additionally, it is not an argument against a theory to show that it makes certain presuppositions, all theories do. Given a theory that explains behavior under the working assumption, that there are rational connections between internal states, one may then wish to explain how it is that human agents are rational. Notice that the task of explaining rationality is one which psychology assumes in addition to explaining behavior. What Dennett might usefully conclude is that intentional explanation makes for incomplete psychology -- as opposed to vacuous psychology -- unless it is augmented with a theory of rationality.

3. Unless one can show that all interesting explanatory generalizations can be recast in the idiom of physical theory, there is not much point in embracing the dogma that physical theory is completely universal. It is enough to acknowledge the universality of physical theory in the properly restricted sense that no physically described phenomenon violates physical law. Yet physics does not appear to possess the resources to make certain distinctions, between types of overt behavior for example, that we would like to make. What is surprising is that Dennett abandons psychological reduction [1978, p. xvii] and yet embraces the universality dogma in what appears to be an unrestricted sense, but see footnote #5 and accompanying text.

4. The specification of the beliefs that an agent "ought to have" is relative to a specification of the agent's goals. Given a goal or set of goals we can attribute beliefs on a normative basis: The system desires '$G_1$, $G_2$, $G_3$'; information '$\underline{B}_1$, $\underline{B}_2$, $\underline{B}_3$...$\underline{B}_N$ is information that is crucial to the fulfillment of '$\underline{G}_1$, $\underline{G}_2$, $\underline{G}_3$; therefore, '$\underline{B}_1$, $\underline{B}_2$, $\underline{B}_3$...$\underline{B}_N$' are beliefs that the agent *ought* to have. But, it is extremely hard to see how there could be a normative basis for the attribution of basic desires. Dennett would likely agree, since he does revert to empirical considerations when attributing basic desires. This is, of course, as it should be, but Dennett is slightly misleading when he appears to claim that we could somehow attribute basic goals normatively. In any case, these considerations show that the intentional stance is not fully normative, it contains empirically considerations *essentially*.

5. The indeterminacy in specifications of design is generated by an inescapable trade-off between the functions assigned to a natural system and idealizations of the system's structure: Different function assignments "reveal" different structures. This is not simply a case of the underdetermination of theory by the available data, in my view, for it is not so much a question of the compatibility of competing theories with all available data, as it is a question of the testability of any interpretation of a system's purpose. One might observe that a system keeps accurate time, but to claim that "timekeeping is the system's purpose" adds no empirical content to the theory. Consider the moon in its orbit; it is highly

regular and thus might be thought of as a time piece. But clearly, there can be no question of testing the "hypothesis" that the moon's purpose is to keep time. Plato, incidentally, seems to have thought of heavenly bodies as endowed with such purposes in the Timeaus.

6. Dennett makes this concession in the same lecture in which he claims that although the intentional stance "gives us predictive power we can get by no other method,...this is no difference in nature, but merely a difference that reflects upon our limited capacities as scientists" [1979]. A perfectly proper observation, however, would seem to be that a scientific community which failed to identify the genuine patterns in behavioral phenomena would be severely epistemically limited and, further, that behavioral phenomena constitute natural kinds distinguishable from the natural kinds amenable to physicalistic explanation. It would, hence, appear difficult to provide an interpretation of Dennett's views on (i) the empirical adequacy of the physical stance, and on (ii) the authenticity of patterns of behavior capturable by intentional interpretation, that would render them fully consistent with each other. The salient point, in any case, is that the ability to take cognizance of intentional patterns in behavioral phenomena in no way seems to reflect a limitation of our epistemic resources.

7. Fodor offers another argument against constraining intentional explanation by criteria for rationality. Though offered, apparently, as an independent argument, it seems to simply be a special case of the observation that belief sets are not deductively closed. In any case this, roughly, is the argument: Subjects perform problem-solving tasks with greater facility when given information in an affirmative mode, i.e. they perform better with the information that 'P' is true than with the information that '-P' is false. Fodor reasons that to state this fact we must advert to the opacity of representations. Thus, Fodor concludes that "intentional explanation works...where precisely lapses from optimal rationality...are at issue" [1981, pp. 107-109]. Notice that Fodor uses 'intentional' to mean, variously, "opaque" and "content respecting." That aside, this argument has the same general form as the argument considered in the text: Beliefs are not closed under double negation,

hence not deductively closed; optimal rationality is incompatible with such a belief set, thus the rationality condition is not a legitimate constraint on theories of behavior that refer to beliefs.

# CHAPTER III

# FUNCTIONAL SPECIFICATIONS AND SEMANTIC SPECIFICATIONS OF MENTAL STATES

## THE CONCEPT OF THE FUNCTIONAL ROLES OF MENTAL STATES

Consider the concept of the functional role of a mental state. We might say, truly if somewhat uninformatively, that the functional role of a mental state type, M, consists in the sum of the functional roles of tokens of M. This characterization of the concept of the functional role of a state type will be unappealing to many functionalists, for it entails only that given a way to identify the tokens of a type, the functional role of the type can be determined. Functionalism demands more from the concept of functional role. In particular, on one interpretation, functionalism is the position that the essential nature of a token mental state, i.e. the kind of entity that the token state is, is determined by a set of facts about the functional role of the token. Functionalism, thus, envisages the formulation of criteria for "sameness of functional role", applicable to token mental states, according to which the type-identity of mental states can be determined. Though functionalists may differ regarding the details of functional specifications of mental states, there is a consensus to the effect that the idea of the functional roles of mental states can be assimilated to a conception of the causal roles of mental states. Fodor is explicit about this point and others noted above:

> The intuition that underlies functionalism is that what determines which kind a mental particular belongs to is its causal role in the life of the organism. Functional individuation is individuation in respect of aspects of causal role; for purposes of psychological theory construction only its causes and effects are to count in determining which kind a mental particular belongs to. [Fodor, 1981, p. 11]

If mental states are to be individuated by reference to their causal-functional roles, then the equivalence, or type-identity, of the causal-functional roles of mental tokens must be a necessary and sufficient condition for the type-identity of mental tokens. It will be argued at some length in a following section that functional criteria for the type-identity of mental states are unavailable. In this section and the next we will set the stage with some helpful preliminaries.

What is at stake in arguments over the possibility of functional specifications of mental states is not merely one proposal, among many, for the individuation of mental states, but a fundamental conception of the nature of mental states. The idea, part and parcel of functionalism, is that mental state types, e.g. sensations, pains, beliefs, and desires, can be exhaustively defined in terms of their causal roles which constitute, in a suitably broad sense, the input-output functions that instances of such states compute. Presumably, functional criteria may be adequate to distinguish instances of pains from instances of beliefs, but it is not at all obvious that exclusively causal-functional resources are sufficient to distinguish, for example, a token of the belief that "it's raining" from a token of the belief that "pluvial atmospheric conditions obtain". The ability of functional criteria to distinguish between propositional attitudes of the same family, e.g. between beliefs or desires, has gone largely unexamined in the functionalist literature. Indeed, one philosopher sympathetic to functionalism, S. Shoemaker, has recently claimed that "no functionalist would maintain that each different belief and each different want must be defined separately; in the case of belief, for example, the functionalist will want a definition of '_S_ believes that _P_' which holds _for all values of P_" [Shoemaker, 1981, p. 118, emphasis mine]. Shoemaker's claim is highly contentious. For, if functionalism cannot individuate belief types, this failure constitutes a debilitating limitation on its scope and generality.

Not all functionalists, however, are prepared to forfeit the possibility of the causal-functional individuation of belief types. Fodor notes that though functionalism is troubled by qualitative phenomena like pain, it is much more plausible in application to propositional attitudes:

> Functionalism applies only to kinds whose defining properties are relational. And while it is arguable that what makes a belief--or other propositional attitude--the belief that it is is the pattern of (e.g. inferential) relations that it enters into, many philosophers (I am among them) find it hard to believe that it is <u>relational</u> properties that make a sensation a pain rather than an itch....It makes no sense to speak of my belief being different from yours despite the identity of their inferential (etc.) roles. This asymmetry [between qualitative states and propositional attitudes] is--plausibly--attributable precisely to the relational character of beliefs. [Fodor, 1981, pp. 16-17]

In Fodor's estimation, propositional attitudes provide the domain, <u>par excellence</u>, within which functional individuation shows its greatest promise. If Fodor is correct, functional individuation is only suited to those entities whose essential natures are determined by the set of relations into which they enter. Beliefs, Fodor suggests, may be just such entities. Notice that one set of relations, relevant to the individuation of beliefs, to which Fodor refers are the inferential relations between beliefs. The idea is that it is at least partially definitive of the belief that <u>P</u> that some particular belief that <u>Q</u> is <u>inferable</u> from <u>P</u>. But, functionalism is committed to a kind of causal individuation of beliefs and, thus, the "inferability" relation must be amenable to some sort of causal construal if it is to enjoy a role in the specification of belief types. The functionalist might, for example, refer to a belief's productivity, i.e. its capacity to cause or bring about new beliefs in collaboration with other beliefs, in order to provide a causal analogue of the inferability relation.) In any case, on the causal-functional approach to the individuation of belief types, there are three kinds of causal relations which might be relevant to specifications of the functional roles of beliefs: Beliefs are related to other propositional attitudes; beliefs are related, in some cases, to environmental stimuli; and beliefs are related, in some cases, to instances of overt behavior. But, are all or only some of these relations relevant to specifications of the functional roles of mental states? Actually, this question concerns only one of two parameters of variance which must be fixed if the credentials of the functionalist approach to the individuation of mental states are to be examined. One parameter concerns what we might think of as the breadth of specifications of the causal roles of mental states, i.e. do such specifications refer to external states or only to internal states? The other parameter along which functional specifications of token states may vary has to do with the level of abstraction at which the causes and effects of mental states are described. We will briefly consider each parameter, examining the idea of a proprietary level of abstraction first.

It is important to recognize that the idea that functional individuation is a species of causal individuation does not, by itself, fix any particular level as the one appropriate for the specification of mental state types. As Fodor takes pains to acknowledge, the causes and effects of mental state can be described in various theoretical vocabularies. One might, for example, give neurological descriptions of the causes and effects of mental states. But, to do so would be to collapse "the intended distinction between functionalism and type physicalism" [Fodor, 1981, p. 11]. Fodor's point is that the individuation of mental states in terms of sets of input-output pairs where those inputs and outputs are described in neurological terms is just a variant of type physicalism: According to such an approach, every mental type is identified with some functionally specified neural type. The only difference from a more traditional type physicalism, here, is that the neural states with which mental types are identified are individuated by functional descriptions, rather than by descriptions of certain _intrinsic_ properties of the states, e.g. 'c-fiber firings = pain'. The salient point is that the notion of functional or causal individuation, _per se_, does not rule out neurological, or otherwise physicalistic, descriptions of the causes and effects of mental states.

Functional individuation is individuation in terms of causes and effects, but there are many possible conceptions and descriptions of the causes and effects of internal states.

Descriptions of the causes and effects of mental states cast in neurological terms are ruled out, however, by the considerations that motivate a certain kind of functionalism. This brand of functionalism, most closely identified with Putnam and Fodor, forbids exclusively neurological descriptions of mental states, or of their causes and effects, since systematic use of such descriptions would restrict the attribution of mental state types, e.g. pain and the belief that P, to systems of the same neurological type. This tenet of functionalism is, of course, highly tenable and widely accepted since many different types of systems, e.g. neural-man, neural-Martian, and perhaps electronic, could presumably realize pain or the belief that P. But, the bold functionalist assertion to the effect that mental states, of physically similar or dissimilar systems, are of the same type just in case they "have the same function" is of little help in itself. The problem is that functional roles can be specified at various levels--some of which are of little or no psychological significance.

When we attempt to fix the level at which psychologically relevant functional descriptions of mental states are given, we are pulled in two directions. On the one hand, functional analysis is a species of causal analysis and must advert only to realizable, causally efficacious, properties of mental processes. On the other hand, if psychological theory is to formulate generalizations explanatory of intentional behavioral types--as we observed in Chapter I, such types as arithmetic problem-solving behavior are in this class--then specifications of functional roles must be given on a level at which interpretations of mental states are appropriate. Although these two requirements may appear to point ~~in~~ different directions, it is an important merit of the computational theory of cognitive processes that it proposes a way to satisfy both requirements. According to the computational approach mental operations are determined by the formal, syntactic, or in any case the structural properties of mental symbols. Yet, on this view the formal properties of mental symbols are presumed to be in a correspondence of some kind to the semantic properties of mental symbols. The formal properties of mental symbols are referred to in order to satisfy the requirement that the properties mentioned in functional specifications of mental states are realizable, causally relevant properties, and the semantic endowments of mental symbols are referred to in order to satisfy the requirement that functional specifications of mental states are given at a level at which interpretations of mental states are appropriate.

The merits of the computational view are considerable, but the idea that mental states or mental processes must be formally specified--which is putatively entailed by Fodor's version of the formality condition--can easily mislead the unwary. As we will see in more detail in the next section, formal specifications of mental state can suffer from the same shortcomings that afflict neurological specifications of mental states. At this juncture, it is sufficient to observe that a defining desideratum of functionalism is that specifications of mental states are to be given at a level of abstraction that provides for the possibility of type-identities between token states in neurologically and/or formally distinguishable systems.

The other parameter along which specifications of causal-functional roles can vary has to do with what we might term the breadth of specifications of such roles. Suppose that we have fixed some level of abstraction, a formal level for example, for the specification of causal roles. It remains for the theorist to determine whether those causes and effects temporally "contiguous" with mental states are to be referred to in specifications of their functional roles. Intuitively, the fact that mental states have a role in the production of behavior suggests that tokens of overt behavior are among the outputs of mental states. And, since mental states are sometimes formulated under the pressure of environmental stimuli, it appears that such stimuli are among the inputs to mental states. These considerations suggest what will be termed a wide construal of the functional roles of mental states--to be contrasted with what will be termed a narrow construal. A wide construal of the functional role of a mental state M is given just in case among the causes and effects referred to in specifications of M's functional role there are states external to the systems which internalize M. A narrow construal of the functional role of a mental state M is given just in case all the causes and effects referred to in specifications of M's functional role are states internal to the systems which internalize M. Notice that this use of the terms 'wide' and 'narrow' is analogous to Putnam's use in defining "psychological states in the wide sense" and "psychological states in the narrow sense". For Putnam, a wide characterization or description of a mental state, e.g. a belief, is one which presupposes the existence of some entity other than the agent of the belief. Any other characterization is narrow [Putnam, 1975, pp. 215-271]. We have said that a wide construal of the functional role of a mental state is one which refers to causes and effects that are other than internal states of the agent and that any other characterization is narrow. Once we distinguish between functional roles in the wide sense and functional roles in the narrow sense, certain options appear. For example, the functional role of a belief $\underline{B}_1$ might be narrowly specified by giving as its causes and effects only other beliefs (narrowly specified). As causes we might specify all those beliefs, and sets of beliefs, of a system which, under its principles of inference, entail $\underline{B}_1$, and as effects all those beliefs which $\underline{B}_1$ entails, again given the system's principles of inference. On this construal, the functional role of a belief is given by a specification of the beliefs to which it is inferentially related. We might think of such narrowly specified functional roles as capturing the "inferential roles" of beliefs. Alternatively, the theorist might associate the belief $\underline{B}_1$ with the outputs of input transducers even though several other mental states, or several mental processes, intervene between the outputs of the transducers and the occurrence of $\underline{B}_1$. Thus there are, in a sense, various degrees of narrowness. All narrow specifications of functional roles refer only to internal states of cognitive systems, but the narrow functional role of a state $\underline{M}$ might be given at one extreme by reference only to those other states with which $\underline{M}$ is immediately connected or, at the other extreme, by reference to states including those that occur at the periphery of the internal psychological processes in which $\underline{M}$ occurs.

Wide specifications of the functional roles of mental states, on the other hand, are consonant with the intuition that the role of mental states is, inter alia, to bring about behavior which affects the environment. When it is said that the function of a mental state is to affect a certain overt action, or when it is said that the function of a mental state is to register certain properties of the environment, wide construals of the functional roles of mental states have been adopted. On one possible wide construal, the functional role of a mental state is to transform the environment, i.e. to bring it about that certain environmental conditions obtain upon the occurrence of certain other environmental conditions. For example, an agent may bring it about that there is an umbrella over his head upon the occurrence of precipitation--doing so might be thought of as an aspect of the wide functional role of the belief "that it's raining".

It appears that both notions of the functional roles of mental states face serious problems. Intuitively, functionalism holds that mental states have a function in the production of behavior. But, limiting theory to narrow specifications of functional roles would seem to entail the view that the functions of mental states are to produce only other internal states. On the other hand, if we limit ourselves to wide specifications of functional roles, many distinct mental states will be indistinguishable. This is a rather crucial point and merits close attention.

There are, no doubt, a variety of mental processes that can mediate a given widely construed input-output pair. There are many different mental processes that might serve to keep one indoors contingent upon the occurrence of rain: An agent may remain indoors because she believes that "rain is wet" and desires to stay dry or because she believes "rain is often accompanied by lightning" and fears electricity. Each of these beliefs may contribute to the determination of many other input-output relations. For example, the latter belief may cause one, as it did Ben Franklin, to fly kites during rainstorms. And, thus, one might suppose that the beliefs can be distinguished by a difference in what we might call their "maximally specified wide functional roles", i.e. in terms of descriptions of all the possible, widely specified, input-output pairs that tokens of a given state may mediate. We may think of maximally specified wide functional roles as given by infinite sets of input-output pairs: $\{(I_1,O_1), (I_2,O_2), (I_3,O_3)\ldots\}$. Now, for many pairs of beliefs, P, Q, there will be indeterminately many points at which the wide functional role of the belief that P and the belief that Q coalesce--indeterminately many I-O pairs common to the maximally specified functional roles of each belief. How is the theorist to determine for one of these shared pairs, $(I_2,O_2)$, whether a token state, m, which mediates $(I_2,O_2)$ on a particular occasion, is an instance of the belief that P or an instance of the belief that *Q*? According to the wide functional role theory, the belief that *P* is distinguished from the belief that *Q* because there exists a pair $(I_3,O_3)$ which one, but not the other, mediates. Let us suppose that the belief that *P*, but not the belief that *Q*, has as an aspect of its maximal functional role the pair $(I_3,O_3)$. The token *m*, then, which occurs within an instance of $(I_2,O_2)$, is a member of the type 'belief that *P*' if and only if it belongs to the type whose tokens can occur within instances of $(I_3,O_3)$. But, determining whether these conditions are satisfied requires the *reidentification of m-like tokens*. That is, in order to determine if a token state, *m*, belongs to the type 'belief that *P*', we must determine whether tokens type-identical to *m* can mediate a certain input-output pair, but this presupposes that there are type-identity criteria for token states independent of their wide functional roles. It presupposes, in particular, that there is some mechanism for the identification of occurrences of tokens of a type that does not rely upon reference to the maximally specified wide functional role of the type. Hence, exclusive reference to external input and output states will not provide for the type-identification of beliefs.

Fortunately, there is no reason for the theorist to cleave either to exclusively wide or to exclusively narrow construals of the functional roles of mental states--at least no reason that we have seen thus far. In fact, one form of functionalism, "Ramsey-sentence functionalism" positively demands mixed construals of the functional roles of mental states.[1] We will consider why this is so, very briefly, and then turn to a consideration of an approach to the individuation of mental states constrained by the "formality condition"--an approach that on one interpretation would appear to require exclusively narrow construals of the functional roles of mental states.

Consider the specification of the functional role of a mental state type *M* given by the Ramsey-sentence of a causal-functional theory of *M*. Let us suppose that *M* is a belief with determinate semantic content, the belief that *Q*. The theory *T* of *M's* functional role will specify the causes and effects of *M* in a vocabulary that contains theoretical and/or observational terms. If the theorist cleaves to narrow construals of *M's* functional role, then the vocabulary of *T* will contain no observational component: All internal states, e.g. beliefs and desires, are designated by theoretical terms. The Ramsey sentence for a theory is obtained by replacing every constant designating a theoretical entity with a bound variable. So if *T* is the simple theory that "the belief that *Q* is caused by an inference from any of the internal states, belief that $P_1$... belief that $P_M$, and *Q* causes any of the internal states, belief that $R_1$...belief that $R_N$", then *T* is Ramsified roughly as '$(\exists X_1)\ldots(\exists X_M)(\exists Y)(\exists Z_1)\ldots(\exists Z_N)T(X_1\ldots X_M, Y, Z_1\ldots Z_N)$', where the variable 'Y' replaces 'the belief that Q'. The original theory contained only theoretical terms and, hence, the Ramsey sentence of the theory has *no observational consequences*. Now, one motivation underlying Ramsey-sentence formulations of theories of mental states, perhaps the chief motivation, is to provide a way to give functional construals of mental states in terms of observable events, stimuli and overt movements of the body. Hence, a theory that specifies the narrow functional role of a mental state, such as *T*, will not be acceptable to the Ramsey-sentence functionalist since such a theory provides no way to define theoretical terms in terms of our understanding of observational terms.

## FUNCTIONALISM AND THE FORMAL INDIVIDUATION OF MENTAL STATES

It will be useful to consider, briefly, the possibility of individuating mental states by reference to their formal properties. Actually, if the observations made in Chapter I can be sustained, formal criteria for the type-identity of mental states are not anticipated. But, it is highly instructive to consider the possibility of formal individuation within the context of functionalism. Now, Fodor holds that what he terms the "formality condition" constrains specifications of mental processes [1980, p. 103]. Roughly, the formality condition constrains specifications of mental processes to reference to the formal properties of mental operations, and to the formal properties of the token states to which such operations apply.[2] Either as a consequence of the formality condition or in conjunction with it, Fodor proposes a criterion for the type-identity of mental states that refers only to the formal, or non-semantic, properties of mental states. Since semantic properties are putatively irrelevant to the specification of mental states, semantically equivalent states which are formally distinct are thereby type-distinct, and semantically non-equivalent states which are formally indistinguishable are thereby type-identical [see Fodor, 1980, p. 67]. The proposed necessary and sufficient condition for the type-identity of mental states, then, is formal indistinguishability. Actually, this criterion may be considered only a first approximation, but the crucial point in what follows will be that any purely formal criterion for the type-identity of mental states is bound to be incompatible with the requirements of functional individuation.

Now, if the theorist could assign each token mental state to its type on the basis of an assessment of its formal structure alone, then the formal individuation of mental states would accomplish a kind of functional individuation. Fodor views mental processes as computational processes and computational processes "have access" only to the form, shape, or syntactic structure of the symbols upon which they operate. On this view, mental states enter into mental processes only in virtue of their form. Hence, specifying the form of a mental state is to specify a property of the state which constrains its embeddings in mental processes or in input-output relations. It would appear to follow that mental states are functionally equivalent if and only if they are formally indistinguishable: All functionally relevant distinctions between mental states have their source in formal distinctions between mental states. Fodor holds that this view entails a kind of methodological solipsism--the doctrine, roughly, that specifications of mental states may not advert to conditions or states of affairs external to the agents who internalize the mental states. In these terms, methodological solipsism appears to be equivalent to the view that all specifications of functional roles must be narrow.

Though the intuition underlying the formality condition may be correct--mental operations apply in virtue of non-semantic properties of mental states--the formality condition does not constrain the individuation of mental states in a manner compatible with functionalism. In the first place, as Fodor himself points out, certain formal differences between tokens don't, intuitively, suggest type-differences between the tokens: The belief that 'that is edible' and the belief that 'this is edible' appear to be type-identical when 'this' and 'that' are coreferential [Fodor, 1980, p. 67]. More generally, Fodor has recently allowed that "there could be [as Dennett observes] many, many syntactic types associated with the same propositional attitude" [Fodor, forthcoming]. Thus, the formal distinguishability of token mental states does not entail, by itself, a functional difference between the mental states. Indeed, formally distinguishable mental processes may be functionally equivalent, e.g. there are many algorithms which might be used to compute sums.

Moreover, and this is perhaps the crucial observation, the formal indistinguishability of token mental states does not guarantee their functional equivalence or their type-identity. Here, we might cite a pair of beliefs of different agents of the form 'that is edible' where one belief is about an apple and the other is about an orange. While it is at least an option to insist that eating apples and eating oranges are behaviors of the same type when each is conceptualized merely as "eating something edible", other considerations suggest that this manoeuver will not save the formal approach to the individuation of mental states. For, it is possible that formally indistinguishable mental states have different functional roles in different systems.

In one work, Fodor offers an example--credited to G. Rey--of just this type. Consider a computer that runs different programs on alternate days. One program is designed to simulate, say, the Six Day War, while the other is designed to simulate a chess game: "It's a possible (though, of course, unlikely) accident that these programs should be *indistinguishable when compiled*...so that the internal career of a machine running one program would be identical, step by step, to that of a machine running the other" [Fodor, 1981, p. 207]. Presumably, however, mental states about the Six Day War and mental states about a chess game are not members of the same type, even though they may be realizable by tokens of the same formal type. However, the relevance of this example may be denied since it cites only computational states which have a subject matter as a function of the intentions of the theorist who uses a computer programmed in a certain way for a certain purpose. In principle, the example is little different from the case in which the theorist simply varies his interpretation of a single formally articulated system on alternate days. Nevertheless, if the formal properties of a process do not (fully) determine the processes' interpretation, it should be possible to construct examples in which formally indistinguishable mental states are type-distinct not as a function of the theorist's intentions, but rather as a function of the agent's own intentions.

Let us suppose that there are two gardeners, one hither and one yon, whose duty it is to serve the mineral needs of a patch of grass in their charge. Each has a gardening vocabulary restricted to the terms 'grass', 'green', 'yellow', 'healthy', 'sick', 'apply fertilizer', 'do nothing', and the sentential connective 'If, then'. Both use the term 'grass' to refer to grass and interpret 'If, then' in the same way, but the remainder of their assignments are inverted. Our gardener hither assigns 'green' to green things, 'healthy' to healthy things, and so on; but our gardener yon assigns 'green' to yellow things, 'yellow' to green things, 'healthy' to sick things, 'sick' to healthy things, and so on. It should be clear that on an occasion when each gardener forms a true belief by internalizing a representation of the form 'the grass is green', their beliefs are type-distinct. And, allowing for inferences, their mental processes may be formally indistinguishable. Suppose that each gardener believes that 'if the grass is healthy, then do nothing' and that 'the grass is healthy'. Our gardener hither may formulate a plan of action in agreement with the conclusion that he should "do nothing" and remain resting in the shade, while our gardener yon formulates a plan of action to carry out the instructions "do nothing" and proceeds to broadcast small pellets over the grass. There is an obvious difference between the functional roles of the formally indistinguishable beliefs that 'the grass is green', if we may consider the wide functional role of each belief token. Hence, individuation of mental states by reference to their formal structures alone will not, in certain cases, allow us to distinguish functionally distinct mental states: The wide functional roles of formally type-identical mental states may vary in different systems.

In summation then, purely formal taxonomies of mental states cut across functional taxonomies of mental states: A mental state of a given functional type may have various formal realizations, in a given system or in different systems, and formally indistinguishable internal representations may instantiate functionally different mental states. It is really not surprising that the formal approach to the individuation of mental states and the functional approach are less than fully compatible. Originally, of course, functionalism was conceived by Putnam as a means of providing a level of description for mental states such that psychological equivalences could hold across systems composed of different stuffs, e.g. the stuff of human physiology, the stuff of Martian physiology, or the stuff of a machine's hardware [Putnam, 1975, pp. 291-303]. But, once the formal-syntactic level of analysis is chosen as the one appropriate for specification of mental states, the generality sought by functionalism is placed out of reach. For, the formal, syntactic, or non-semantic properties of a mental state type, e.g. the belief that *P*, may vary across different systems. Recall that one observation which has been widely taken to rule out the possibility of neurological, or simply physicalistic, specifications of mental state *types* is that the neurological properties of the states which realize the belief that *P* may vary across different systems. By parity of reasoning, then, the possibility of exclusively formal-syntactic specifications of mental state types must also be rejected.

What we have revealed is a conflict between the putative implications of the formality condition and the defining desideratum of functionalism. The formality condition directs the theorist to respect only the formal properties of mental states, in assessments of their type, while functionalism is predicated upon the assumption that states of the same type may have realizations of different forms. What we require is a way to specify mental states that is largely non-commital with respect to the particular form of their realizations--at least this is what functionalism requires. This means that neither neurological nor formal-syntactic specifications of mental states are adequate. The functionalist, of course, will contend that some concept of the functional role of mental states is adequate for the purpose of individuating mental states.

It is to this claim that we will now turn our attention.

## FUNCTIONAL CRITERIA FOR THE TYPE-IDENTITY OF MENTAL STATES

The idea that the type to which a token mental state belongs is fixed by the functional role of the token forms the core of the functionalist view. And, as Fodor points out, the construal of the functional roles of mental states in terms of their causal roles is part and parcel of functionalism [see Fodor, 1981, pp. 16-17]. But, one may share with the functionalist the view that mental states have a causal role in the production of behavior without adopting the view that the causal role that a token mental state happens to inhabit determines the token's type. In this section, we will construct a series of arguments against the causal-functional approach to the individuation of mental states. There are many difficulties which beset the functional approach to the individuation of mental states. To anticipate what is perhaps the single most damning problem, functional criteria for the individuation of mental states fail to provide adequate means for the evaluation of type-relations across different <u>de facto</u> causal roles, i.e. the roles in which token states are actually employed on the occasion on which they occur. This issue will become clear as we proceed.

In order to motivate a careful consideration of this investigation of the possibility of causal-functional specifications of mental states, a few acknowledgements are in order. In the first place, throughout the following discussion we will allow that mental states do have causal-functional roles. Further, given comprehensive specifications of the complete functional roles of token states, one could type-distinguish tokens by locating some difference between their functional roles--though this assumes, of course, the availability of a criterion for the type-identity of functional roles. But we will suggest that the functional roles of token mental states can only be determined by reference to their types. That is, it may be the case, and it will be argued that it is the case, that the functional roles of mental states can only be determined once their types are (non-functionally) specified. On this view, the functional roles of mental states are, in a sense, "read off", or generated in accordance with, specifications of their types. If this view is correct, we are in a position to distinguish mental states by a comparison of their functional roles only subsequent to a determination of their types, in which case functional individuation is redundant and gratuitous.

Let us also acknowledge an observation that might be taken to suggest the possibility of causal-functional specifications of mental states, though one that is, by itself, insufficient to guarantee the possibility of such specifications of mental states: Token mental states of complete causal Doppelgängers, systems whose complete causal histories are type-identical, must be type-identical. That is, if we fix two systems so that every aspect of the wide and narrow causal history of occurrent beliefs arising in each at t is the same, and likewise for the remainder of the systems' belief and desire sets, then the tokens at t must be of the same type. Though this observation is congenial to functionalism and the causal analysis of mental states, as stated it asserts nothing more than the supervenience of mental states upon their causal embeddings. The concept of supervenience is commonly portrayed as a thesis concerning the relation of mental states to their physical realizations.[3] But, it will be helpful if we may say, more broadly, that mental states are supervenient with respect to a level of analysis, L, just in case a type-difference between token mental states necessitates a difference specifiable at L. Notice that, in these terms, there is nothing particularly surprising about the supervenience of mental states upon their causal embeddings. In particular, the discovery of a supervenience relation between mental states and some level of analysis L, does not pick out L as that level on which criteria for the type-identity of mental states are to be defined. This point can be illustrated by example. Mental states are supervenient on their physical realizations, i.e. supervenient with respect to the physical level of analysis, but physically dissimilar states may realize type-identical mental states. The observation, here, is that supervenience is an asymmetric relation in many, if not most, instances. Presumably, the correct level of analysis for mental states is just that level on which the supervenience relation is symmetric: A type-difference between mental states must necessitate a difference specifiable at L, and a type-difference specifiable at L, for realizations of token mental states, must necessitate a type-difference between the token mental states. Now, if it can be shown that type-identical mental states can possess different causal-functional roles, then we will have good reason for believing that causal-functional specifications of mental states are inadequate. Consequently, the functionalist must maintain that type-identical mental states cannot possess different causal roles and must provide a construal of the causal roles of mental states that justifies this view.

Let us anticipate, once more, certain conclusions toward which the following examination of the causal-functional analysis of mental states points. One level on which the supervenience relation appears to be symmetric is the semantic level. Arguably, type-identical mental states have equivalent semantic contents, and semantically equivalent mental states are type-identical. If this is so, the semantic level for the specification of mental states overcomes the defects that may be charged against the physicalistic level, the formal level, and the causal-functional level. For, type-identical mental states may have different physical realizations, different formal structures, and, on one plausible interpretation, different causal roles. Just possibly, the semantic level may not suffer the analogous defect. In any case, this is the direction in which the considerations to follow, together with those of the preceding sections, would appear to point. Our procedure in what follows will be to examine various formulations of causal-functional criteria for the type-identity of mental states in order to reveal certain fundamental problems confronting this approach to the individuation of mental states.[4]

Suppose that, as a first approximation, the functionalist attempts to define type-identity relations between beliefs over the actual, dated causes and effects of token mental states:

> (CR.1) Token mental states are type identical if and only if there exists a type-identity relation (a) between the token causes of the token mental states, and (b) between the token effects of the token mental states.

According to (CR.1), on its intended reading, token beliefs of the form 'the building is on fire' are type-distinct when, for example, one token is the causal consequence of the perception of smoke in the air and another token is the causal consequence of the interpretation of an English sentence, e.g. 'The fire alarm is ringing'. Thus, (CR.1) is incompatible with the intuition that beliefs of the same type may have causes and/or effects of different sorts on different occasions of occurrence. The belief that 'the building is on fire', intuitively, may be caused by any one of an indeterminately large number of circumstances, e.g. the sound of alarms, the smell of smoke, the sight of flames, or the interpretation of the utterance "The building is on fire". If we move from such wide specifications of the causes of beliefs to narrow specifications, the same circumstance obtains. A belief of a given type may form the conclusion of infinitely many different arguments, or reasoning processes, carried out by an agent. Further, a belief of a given type may underwrite a variety of different actions. For example, the belief "that it's raining" may function, in collaboration with other mental states, so as to affect the decision to stay inside or so as to affect the decision to locate an umbrella. If we were to adopt (CR.1), all tokens of a mental state type would, by definition, bring about type-identical effects on every occasion of occurrence, and no behavioral plasticity would be exhibited by the agents of belief.

Hence, a type-distinction between the actual effects, or between the actual causes, of token beliefs cannot, by itself, establish a type-distinction between the beliefs. We may put this by saying that differences between the _de facto_ causal roles of beliefs do not entail type-differences between the beliefs. One way to illustrate the difficulty which faces the causal-functional approach is to fix a token state of a system and update the system's knowledge in such a way that the repertoire of actions that the token underwrites is either broadened or narrowed. That is, we may change the functional role that a token mental state actually inhabits by changing other elements, other beliefs for example, of the system in which the token mental state occurs. Suppose that Alfred believes at _t_ that "there are blackbirds in the tree" and that, while holding this belief constant, Alfred is told at _t+l_ that "Blackbirds are good to eat". The original token underwrites an extended behavioral repertoire when a belief about the palatability of blackbirds is appended to Alfred's belief set, even though the original token is, _by hypothesis_, sustained through and past _t+l_. Similarly, Alfred may be disposed to endanger the well-being of blackbirds when he believes that "there are blackbirds in the tree" until he is informed that "Blackbirds belong to an endangered species". Being so informed, Alfred desists from his previous malicious, blackbird-directed behavior. Analogously, the causes of a token belief may change subsequent to the addition of new beliefs. Suppose that at *t* Alfred is caused to believe that "there are blackbirds in the tree" under conditions affording clear vision of a nearby tree. At *t+l* an ornithologist provides Alfred with a description of the characteristic calls of blackbirds that allows Alfred to distinguish the calls of blackbirds from those of other birds. Subsequent to *t+l*, with his line of sight directed away from the tree, Alfred's belief that "there are blackbirds in the tree" may be supported by stimuli of a sort quite different from those relevant at *t*.

The point illustrated by these examples is, quite simply, that a change in what we have termed the *de facto* causal or functional role of a token belief does not necessarily affect a reassignment of the token to a new or different type. Given th~~at a t~~oken belief's type-membership does not change with every change in the set of actions that the token underwrites, or with every change in the token's causes, what functional properties of a token determine its type-membership? The functionalist answer must be that the "causal-functional role" occupied by a token determines its type. But, on the most natural understanding of the notion of the causal roles of mental states, type-identical mental states may be deployed in different *de facto* causal roles. Thus, the functionalist requires an (unnatural) understanding of the causal roles of mental states according to which the causal roles of type-identical mental states may never vary.

Since the de facto causes and effects of type-identical tokens may belong to different types, the functionalist might advert to the possible causes and possible effects of mental states in an attempt to construct a concept of the causal role of a mental state appropriate to the program he envisages. Some formulation such as the following might be tendered:

> (CR.2) Token mental states are type-identical if and only if (a) for each counterfactually possible cause of one token there exists a type-identical counterfactually possible cause of the other token, and vice versa; and (b) for each counterfactually possible effect of one token there exists a type-identical counterfactually possible effect of the other token and vice versa.

The idea underlying (CR.2) is that mental states belong to the same type just in case they can be brought about by the same set of circumstances and have the capacity to bring about the same set of circumstances; i.e. just in case they share the same set of counterfactually possible causes and effects. Note that it is necessary to consider a token mental state's counterfactually possible causes and effects. Doing so allows us to circumvent the problem engendered by the fact that the actually possible causes and effects of type-identical beliefs may differ. That is, given the vagaries of different belief sets, a possible cause of the belief that *P* in one system is not necessarily a possible cause of the belief that *P* in the other system. For example, one system may believe that '*Q* → *P*', thus establishing the belief that *Q* among the possible causes of the belief that *P*, while the other system believes, independently, that *Q* is false, thus removing the belief that *Q* from the possible causes of the belief that *P*.

The problem with (CR.2) isn't that it is flatly false -- in our introductory remarks in this section we allowed that something of this sort might be true. The problem is that the resources of the causal-functional approach are inadequate for determining when the conditions posed by (CR.2) are satisfied. Ideally, to employ (CR.2) in the individuation of mental states the functionalist requires access to an effective method for enumerating the infinitely many possible causes and effects of token mental states. This asks a great deal from functionalism and there is good reason for believing that it is more than functionalism can provide. In fact, we will suggest that given a token mental state *M*, and a specification of its *de facto* causes and effects, the functionalist has no means for identifying reoccurrences of tokens of the type to which *M* belongs. If this is correct, then the functionalist cannot determine whether or not a particular token is among the counterfactually possible causes of a token mental state M -- hence there can be no effective method for enumerating the infinitely many possible causes and effects of M.

When do two tokens possess the same counterfactual causal role? In our criticism of the criterion (CR.1) it was only necessary to point out that the type-identity of the *de facto* causes and effects of a pair of mental tokens is not a *necessary* condition for their type-identity: Mental states may, thus, possess equivalent counterfactual causal roles even though their actual causes and effects are different. The functionalist's attempt to determine if a pair of tokens possess equivalent counterfactual causal roles is further frustrated by the following fact. The type-identity of the *de facto* causes and effects of a pair of tokens is not a *sufficient* condition for their type-identity. If we specify only equivalent *de facto* causes and effects for a pair of mental states, we will have specified conditions that are compatible with the realization of type-distinct mental states. For example, I might believe that P (= it's dark outside) or Q (= it's nightime) under the pressure of a common cause, e.g. the belief that "it's 12:00 midnight". Both P and Q may cause me to do the same thing, e.g. to carry a flashlight outside, on some occasion. But, there are possible causes of P that are not among the possible causes of Q. The belief that "it's dark outside" might be caused, on one occasion, by the belief that one's locale is experiencing the heaviest cloud cover ever seen in the world. Thus, the type-identity of the *de facto* causes and effects of a pair of tokens does not guarantee the type-identity of their counterfactual causal roles. In its fully general form, the point here is that the actual causes and effects of token mental states may agree within a given sample while the counterfactually possible causes and effects of those tokens diverge outside the sample.

These points do not argue against the metaphysical picture presented by (CR.2), but they do argue against the claim that the functionalist can *employ* (CR.2) successfully when faced with the task of individuating actual mental tokens. According to (CR.2), the functionalist must in some way specify the counterfactual causal role of token mental states if he is to evaluate the type relations which lie between them. But the only functional properties of a token to which the functionalist may claim access are those properties constituted by the *de facto* causes and effects of the token, and the equivalence of these properties for a pair of tokens is compatible with a type-difference between the tokens. If this is correct, the functionalist cannot effectively specify all the possible causes and effects of a token M since the functional properties he may justifiably attribute to M are compatible with different counterfactual causal roles.

The functionalist might attempt to avoid these problems by formulating a functional type-identity criterion that does not advert to all the counterfactually possible causes and effects of particular mental states. However, there is a highly general problem associated with (CR.1) and (CR.2) that we should mention before proceeding to an examination of functional criteria for the individuation of mental states which, putatively, do not require the effective enumeration of all the possible causes and effects of token states. If the type-identity of token beliefs, occurring at different times or in different agents, rests on "sameness of causal role" (actual or possible), then a criterion for the type-identity of the causal roles of beliefs is required. We can look at the situation in the following way. Suppose that we are interested in the set of possible causes of a belief. Such a set is a set of types. For example, we might claim that tokens of the belief that P can be caused by token beliefs of any of the types {Q, R, S,...}. In order to determine the type relation between two arbitrary beliefs A and B, then, we must determine if each token possible cause of one belief is type-identical to a token possible cause of the other belief. Since exclusively wide specifications of the causal roles of mental states are inadequate, reference must be made to the internal causes of mental states as well. But, *beliefs* are among the possible internal causes of beliefs. Hence, both (CR.1) and (CR.2) define the type-identity of beliefs in terms of the type-identity of beliefs. Since type-identical beliefs must have the same possible causes, and since beliefs are among the causes of beliefs, the type-identity of the token beliefs *A* and *B* depends in part, upon the type-identity of their token causes *A** and *B**, which are themselves instances of belief. This problem is one that iterates without limit. The type-identity of *A** and *B** depends upon the existence of a type-identity between the tokens which represent possible causes of each belief, and *A** and *B** have other beliefs among their possible causes and effects.

Although it has been suggested otherwise, the embarrassment of individuating mental states by reference to the mental state types to which they are related cannot be avoided by Ramsification. S. Shoemaker has recently claimed that the Ramsey-sentence for a psychological theory avoids the circularity problems which may beset the original formulation of the theory. Interestingly, if Shoemaker were correct we could provide just what (CR.2) lacks, i.e. a way to specify the type of a token, *M*, that does not require specification of the type of every other token to which *M* may be related. However, the Ramsey-sentence for a theory will preserve any circularity endemic to the theory's original formulation. Since this point is sometimes missed, consider Shoemaker's construal of the belief that "it's raining" and the desire to "keep dry" on a psychological theory $\underline{T}$:

> The Ramsey-sentence of $\underline{T}$ can be written as
>
> $\exists F_1 \ldots \exists F_N\ T(F_1 \ldots F_N)$
>
> If '$F_1$' is the variable that replaced 'believes that it is raining' in the formulation of the Ramsey-sentence, and '$F_2$' is the variable that replaced 'wants to keep dry', then the following biconditionals will hold:
>
> (1) $\underline{x}$ believes that it is raining $\leftrightarrow \exists F_1 \ldots F_N[T(F_1 \ldots F_N)\ \&\ \underline{x}$ has $F_1]$ and
>
> (2) $\underline{x}$ wants to keep dry $\leftrightarrow \exists F_1 \ldots F_N[T(F_1 \ldots F_N)\ \&\ \underline{x}$ has $F_2]$. [Shoemaker, 1981, pp. 93-94]

Shoemaker claims that the predicates on the right-hand side of the biconditionals "quantify over mental properties, or states, but do not mention any specific ones (the Ramsey-sentence having been purged of mental predicates); so no circularity is involved in defining the belief in terms of (1) and the desire in terms of (2)" [p. 94]. The appearance that all is well with definitions like (1) and (2) is purely specious and is the product, in part, of the excessively skeletal form of the Ramsey-sentence formulations offered[5]. If the non-Ramsified theory T functionally specifies the belief "that it's raining" by some description such as "the belief which in collaboration with the desire 'to stay dry' causes an agent to search for an umbrella or to stay inside," then the Ramsified theory will specify the state $\underline{F}_1$ by some description such as "the state which in collaboration with the state $\underline{F}_2$ causes an agent to search for an umbrella or to stay inside". Further, reference to $\underline{F}_1$ is similarly essential in the specification of $\underline{F}_2$ as long as the original theory interdefines the relevant beliefs and desires--which Shoemaker assumes to be the case. The crucial point is that if the original theory T interdefines 'belief' and 'desire' in any way, then the Ramsey-sentence for T interdefines the states that satisfy the variables which uniformly replace the theoretical terms 'belief that P' and 'desire that Q'. Functional individuation, by nature, picks out mental states by reference to their relations to other mental states, and to inputs and outputs. Since reference only to inputs and outputs is insufficient to individuate mental state types, reference to internal states is required. Thus, a kind of interdefinition of state-types is unavoidable in functional specifications of mental states, whether given in a Ramsified form or not.

How, then, might reference to the causes and effects of mental states provide for their individuation? We cannot require, as in (CR.1), complete equivalence of the actual causes and actual effects of token beliefs on pain of distinguishing obviously equivalent beliefs, e.g. logically equivalent beliefs arrived at by the formulation of different arguments. We cannot, as in (CR.2), require the type-identity of every counterfactually possible cause and effect of token beliefs on pain of launching an infinite regress, i.e. determining the type-identity of beliefs by determining the type-identity of the beliefs which serve as their causes and iterating this step indefinitely many times. Moreover, evaluation of a token's type in terms of its infinite set of possible causes and effects requires the effective enumeration of its causes and effects and we have argued that functionalist resources are inadequate for this task. However, we will consider two approaches to the functional individuation of mental states which, putatively, do not require the effective enumeration of a token's infinite set of causes and effects. One approach is proposed by Shoemaker and the other, involving the notion of the proprietary role of a mental state, is easily anticipated.

Now, token mental states typically have a behavior guiding potential that is not fully utilized by any of the de facto systems in which they occur; no system has all the beliefs and desires with which tokens of a type, e.g. the belief that P, could collaborate. Nonetheless, we might attribute to a token mental state, in a spirit similar to that of (CR.2), a causal potential constituted by the sum of its causal properties in all logically possible systems. Mental states will, typically, have infinitely many possible causes and effects, but if we cannot effectively enumerate the relevant infinite sets of causes and effects, by applying some functional theory, then the appeal to the causal potentials of mental states would appear to be empty. Shoemaker proposes a treatment of a closely related problem which does not require, he holds, the specification of the infinite sets of causes and effects of each token mental state. Shoemaker observes that properties or states--he uses the term interchangeably in an unduly confusing fashion--have two sorts of causal features:

> One sort are causal potentialities; a property has a causal potentiality in virtue of being such that its instantiation in a thing contributes, when combined with the instantiation of certain other properties, to the possession by that thing of a certain causal power....The second sort of causal feature has to do with the ways in which the instantiation of a property can be caused....Every property has many, perhaps in some cases uncountably many, causal features of each of these kinds. [Shoemaker, 1981, pp. 105-106]

Shoemaker attempts to define a functional property attributable to states in virtue of their possession of these two sorts of causal features, even though "there is no guarantee that...a finite specification of the causal features of a property is possible" [p. 106]. Shoemaker takes the fact that the causal properties of states cannot be finitely specified, i.e. cannot be given by the specification of finite sets of properties, to pose the central problem for functional specifications of mental states. Shoemaker doesn't entertain the possibility of effectively enumerating a state's infinitely many causal properties, but he proposes, instead, that functional individuation be carried out by reference to a certain second-order property of states, a property for which he reserves the term _functional_ _correlate_:

> It is obviously true that if all the members of a set can belong to the same property [state], there is a functional property which something has just in case it has a property having all the causal features in the set. I propose that we extend the notion of functional property by stipulating that this is true of infinite and uncountable sets of features as well; the functional property corresponding to such a set will be the property of having a property having all of the causal features in the set...we may say that corresponding to any property _P_ there is a functional property, its functional correlate, which something has just in case it has a property having the totality of the causal features possessed by _P_. [Shoemaker, 1981, p. 106]

The idea is that since we are not in a position to specify the infinitely many causal features of a functional state, we must instead refer to a determinable functional property of such states: According to Shoemaker, the relevant property, the functional correlate, of a state, S, is the property which S possesses if and only if some property of S has all the causal features of S.

The notion of the functional correlate of a state is employed in a criterion for the type-identity of token states. Shoemaker actually holds that type-identity relations can be defined for token states across possible worlds in terms of the transworld identity of functional correlates, but this detail need not concern us here [see p. 107]. In any case, a criterion for the type-identity of mental states of the following sort is strongly suggested:

> (CR.3) Token mental states are type-identical if and only if their functional correlates are identical.

Reference to the functional correlates of mental states is proposed in an attempt to avoid the need to actually specify all the possible causes and effects of mental tokens. Now, if mental states, or any other sorts of states, can be <u>fully</u> individuated by reference to their causal properties, then it follows that the "property" of a state S that has a certain infinite set of causal properties is just S itself.

Shoemaker, in fact, will agree, for he maintains that if the causal theory of properties is true, i.e. the theory that "it is a necessary and sufficient condition for the identity of properties _A_ and _B_...that _A_ and _B_ share all of the same causal features," then "every property [state] will be identical to its functional correlate" [p. 107]. But, if token states are identical to their functional correlates, then (CR.3) has it that token mental states belong to the same type just in case they are (type) identical. Hence, the functional correlate theory of mental state types, _a la_ Shoemaker, is utterly uninformative unless an independent procedure for the specification of the functional correlates of mental states can be provided. Shoemaker explains, in an abstract way, what sort of property a functional correlate is, i.e. that property of a state, _S_, which _S_ has if and only if _S_ has a property having all the causal properties of _S_, but he does not offer criteria for distinguishing the functional correlates of token states.

In order to employ (CR.3) to advantage, we need to know the conditions under which the functional correlates of mental states are different. Without question, differences between the functional correlates of token states depend upon differences between the infinite sets of causal properties associated with the tokens. Thus, we return, full circle, to the problem the functional correlate is intended to circumvent: There is no available method for specifying the infinitely many possible causes and effects of token mental states--at least no such method based merely upon the details of a token's actual causes and effects on the occasion on which it occurs.

There is another approach to the functional individuation of mental states which may be attractive to some functionalists. Like the functional correlate theory, this approach, which we will term the proprietary role theory, attempts to avoid the need to specify the infinitely many possible causes and effects of tokens of a given mental state type. But, unlike the criteria we have considered to this point, the proprietary role theory rests on an idealization. Consider an idealization of a system which attributes perfect rationality to the system along with all the beliefs the system requires in order to work toward its fixed goals in the best possible way. The functionalist might claim that the type-identity of mental states is to be evaluated in the context provided by such a system. According to this idea, the functional individuation of mental states is to be accomplished by reference to a putatively definitive aspect of the possible causal roles of mental states, rather than by reference to the infinitely many causes and effects of tokens of a type in all possible systems. As a first approximation, consider the following criterion:

> (CR.4) Token mental states are type-identical if and only if they would possess the same causal role in the ideal system.

Presumably, the ideal system is perfectly rational and perfectly informed. Thus, the proprietary role of the belief that "the building is on fire" is to represent the fact that the building is on fire and to effect an appropriate response, e.g. evacuation from the building, to that fact. According to (CR.4), any token which would play this role in the ideal system is a token of the belief type "the building is on fire". But (CR.4) would force us to type-identify beliefs that involve the misuse or misunderstanding of certain concepts with ideally realized beliefs in which no misuse or misunderstanding of the concepts is involved. Suppose that an agent incorrectly uses the term 'green' to refer to yellow objects. We will not, contrary to (CR.4) type-identify occurrences of 'this is green' across the ideal system and the anomalous system. That is, although all tokens of the form 'this is green' would possess the same functional role in the ideal system, they are not thereby type-identical.

It is difficult, perhaps impossible, to construct a proprietary role theory of mental state types which avoids this difficulty. We might attempt to improve upon (CR.4) by avoiding the counterfactual condition to the effect that tokens belong to the same type just in case they would possess the same roles in the ideal system. One way to do this might be to define a set of proprietary roles by reference to pairings of inputs and outputs and to specify the conditions under which a token occurs in its proprietary role.

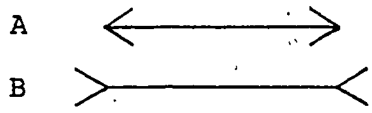
> (CR.5) Token mental states are type-identical if and only if they possess the same proprietary role: A token mental state _M_ occurs in its proprietary role _R_ if and only if, under ideal conditions, _M_ occurs as a consequence of _C_ and _M_ causes _B_.

The idea, here, is to specify causes, _C_, and effects, _B_, that are _definitive_ of a token mental state's type without being _exhaustive_ of the causal role of the type. For example, those tokens that occur under ideal conditions as a consequence of _C_ =(a fire in the building), and which cause _B_ =(evacuation of the building), belong to the type "the building is on fire". Of course, reference to internal causes and effects in the specification of the relevant _C_'s and _B_'s is also required in order to rule out the type-identification, for example, of tokens of the forms 'the building is on fire' and 'it is prudent to act on the assumption that the building is on fire'. In this connection, (CR.5) can likely be improved, but no proprietary role theory can evade the problems associated with non-proprietary occurrences of mental states.

Before considering the problems that non-proprietary occurrences of mental states pose for (CR.5), and similar criteria, it is perhaps worth noting that the idea underlying (CR.5) is analogous to one that Fodor employs for a slightly different purpose. Fodor wants to provide a theory of the semantic endowability of mental states [forthcoming b]. Toward that end, Fodor suggests that there are teleological facts about cognitive systems of roughly the following sort. The function, i.e. purpose, of a cognitive system is to believe only those propositions whose truth conditions are satisfied. Given such a purpose, there must exist a certain psychological function which provides for the fulfillment of that purpose. Fodor maintains that this function, dubbed the "yes-box function", has a special place in a story about the semantic endowability of mental states since the relation '_M_ is in the yes-box' is coextensive with the relation '_S_ is the truth condition of _M_', where _M_ is any representable proposition. What Fodor's intriguing study, which we have merely touched on in the briefest possible way, shares with (CR.5) is the use of a severe idealization. According to (CR.5), the type-relations of beliefs are to be assessed under idealized conditions. According to Fodor, "an organism _O_ believes _M_ iff _M_ is true" and, thus, beliefs by definition never arise under less than the most auspicious of conditions [see Fodor, forthcoming b]. But, to paraphrase a comment that Dennett makes in a completely different context, an idealization that has the consequence that believers believe all and only the truths would appear to make a better cliff over which one may push one's opponents than a live option.

Consider, then, the possibility of live options. Can beliefs be individuated by reference to their proprietary roles? On a proprietary role construal of mental state types, one must hold either (a) that tokens which occur in non-proprietary roles may be type-identical to tokens which occur in proprietary roles, or (b) that tokens which occur in non-proprietary roles are never type-identical to tokens which occur in proprietary roles. We may quickly see that (a) is not a genuine option for the proprietary role theory of mental state types. Although it is intuitively correct to hold that mental states can be type-identified across proprietary and non-proprietary occurrences, if this alternative is chosen, then reference to the proprietary roles of mental states will do no work in the theory of state types. Given a belief which constitutes an illicit conclusion of an internal inference we cannot, following (CR.5), maintain that its type is fixed by the token beliefs to which it is actually causally related since they provide a non-proprietary context for the belief. Nevertheless, if we know precisely how an inference goes awry, e.g. what non-truth-preserving rule is applied, we can calculate the content of the inferred belief: '$P \rightarrow Q$', '$Q$', and mistaken reliance on affirmation of the consequent delivers a belief type-identical to the antecedent of '$P \rightarrow Q$'. Thus, the non-proprietariness of the causes of the belief that *P* may have no effect upon the type-membership of the belief, i.e. the token belongs to the type 'that *P*' even though its causes are less than ideal. But, if tokens are sometimes type-identical across proprietary and non-proprietary roles, then on what does their type-identity depend? One might attempt to hold, for example, that the type-identification of tokens across different causal roles is determined by the indistinguishability of their formal structures or by the equivalence of their contents. To revert to either condition is to abandon the proprietary role theory of mental state types since the formal structures and the contents of mental states are, in many cases, independent of the proprietariness of their derivations.

The other alternative fairs no better. Maintaining (b), that tokens which occur in anomolous roles may never be type-identified with tokens which occur in ideal roles, is worse than simply counterintuitive. What makes a causal sequence non-proprietary, presumably, is that the belief which occurs in such a sequence receives an inadequate epistemic warrant from the circumstances which bring it about, or that the belief provides an inadequate epistemic warrant for the beliefs and actions which it brings about. In any case, it will not do to maintain that the proprietary sequences can be picked out statistically, or identified with what is "normal" for a population. For, consider any of the well known illusions, e.g. the Muller-Lyer illusion:



The belief that is normally produced in subjects in this illusion, "line B is longer than line A", occurs outside its proprietary role. Illusions provide only one case among many. For example, inductive inferences based upon unreasonably small samples, deductive errors, heuristic rules that are misapplied, sheer guesswork, and sheer associative connections all have the capacity to produce inadequately warranted, or unjustified, beliefs. But to capture the idea that a token belief is unjustified, it is necessary to contrast the (anomolous) conditions which give rise to the token with the (ideal) conditions which justify tokens of the same type. The notion of proprietary causal embeddings for beliefs, thus, presupposes the definability of type-identities between tokens occurring in anomolous roles and tokens occurring in ideal roles. For if one forfeits the possibility of type-identities across anomolous and ideal circumstances, then there can be no basis for the evaluation of a particular role as anomolous and, hence, no basis for a contrast between anomolous and ideal roles, non-proprietary and proprietary roles.

## REJECTION OF FUNCTIONAL SPECIFICATIONS OF MENTAL STATES

The basic problem with (CR.5), and all versions of proprietary role theories of mental state types, is that it provides no means for assessing the type-membership of an arbitrary token in an arbitrary system. Type-identical mental states will occupy different de facto causal roles in systems with different belief-desire sets or with different inferential performance abilities. For this reason, among others, (CR.5) appeals to an idealized system with a fixed belief-desire set and fixed inferential rules. Now, in a completely fixed system, a belief of a given type will always have the same causes and effects. But, this result is a consequence of the supervenience of mental states upon their causal embeddings: If mental states are supervenient with respect to phenomena describable on level L, then type-identical conditions describable on L will realize type-identical mental tokens, if those conditions realize any mental states at all. In addition to this, a theory of mental state types must specify a level L such that the type-identity of mental states entails the type-identity of the phenomena describable on L. In the present context, this means that the type-identity of mental states must entail the type-identity of their causal roles. This demand for a construal of the type-identity conditions of token mental states that entails the type-identity of the causal roles of the tokens vitiates the causal-functional approach to the individuation of mental states. For, unless we fix all the beliefs, desires, and inferential practices of two systems, the causal-functional roles of type-identical states may differ; but if we fix all the beliefs, desires, and inferential practices of two systems, then we relativize type-identity relations to systems of particular fixed types--in effect, we relativize to particular systems. In either case, we fail to provide criteria for the type-identity of mental states across systems. Reference to the ideal system, for example, is just reference to a particular system with fixed beliefs, desires, and rules of inference and, hence, fails to avoid this dilemma.

It will be helpful to review some of the problems that give rise to this dilemma, i.e. some of the problems that the proprietary role theory seeks to circumvent. Since the actual token causes and effects of a token mental state do not fix the state's type, a tenable causal criterion might specify the counterfactually possible embedding of a token state. But, there are infinitely many possible causes and effects associated with the tokens of a type. In lieu of the effective enumeration of a token state's possible causes and effects, the theorist must specify functional properties of the token's causal role that fix the token's type. For, what we have called the counterfactual causal roles of token mental states cannot be specified by reference to the _de facto_ causes and effects of token mental states. The functionalist must, then, specify some determinate functional property of each token over which type-identity conditions are to be defined. There are two approaches to the specification of such a property which serve to illustrate the difficulty that faces the functionalist. One approach to this problem, due to Shoemaker, adverts to a property which mental tokens possess in virtue of possessing a particular, but unspecified, set of causal features. This approach fails since there is no way to distinguish the "functional correlates" of mental states short of (a) specifying their infinitely many causes and effects, or (b) specifying the types to which they belong. Another approach adverts to the putative causal-functional property "has the proprietary role R". But, idealizing in the manner required by the proprietary role theory doesn't help since actual mental states often occur under less than ideal conditions and the proprietary role theory can provide no mechanism to accommodate tokens which occur outside their proprietary roles.

The lesson associated with these observations seems to be the following. Given a mental token of a _specified type_, e.g. the belief that "the moon is blue", it is a simple enough matter to evaluate aspects of its functional role in various possible systems, or in a system whose beliefs undergo various changes. But, the functionalist requires a way to specify the functional role of a token belief _without reference to its type_. And, the only properties of a token's functional role to which the theorist has access are those _de facto_ causes and effects associated with the token on the occasion on which it occurs. If we are to calculate further aspects of a token's possible, or dispositional, functional role, we require access to a procedure for identifying occurrences of tokens of the same type--just what functionalism seeks, but fails, to provide.

These considerations suggest a non-functional approach to the individuation of certain mental state types, i.e. belief types, desire types, or the subtypes of all propositional attitude types, e.g. the belief that _P_. Since we require a way to identify occurrences of tokens of a type deployed in various different _de facto_ causal roles, ideally, we require access to a property of a token that is invariable throughout its various occurrences. Unlike Shoemaker who constructs an abstract property attributable to mental states in virtue of their possession of infinitely many causal features, we envisage reference to a property, C, such that the possession of an infinite set of functional features is a function of C. This approach can be motivated by example. A carburetor, to use a well-worn example, is a mechanism possessing a certain functional role, but a token carburetor has the capacity to serve in a certain functional role only by virtue of certain of its intrinsic properties. On the physicalistic level of analysis, the intrinsic properties of token carburetors may vary widely without intruding upon the functional role that defines the type to which they belong. Arguably, the class of physical properties of a thing by virtue of which it could serve as a carburetor is both disjunctive and open. Thus, the question "By virtue of what does a carburetor possess the capacity to fill a certain functional role?" admits of infinitely many different answers. But mental states aren't carburetors, and it is just possible that there is a level of analysis appropriate to mental states--though not, alas, appropriate to carburetors, poor things--on which the analogous question for mental states has a determinate answer. If such a level of analysis is available, then we might capture the idea that mental tokens have causal-functional roles without requiring the actual specification of the infinitely many functional features associated with mental states. Given such a level of analysis, type-identity relations will be specified by reference to that property, or those properties, in virtue of which tokens have the capacity to fill various functional roles. It is to this idea that we now turn.

THE SEMANTIC EQUIVALENCE OF MENTAL STATES AND OPACITY

In the remaining portion of this chapter we will argue that semantic conditions for the equivalence of mental states can, and should, be given. In essence, the view advocated is that token mental states are type-identical if and only if their contents are equivalent. Now, before leaving functionalist speculations altogether, notice that a certain type of (hypothetical) functionalist might claim to embrace the view that mental states are type-identical just in case their contents are equivalent. For, he will argue, the contents of mental states can be "functionally" determined. Although the view is usually vague when it comes to details, some philosophers appear to take the view that the functional roles of mental states somehow exhaust the contents of mental states. G. Harman, for example, has recently alluded to a semantic theory according to which "the contents of concepts and thoughts are determined by their functional role in a person's psychology" [1982, p. 242]. Harman uses the idea of the functional role of a thought very loosely, but if we insist on a causal version of functionalism it can easily be seen that there can be no "functional theory of content". The functionalist with the interesting position is, after all, the one who maintains that mental states are to be individuated causally. In this connection, our hypothetical functionalist must claim that the causal role of a mental token fixes its content. But, under what sorts of descriptions are the causal inputs and output of a mental token to be given? A stimulus event might be described in a physical idiom, i.e. described by predicates which are projectible given the body of physical theory. However, phenomena so described are amenable to a variety of internally produced interpretations. This observation enjoys a host of simple but highly effective illustrations, e.g. the duck-rabbit figure, the faces-vaces figure, and the Necker cube. The same circumstance holds for the outputs of mental processes, as we have taken pains to point out in Chapter I. As Pylyshyn points out, physically described stimuli are inherently ambiguous [1980, p. 112]. We cannot expect to fix, or even to constrain, the interpretation of mental tokens by reference to descriptions under which the inputs and outputs associated with such tokens are inherently ambiguous.

The seductive appeal of the idea that, to paraphrase Harman only slightly, the contents of thoughts are determined by their functional role is an artifact of allowing oneself recourse to fully interpreted specifications of the inputs and outputs of mental states. Importantly, to appeal to such fully interpreted specifications of inputs and outputs is to adopt a semantic approach, of one kind, to the individuation of mental states -- though on the view urged here it is a "backwards" semantic approach. Once fully interpreted inputs and outputs are given, it is often a trivial task to specify the content of a mental token. In fact, even narrow construals of the functional roles of mental states can fix the content of mental states once such interpretations are provided. For example, if we know that a token belief that expresses the content that '_P_ and _Q_' is supplied as input to the _simplification_ operation, i.e. '_P_ and _Q_ → _P_', then we know that the resulting token expresses the content that _P_. Thus, we can ask for the conditions which fix the interpretation of inputs and outputs, and attempt to fix the contents of mental states in terms of interpreted inputs and outputs; or we can ask for the conditions which fix the interpretation of mental states, and then interpret inputs and outputs in terms of the contents of mental states. Yet, these options are not equally plausible. Only the latter is compatible with the conception of cognitive systems as _interpretive_ systems and is, thus, strongly preferred.

Let us turn to an investigation of some of the problems which face the semantic approach to the individuation of mental states. If the semantic level is appropriate for the individuation of mental states, then (i) a type-difference between token mental states requires a difference between the semantic contents of those tokens, and (ii) a difference between the semantic contents of token mental states requires a type-difference between those tokens. One apparent problem for this view is presented by the opacity of mental state attributions: The belief that '_a_ is _F_' may be correctly attributable to $\underline{S}_1$, while the belief that '_b_ is _F_' is not correctly attributable to $\underline{S}_1$ even though '_a_ = _b_'. A striking example of the opacity of the belief attribution context is presented by the case of Oedipus. Oedipus believes of himself that "I want to marry Jocasta", but not that "I want to marry my mother" [see Fodor, 1980, p. 66]. Of course, at the stage where the former belief is correctly attributable to Oedipus, the coreference of 'Jocasta' and 'my mother' is not something that Oedipus recognizes. Thus, token beliefs which differ only with respect to the occurrence of coreferential terms are not always jointly attributable to an agent -- such tokens are often type-distinct. If this is correct, then we must be prepared to specify a semantic difference between certain extensionally equivalent beliefs. This is precisely what we propose to do.

However, it is crucial to decide whether extensionally equivalent beliefs which implicate different coreferential terms should always be viewed as necessarily type-distinct. That is, does the occurrence of syntactically or formally distinguishable names in token beliefs entail a type-difference between the beliefs? As we noted in Chapter I., Fodor has recently allowed that formally distinguishable beliefs may belong to the same semantic type. In fact, Fodor employs this observation in an argument for the inclusion of a semantic component in psychological theory [Fodor, forthcoming]. Thus, the mere formal difference between '_a_ is _F_' and '_b_ is _F_' does not, on this view, establish their type-distinctness. Are there specifiable conditions under which this pair of belief tokens belong to the same type?

Now, though the belief attribution context is opaque, the substitution of identity is an operation that a cognitive system might apply to an internal representation. For example, an agent who is informed that "The current President of the United States will arrive at 8:00 P.M." may come to believe that "Ronald Reagan will arrive at 8:00 P.M." _via_ the application of a rule for the intersubstitution of coreferring terms. Presumably, such an operation on an internal representation is not only possible but extremely common -- the substitution of coreferring terms into internal representations may constitute an empirical norm for human agents. In the context of belief attribution, if we can justify the attribution of '*a* is *F*' and the coreference assignment '*a* = *b*' to an agent, then it would appear that we can justify the attribution of '*b* is *F*' to that agent. This suggests that we needn't hold that all tokens which differ with respect to the occurrence of coreferential terms are type-distinct. Hence, we might offer the following criterion for the type-identity of certain beliefs.

> (SC.1) Token beliefs which differ only with respect to the occurrence of the coreferential terms '*a*' and '*b*' are type-identical, for a system $\underline{S}_1$, if and only if $\underline{S}_1$ internalizes a representation of the coreference of '*a*' and '*b*'.[6]

The idea is to provide, first, a criterion that rules '*a* is *F*' and '*b* is *F*' type-identical when, *from the perspective of the agent who internalizes those tokens*, the contribution of '*a*' and the contribution of '*b*' to the content of those tokens is the same. The intuition underlying (SC.1) is that the equivalence of the semantic contributions of different names to token mental states depends upon the internal representation of the coreferentiality of those names. This view is particularly plausible from the perspective of what J. Searle calls "no sense" theories of names [Searle, 1965, p. 487]. According to such theories, the semantic contribution of a name to a statement is exhausted by the referential properties of the name, i.e. by the reference that the name determines.

It would be foolhardy to underestimate the considerable difficulties and the notorious subtilities that confront the formulation of a general theory of names. But, something like (SC.1) is, arguably, preferable to a theory which determines the type-relation between the beliefs that '*a* is *F*' and that '*b* is *F*', in part, by reference to the senses associated with the coreferring terms '*a*' and '*b*'. Consider, briefly, the Fregean observation that the cognitive significance of '*a* = *a*' is not on a par with the cognitive significance of '*a* = *b*', even where the two statements are extensionally equivalent. Frege may be interpreted to hold that explaining the relevant difference between '*a* = *a*' and '*a* = *b*' confronts the theorist with a dilemma: If '*a* = *b*' is interpreted as asserting a relation between objects, then, if true, it asserts only a relation of an object to itself. If, on the other hand, we take '*a* = *b*' to convey a relation between signs, i.e. between '*a*' and '*b*', then "the signs themselves would be under discussion...[and] we would express no proper knowledge by its means" [Frege, in Geach and Black, 1952, pp. 56-57]. The latter horn of the dilemma conveys the need to capture the notion of what, in more modern terms, is referred to as a theoretical identity: The assertion that "Hesperus is Phosphorus" conveys a discovery about the world, not a discovery about a pair of names. However, as we will see, there are purposes for which the interpretation of '*a* = *b*' presented by the latter horn of the dilemma is most appropriate.

Frege's solution to the apparent dilemma involves an appeal to a *sense* or *intension* associated with each name: The reason that '*a* = *b*' has a significance that '*a* = *a*' lacks is that '*a*' and '*b*' are associated with *different* senses which, nevertheless, determine the same reference. Now there is no need to deny any insight that may lie in this direction, but consider the following case. Suppose that the term 'caw' is used in a certain language group to refer to some sacred relic housed by the high priest. For the sake of picturesqueness, we can imagine that 'caw' is used to refer to the hammer that legend proclaims was used to slay the One True Prophet recognized by the group. One day the high priest lets it be known that the caw may be referred to by the name 'bop' (maybe even 'caw-bop'). No justification is given for the introduction of the new name, nor is the new name associated with any new description which the caw satisfies. Under such circumstances it would appear that the senses associated with 'caw' and 'bop' are equivalent. Nevertheless, the "cognitive significance" of 'caw = caw' and 'caw = bop' may still differ for some arbitrary member of the group. The fact that, by hypothesis, the sense of 'caw' is also the sense of 'bop' does not rule out the possibility of a difference between the significance of 'caw = caw' and 'caw = bop' for some cognitive agent.

Notice that the tenability of this story depends only upon the following thesis. A difference between the syntactic form of two names does not necessarily indicate a difference between the senses associated with the names. It is extremely difficult to see how an advocate of the sense theory of names could deny this thesis. Moreover, if the thesis is accepted, then the puzzle presented by the fact that '_a_ = _a_' and '_a_ = _b_' differ in significance is not wholly resolved by the sense theory of names. For, even if the formally distinguishable names 'caw' and 'bop' are associated with precisely the same senses, the information that there is a coreference relation between those names, caw = bop, is to be distinguished from the information that 'caw' refers to whatever it refers to, i.e. the information which is conveyed by 'caw = caw'. What we are suggesting is that even after the appeal to the senses of names, a residual puzzle for a theory of names remains if formally distinguishable names may be associated with the same sense. And, it appears to be completely within the realm of possibility that linguistic convention, for example, could associate the same sense, or the same descriptions, with formally distinguishable names -- this may happen, for all we know, across different natural languages.

These observations suggest a concept of the _informativeness_ of coreference assignments that does not appeal to the senses associated with names. Presumably, it is potentially informative to be told that the reference of '*a*' is the same as the reference of '*b*', but it is never informative to be told that the reference of '*a*' is the same as the reference of '*a*'. Thus, we might construe coreference assignments as conveying information (beliefs) about the correct usage of terms. When an agent learns that a coreference relation holds between formally distinguishable terms part of the information acquired is that each term is used, or can be used, to refer to the same object. Obviously, we cannot construe identity statements relating terms of the same formal type in this way. An agent who has acquired the use of the name '*a*', is not in a position to be informed by the "information" that '*a*' can be used to refer to what '*a*' refers to. Statements of the form '*a* = *a*' are merely substitution instances of some rule for self-identity, e.g. (∀x) (x = x). Notice that if these observations can be sustained, then '*a* = *a*' and '*a* = *b*' are not a pair of statements which differ *only* with respect to the occurrence of coreferential terms. The latter statement, unlike the former, has the capacity to convey information about linguistic usage. Hence, (SC.1) need not assign the beliefs that '*a* = *a*' and that '*a* = *b*' to the same semantic type when both occur within a given system -- recall that (SC.1) is qualified in such a way to apply only to beliefs which differ only with respect to the occurrence of coreferring terms.[7]

These brief reflections concerning sense theories of names are intended to provide support for (SC.1). Presumably, the token beliefs '_a_ is _F_' and '_b_ is _F_' are semantically equivalent according to sense theories of names only if the sense associated with '_a_' is identical with the sense associated with '_b_'. But, according to (SC.1), a type of semantic equivalence holds between these two tokens within any system which represents the coreference of '_a_' and '_b_', whether or not the senses associated with the terms are synonymous. Sense theories are troubled by the fact that conveying a coreference assignment to an agent, even where the coreferring terms have identical senses, is potentially informative to that agent. This suggests that an important aspect of the semantics of a name is the coreference assignments into which it enters. Thus, (SC.1) has it that the semantic contributions of formally distinguishable names to internal representations are equivalent just in case each name enters into the same set of coreference assignments.

Now, in order to defend the semantic approach to the individuation of mental states from the "opacity objection" we require a systematic procedure for specifying a semantic difference between certain pairs of extensionally equivalent beliefs. The procedure which we suggest is quite simple:

Extensionally equivalent beliefs are semantically distinguishable in a system just in case that system internalizes no coreference assignment that determines the extensional equivalence of the beliefs. The idea is that the semantic contributions for the names '_a_' and '_b_' are not equivalent for an agent unless he represents the coreference of those names and, thus, the beliefs that '_a_ is _F_' and that '_b_ is _F_' belong to different semantic types just in case no such coreference assignment is internalized.

We have, thus far, limited application of the idea that internalized coreference assignments are important constituents of the semantics of referential terms to intrasubjective type relations among beliefs. This has been done only to simplify the presentation, by no means do we seek to relativize type-identity relations among token mental states to particular systems. Fortunately, (SC.1) is generalizable in such a way as to accommodate type relations between token mental states occurring in different systems.

(SC.2) For any two systems $\underline{S}_1$ and $\underline{S}_2$, token mental states which differ only with respect to the occurrence of coreferential terms, '_a_', '_b_', '_c_',..., are type-identical if and only if (a) $\underline{S}_1$ and $\underline{S}_2$ each internalize a coreference assignment between the relevant terms, and (b) every coreference assignment, involving any of the relevant terms, internalized by either $\underline{S}_1$ or $\underline{S}_2$ is also internalized by the other.

As one can easily appreciate, stating the appropriate conditions for the fully generalized case requires care. In the first place, we want to allow that coreference relations can hold between pairs of terms, triplets of terms, quadruplets of terms, and so on. Further, it is insufficient to demand that the agents, $\underline{S}_1$ and $\underline{S}_2$, internalize the coreference assignment '$\underline{a}$ = $\underline{b}$' in order to warrant the assignment of their respective beliefs '$\underline{a}$ is $\underline{F}$' and '$\underline{b}$ is $\underline{F}$' to the same type. For, one but not the other might also internalize '$\underline{a}$ = $\underline{b}$ = $\underline{c}$'. In particular, we do not want to maintain that Bob's belief that "The Morning Star is bright" is type-identical with Bill's belief that "Phosphorus is bright" when each agent represents it to himself that 'The Morning Star = Phosphorus' but when Bill, unlike Bob, also internalizes 'Phosphorus = Hesperus '. If we were to do so, then by the transitivity of identity we would get the type-identity of Bob's belief that "The Morning Star is bright" with Bill's belief that "Hesperus is bright" even though Bob has never acquired use of the term 'Hesperus '.

With (SC.2) in hand, we can point to a semantic distinction between certain token mental states which differ only with respect to the occurrence of coreferring terms and to a semantic distinction between certain token mental states of the same formal type. In both cases, the semantic distinction depends upon the coreference assignments internalized by the agents in which the mental tokens occur. For example, tokens of the syntactic or formal type '<u>a</u>' may be associated with different semantic properties in different systems insofar as the coreference assignments for '<u>a</u>' vary across different systems. These considerations make contact with so-called "conceptual role" or "inferential role" semantic theories [Field, 1977; Harman, 1982]. It may be enlightening to close this section with a comment on inferential role semantics. The idea behind such theories is that the inferential connections, and perhaps other conceptual connections, between a particular mental state <u>M</u> and all other mental states of the system in which <u>M</u> occurs are, in part, constitutive of the semantic content of <u>M</u>. Unfortunately these theories, with the single exception provided by Field, have not been sketched in any detail. But, Field acknowledges that his proposal for a semantics based on the subjective probabilities of beliefs will not allow type-identities between mental states to be defined across systems, i.e. for Field there can be no concept of the intersubjective type-identity of mental states [Field, p. 398]. For this reason the scheme that Field provides is unsuited to the task of individuating mental states in such a manner that mental state types may be cited in psychological generalizations. But suppose that the inferential role theorists are essentially correct at least in the observation that the inferential connections among beliefs constrain the semantic interpretation of any particular belief. If this is correct, then equivalent constraints on the semantic interpretation of any two belief tokens are imposed by inferential role semantics just in case the inferential roles of the two tokens are equivalent. But under what conditions are the two beliefs '_a_ is _F_' and '_b_ is _F_' inferentially equivalent? It seems very clear that a necessary condition for the inferential equivalence of this pair of beliefs is the _intersubstitutability_ of '_a_' and '_b_'. Significantly, '_a_' and '_b_' are intersubstitutable for a system $\underline{S_1}$ just in case $\underline{S_1}$ represents the coreference of '_a_' and '_b_'. If $\underline{S_1}$ does not represent the coreference of '_a_' and '_b_', then there will be beliefs that $\underline{S_1}$ can derive from '_a_ is _F_', in collaboration with other of its beliefs, that it cannot derive from '_b_ is _F_'. Moreover, the representation of the coreference assignment '_a_ = _b_' is, arguably, a sufficient condition for the equivalence of the inferential roles of '_a_ is _F_' and '_b_ is _F_' in $\underline{S_1}$. For given that coreference assignment, any belief derivable from one of this pair of tokens, in collaboration with other internalized beliefs, will be derivable from the other token. The inferential role theorist claims that beliefs are semantically equivalent just in case they share a common inferential role. Thus, if the above observations are correct, a necessary and sufficient condition for the equivalence of the inferential roles of beliefs in a system, which differ only with respect to the occurrence of coreferring terms, is the internalization of the same set of coreference assignments.

ASPECTS OF THE CONTENTS OF MENTAL STATES

In this section we will further generalize the conditions for the semantic equivalence of beliefs given by (SC.2). In order to provide some useful motivation for the proposed generalization, we will briefly consider two approaches to the interpretation of mental states which represent the extremes between which most, perhaps all, plausible theories of the contents of mental states lie. Essentially, one approach which we will consider has it that the totality of an agent's (consistent) beliefs associated with a term are constitutive of the concept which that term conveys. The other approach has it that sociolinguistic conventions which determine the correct application of a term are constitutive of the concept which that term conveys when it occurs in a mental representation.

We have already seen that the content of a belief token may depend, in part, on the internalized coreference assignments for the terms occurring in referential positions in that token. This view has the advantage of allowing us to distinguish certain coextensive beliefs _semantically_ without appealing, in any explicit way, to the senses of the referential terms invoked in those beliefs. This result, incidentally, is especially interesting since Fodor constructs an argument for the formality condition, and thus for the non-semantic individuation of mental states, upon the observation that attributions of mental states are opaque [see Ch. III, note #2]. On the view urged here, the content of a belief is sensitive to any coreference assignments involving terms invoked in the belief. This might suggest to some that the content of a representation is fixed by an agent's information about the objects designated by the referential terms which occur in the representation. We will refer to this view as the _subjectivist theory of content_ or semantic subjectivism. One mark of semantic subjectivism is that it takes the referent of a name occurring in a belief to be determined by the set of an agent's beliefs which invoke that name.

In order to illustrate the shortcomings of the subjectivist view we may adopt an example which T. Burge employs in a somewhat similar role [Burge, 1979]. The subjectivist holds that the extensions of referring terms are fixed by aspects of an agent's conceptual system. Hence, if Jones believes that "Smith entered into a contract," Jones' conceptual system must, _inter alia_, (a) assign the proper name 'Smith' to some unique individual, and (b) specify the extension of 'contract'. If we adopt the subjectivist viewpoint, we may assume, for example, that Jones' conceptual system fixes the extension of 'contract' by description. Suppose that Jones believes that all and only contracts jointly satisfy the conditions (i) a legal arrangement fixing an exchange of valued commodities, and (ii) a written document. The intersection of (i) and (ii) constitutes the "subjective extension" of 'contract' for Jones. Now, suppose that Jones sincerely avows the belief that "verbal contracts do not bind" or that "contracts must be written". Such beliefs are false if Jones is a member of a legal-linguistic community that uses 'contract' as we use that term. The subjectivist theory of content, however, must rule Jones' beliefs analytically true. If (i) and (ii) fix the extension of 'contract' relevant to the determination of the content of Jones' belief that "contracts must be written," then there is no basis for the evaluation of that belief as false.

The point illustrated by this example is obvious enough. A fully subjectivist theory of content makes the believer the arbiter of the truth of many of his beliefs. In the extreme, the problem is that if an agent's beliefs invoking the term '<u>a</u>' fix the agent's concept of <u>a</u>, then any new belief invoking '<u>a</u>' appended to the agent's belief set must be considered "true" by the subjectivist just in case the new belief is consistent with the original belief set. For if each of the beliefs of a system implicating a certain term conjoin to determine the concept associated with the term, then any consistent addition to a belief set in some domain, e.g. in the domain of "contract beliefs," is either entailed by the belief set or constitutive of a change in the concept that the original belief set determines. This is a coherence theory, and perhaps a holism, with a vengeance. Putnam has voiced a closely analogous concern with certain theories of content. Putnam suggests that an adequate theory of content requires a distinction between _concepts_ and _collateral information_. Putnam may be interpreted to hold that subjectivism is bound to conflate the notions of concepts and collateral information. Putnam asserts that:

> The reason that we cannot count every difference in collateral information we have as a difference in the meaning of a _word_, is that to do so abandons the distinctions between our "concepts" and what beliefs we have that contain those concepts, and just this distinction is the _basis_ of the intuitive notions of meaning, synonymy, analyticity, etc..."content" must remain stable under some changes of belief. [Putnam, forthcoming]

A subjectivist theory counts every belief in an agent's belief set implicating the symbol '_a_' as instrumental in determining the content conveyed by the symbol '_a_'. According to such a theory, then, every new belief not entailed by the original belief set constitutes a change in some concept. If Putnam is correct, what we have termed the subjectivist theory is no theory of content at all.

How might the problems confronting the subjectivist approach be overcome? One apparent alternative to semantic subjectivism lies at the other extreme and has been embraced by Burge and perhaps by Putnam -- it is, admittedly, difficult to tell what Putnam's position is. Burge, in any case, suggests that the concept of contract predominant in an agent's sociolinguistic community should be attributed to the agent when he forms a mental state invoking the term 'contract' [Burge, 1979, pp. 78-79; see Fodor, forthcoming]. We could term such a view the "socialist" theory of content had not that term been preempted by other forces. Whatever we call it--perhaps the *social theory of content*--it is easy to see that this view suffers serious defects. Fodor has pointed to one grave defect: If we refer only to social-legal-linguistic conventions when assigning contents to mental states, then we must rule the contents of many mental states self-contradictory [see Fodor, forthcoming]. Recall Jones' belief that "contracts must be written". For Burge it is true, by hypothesis if not in fact, that it is constitutive of our social-legal-linguistic concept of contracts that "contracts need not be written". But, if we proceed to interpret Jones' belief in accordance with this social rule, then the content of Jones' belief is roughly "legal agreements which need not be written must be written". Hence, making social conventions the sole arbiters of the contents of beliefs has the consequence of forcing self-contradictory interpretations of the contents of many beliefs. Notice that this situation is really quite severe. The social theory of content cannot hold merely that an agent unknowingly forms many self-contradictory beliefs, but must hold that the only contents which could be known to an agent are, in many cases, self-contradictory.

Advocates of the subjectivist theory of content and advocates of the social theory of content may attempt various manoeuvers to avoid the problems briefly rehearsed here. The crucial lesson provided by these illustrations is that both an unrefined subjectivist theory and an unrefined social theory of content will fall short of providing a stable basis for the interpretation of the contents of mental states. The problems encountered by either extreme suggest that a middle ground of some sort might be usefully occupied. On the one hand, it would appear that an agent's communicative or linguistic intentions contribute to the determination of the contents of his beliefs. On the other hand, it would appear that an agent's place in a community of speakers contributes to the determination of the contents of his beliefs. Ideally, we want a theory for the interpretation of beliefs which captures both kinds of contributions. A complete account of such a theory is obviously a major task. But an important part of any appropriate theory~~, no d~~oubt, is the recognition that in the typical case an agent's communicative intentions include the intention to conform to sociolinguistic rules. We envisage an empirical generalization which relates the fact that cognitive agents in a community of speakers intend their usages of terms to conform to the sociolinguistic standard in their community. Obviously, agents will not always succeed in acquiring the appropriate standard -- as in the case of Jones' contract beliefs. But insofar as an agent believes that his usage is correct, when it in fact is not, a belief of the form 'contracts must be written' need not be thought of as either analytically true or self-contradictory. We propose that it is, in part, constitutive of the content of such a belief that the agent who formulates it takes himself to be in conformity with the sociolinguistic rule for the term 'contract'. Thus, the analysis appropriate to 'contracts must be written' is something like "what are called 'contracts' around here must be written". The indexical phrase "around here" serves to fix the agent's sociolinguistic group. If we may proceed in this fashion, then the belief that Jones' forms is empirically false if it misrepresents the linguistic usage of Jones' group. To this proposal some philosophers will no doubt want to reply, "But Jones' belief is about contracts, not linguistic usage!" And, naturally, this observation is partially correct, but Jones may have a belief about contracts via his intention to refer to what members of his group designate when they use 'contract' correctly, i.e. when they conform to the appropriate sociolinguistic rule.

These remarks are admittedly sketchy ; we will return to them later. For now, it is important that we provide a characterization of the conditions for the equivalence of the contents of beliefs that is neither purely subjectivistic nor purely social. This may be considered a somewhat limited task when compared to the task of constructing a theory that explains the contributions to the contents of mental states made by an agent's own conceptual system and by the rules which structure his sociolinguistic community. But, even this limited task is a considerable one. Putnam, in particular, has claimed that there can be no notion of the equivalence of the contents of mental states adequate for the purposes of cognitive psychology. Indeed, Putnam has held the role of chief sceptic with respect to the concept of the contents of mental states since the presentation of his "Twin Earth" argument in "The Meaning of Meaning" [Putnam, 1975, pp. 215-271].

Before turning to a consideration of Putnam's arguments, we must do what was promised earlier, i.e. to generalize (SC.2) and the conditions it offers for the semantic equivalence of a certain class of belief tokens. Now, the criterion that will be offered will appear susceptible to Putnam-like counterexamples. The form that the new criterion will take is dictated only by considerations of clarity, and its apparent susceptibility to Putnam's arguments is a defect that will be remedied in the next section.

How are the contents of beliefs to be specified? A host of different views have been entertained by philosophers as answers to this question. For example, one might hold that the truth conditions of a belief, or perhaps those conditions *in virtue of which* a belief is true, determine its content [see Fodor, forthcoming b]. One might hold that the conditions of verification for a belief determine its content [see Putnam, forthcoming]. One might hold that the conceptual role of a belief determines its content. Or, one might try some combination of these views and hold that the truth conditions plus the inferential and/or conceptual role of a belief determine its content [Fodor, forthcoming b; Field, 1977]. This last alternative is perhaps the most promising, though our version of it here makes only a limited appeal to the idea of the "conceptual role" of a belief. One framework which is readily employable in the task of specifying the contents or meanings of mental states has been constructed by Field. Actually, the scheme to which we refer constitutes only one of two parts of Field's approach to the semantic equivalence of mental states. In any case, Field constructs what he calls a referential version of truth-theoretic semantics on the following paradigm:

> 'Beethoven lived in Germany' is true if and only if there are objects *x* and *y* and a relation *R* such that 'Beethoven' stands for *x*, 'Germany' stands for *y*, 'lived in' stands for *R*, and *x* bears *R* to *y*. [Field, 1977, p. 389]

The *referential truth conditions* for 'Beethoven lived in Germany' are given in accordance with the above schema *once the referents for x, y, and R are given*. Now this paradigm for the statement of the truth conditions of propositions is superior to the traditional formulation, in one way, since it exposes more structure than its "disquotational" counterpart: "'Beethoven lived in Germany' is true if and only if Beethoven lived in Germany". Additionally, Field's schema makes it clear that the content of a mental state depends upon its referential conditions. Notice also that the paradigm offered by Field could be given a kind of verificationist interpretation, i.e. the paradigm has a verificationist version.[9] For it would be possible on a certain kind of verificationist semantic theory to cleave to a version of the above paradigm by interpreting the relation of designation, i.e. "stands for", in some verificationist fashion. For example, the content of "'Beethoven' stands for *x*" might be given, following a suggestion of Putnam's in a different context, by a specification of the conditions under which "'Beethoven' stands for *x*" is acceptable or warrantedly assertable [Putnam, forthcoming]. Putnam does not actually say that the designation-relation is to be understood in terms of conditions for acceptability or warranted assertability, but he does appeal to these notions as paradigmatic verificationist concepts. Now if appeals to the referential conditions of beliefs are countenanced, at least in principle, by both truth-theoretic and verificationist semantic theories, then the idea underlying the appeal to the referential conditions of beliefs should be relatively uncontentious. According to this idea, the content of a mental state is, roughly, a function of the symbolic attribution of some designated property to some designated object. Thus, to a first approximation, representations that attribute the same property to the same object are semantically equivalent, i.e. representations with the same referential truth conditions. However, we have already effectively denied the sufficiency of this principle in asserting that internalized coreference assignments have an important role to play in the determination of the contents of mental states.

Similarly, Field holds that sharing the same referential truth conditions is a necessary but not sufficient condition for the synonomy of beliefs. Some representations with equivalent referential truth conditions are sometimes semantically distinct, e.g. 'Hesperus is Phosphorus' and

'Phosphorous is Phosphorus'. In order to resolve this problem Field appeals, as we have indicated, to the conceptual role of mental states. According to Field two arbitrary beliefs, A and B, internalized by an agent have different conceptual roles just in case there is some belief C in the agent's belief set such that the subjective probability of A given C is higher or lower than the subjective probability of B given C. Field suggests that the equivalence of conceptual roles, so understood, is a necessary condition for the semantic equivalence of beliefs. This approach tends to assume that a "probability function" is a psychological mechanism realized in all believers. Harman, incidentally, claims to have shown that this is unlikely [Harman, 1982, p. 247]. But, as mentioned earlier, Field's concept of the conceptual roles of beliefs is highly unattractive from the perspective of psychological theory because it is inapplicable across systems. The conceptual role of a belief, a la Field, can be determined only relative to an individual agent since different agents will have different probability functions [Field, 1977, p. 398].

In order to avoid this consequence of Field's approach, we have proposed that the contents of token beliefs depend, in part, upon the internalized coreference assignments for terms which occur in referential positions in those tokens. The semantic equivalence criterion suggested here has two components. We will adopt Field's notion of the referential truth conditions of beliefs, but replace Field's reference to the conceptual roles of beliefs with a reference to the mutual internalization of coreference assignments: According to this view, in essence, mental states have equivalent contents just in case (i) they have the same referential truth conditions, and (ii) the terms which occur in referential positions in those tokens are implicated in the same set of internalized coreference assignments. But we must take care in stating these conditions precisely.

> (SC.3) Token mental states occurring in any two systems, $\underline{S}_1$ and $\underline{S}_2$, are type-identical if and only if (a) (i) the tokens have the same referential truth conditions, and (ii) every coreference assignment internalized by either $\underline{S}_1$ or $\underline{S}_2$ for terms occurring in referential positions in those tokens is internalized by the other, and (b) if the tokens differ with respect to the occurrence of coreferential terms then $\underline{S}_1$ and $\underline{S}_2$ each internalize a coreference assignment between the terms with respect to which they differ.

Actually, the conditions given by (SC.3) are not as complex as they may appear. In condition (a) of (SC.3) it is necessary to require that referentially equivalent representations, even when they are formally identical, are associated with the same sets of coreference assignments. Otherwise, by the transitivity of identity, the token '$\underline{a}$ is $\underline{F}$' in $\underline{S}_1$ would be type-identical to the token '$\underline{b}$ is $\underline{F}$' in $\underline{S}_2$ when there is a system $\underline{S}_3$ that forms '$\underline{a}$ is $\underline{F}$' and $\underline{S}_2$ and $\underline{S}_3$

internalize '<u>a</u> = <u>b</u>'. Condition (b) of (SC.3) is constructed, on the model of (SC.2), to accommodate those cases in which tokens with the same referential truth conditions employ formally distinct referring terms. Thus, if a pair of tokens satisfies the two conditions of (a) of (SC.3), but they invoke different referring terms, then it is only necessary that the systems in which they occur each internalize a coreference assignment between the terms -- it is necessary to include this condition explicitly since $\underline{S}_1$ and $\underline{S}_2$ might satisfy (a)(ii) by internalizing no coreference assignments at all.

All token mental states which satisfy the conditions of (SC.3) are arguably type-identical. Moreover, (SC.3) offers <u>semantic</u> conditions for the type-identity of mental states and this is as it should be since it is extremely difficult to envisage adequate non-semantic criteria for the type-identity of mental states. Offering semantic conditions for the type-identity of mental states is highly advantageous since it gives us considerable room to manoeuver. We might just mention two areas in which this is possible. Some may object that the standards that (SC.3) sets for the type-identity of mental states are too high and the resulting "grain-size" is too small. Condition (ii) of (a) requires that the token beliefs of two agents are semantically equivalent only if the agents internalize precisely the same set of coreference assignment involving all the referring terms which occur in their respective belief tokens. We could modify this requirement in a number of ways. For example, we might maintain that $\underline{S}_1$'s belief is equivalent to $\underline{S}_2$'s belief only if neither agent internalizes an idiosyncratic coreference assignment for any referring term which occurs in their belief tokens. Though this approach is not recommended here, it illustrates the possibility of adjusting the grain size of (SC.3). Additionally, one might toy with the idea that an agent needn't actually internalize a coreference assignment, but only be disposed to accept a coreference assignment if it is conveyed to him. Such a "disposition to accept" a coreference assignment might be analyzed in terms of the "degree of belief revision" required by a system in order to accommodate the coreference assignment. Presumably, the problem of evaluating the semantic contents of the beliefs of members of different language groups might be treated in this way: If an agent must overturn a standing belief when he learns of a <u>translational</u> coreference assignment, e.g. London = <u>Londre</u>, then we might say that his disposition to accept the coreference assignment is too weak to warrant the type-identification of any belief of the agent implicating one of the terms with the belief of another agent implicating the other term.

These speculations may be worth pursuing, but we won't follow this course here. Instead, we will treat Putnam's objections to the concept of the equivalence of the contents of mental states. After all, (SC.3) purports to specify semantic conditions for the equivalence of mental states and this is something that Putnam claims cannot be done.

## PUTNAM'S PROBLEMS

One of Putnam's self-proclaimed problems is an inability to distinguish between elm trees and beech trees. Putnam maintains that this inability demonstrates that his concept of "elm" and his concept of "beech" are equivalent [1975, p. 226]. The lesson that Putnam intends this illustration to convey is that an agent's "concepts" do not fix the extensions of the terms he employs. For although Putnam's concept of elm is putatively the same as his concept of beech, the terms 'elm' and 'beech' have different extensions. However, this illustration fails to convey the lesson that Putnam intends. For, if we proceed according to the proposals of the last two sections, then none of Putnam's beliefs about elm trees are equivalent in content to his beliefs about beech trees. The criterion (SC.3), and (SC.1) for that matter, distinguishes between beliefs invoking 'elm', in a referential position, and beliefs likewise invoking 'beech' in all those cases in which the coreference of 'elm' and 'beech' is not represented by the system in which the beliefs occur. In fact, it is wildly implausible to hold that H.P.'s concept of elm is identical to H.P.'s concept of beech even though H.P. believes "that elm trees are not beech trees". Yet, this is just what Putnam asks us to believe, i.e. that his concept of "elm" is equivalent to his concept of "beech" even though he knows that 'elm' and 'beech' are non-coreferential. According to our proposal, a crucial part of Putnam's <u>concept of elm</u> is the belief that "beech trees are not elm trees". Quite simply, one cannot have the same concept of "elm" and "beech" unless one internalizes a coreference assignment for 'elm' and 'beech'.

This approach leaves us with the following question. What does Putnam believe when he thinks to himself "that's an elm tree"? We can distinguish this belief from "that's a beech tree", but Putnam's inability to distinguish elms from beeches does show that he does not possess criteria for determining if something is an elm or a beech. Earlier we suggested that an agent's intension to use terms in such a way as to conform to the appropriate sociolinguistic norms should be taken into consideration when evaluating the content of certain of the agent's mental states. This idea is particularly useful when we must evaluate the contents of mental states formed by non-experts. If we apply this idea in the present context, the <u>analysis</u> appropriate for Putnam's belief of the form 'that's an elm tree' is something like "that's a tree that experts include in the class of trees they call 'elm'". Of course, Putnam's belief is false just in case the appropriate experts would not so classify the tree to which Putnam refers. This approach seems all the more appropriate since Putnam actually advocates a special role for experts in a sociolinguistic community [1975, pp. 227-229].

All of Putnam's problems cannot be dealt with so easily. Putnam's most widely known argument which presents a challenge to the idea of the semantic equivalence of mental states draws the highly general conclusion, echoed by the elm-beech example, that mental states do not determine the meanings of the terms they employ. In the context in which Putnam's so-called "Twin Earth" argument is placed, reference to the Twin Earth example is made in support of one premise of a broad sweeping argument. Very briefly, it is Putnam's contention (a) that the extension of a term is a necessary constituent of the meaning of the term, and (b) that the extension of a term is not fixed by any aspect of the conceptual system of an agent who employs that term. Together, (a) and (b) are said to entail that (c) meanings are not determined by mental states, or in Putnam's more colorful language, that "'meanings' just ain't in the head" [Putnam, 1975, p. 227]. The notorious Twin Earth argument is enlisted in support of premise (b) -- extensions are not determined by mental states. This sketch of the context in which the Twin Earth argument is embedded leaves considerable room for interpretation. In particular, we must post a warning before actually rehearsing Putnam's piece of science fiction. Reading Putnam, one is often impressed by the apparent generality of the views which are presented. The idea that meanings are not in the head seems to be portrayed as a fully general thesis about the nature of mental states and meanings. But the argument actually given for a crucial premise supporting this thesis relies heavily upon illustrations that are slight variations on a single theme. Crucially, it is far from clear that the argument is generalizable in such a way as to lend the thesis about mental states the full generality that is apparently claimed for it. Putnam purports to show that in certain cases token mental states may share the same physical, neurological, and psychological properties even though the tokens are associated with different natural kinds. But, does it follow from this that cognitive systems are inherently incapable of internalizing information that would fix the extension of a kind term? As we will see, it will be helpful to keep this question in mind.

Now, in the Twin Earth thought-experiment Putnam asks us to imagine that there are two planets, Earth and Twin

Earth, which resemble each other in virtually every major and minor detail. The resemblance between the two planets extends even to the human inhabitants, and to the languages, of the two planets. In fact, we may assume that each resident of Earth has a Doppelgänger on Twin Earth, i.e. an exact physical duplicate. The single difference between Earth and Twin Earth is that, by hypothesis, water on Earth has the chemical formula $H_2O$ and what is called 'water' on Twin Earth has a different molecular structure given by the formula XYZ. But, $H_2O$ and XYZ resemble each other precisely in every macrophysical and every observable detail. The substance that is called 'water' on each planet fills all lakes and oceans, it quenchs thirst and fire, and it falls from the clouds in the form of rain and snow.

Now an individual on earth, Oscar, and his Doppelgänger on Twin Earth, who we will call Elmer for convenience, are in type-identical total states on every occasion. Thus, an exhaustive comparison of the token mental states of Oscar and Elmer at any time t will reveal absolutely no difference -- physical, chemical, neurological, computational, psychological, or intentional. Actually, this holds only for all times t prior to the divergence of the concepts of "water" acquired by Oscar and Elmer. Hence, Putnam asks us to consider a pair of mental states, internalized by Oscar and Elmer respectively, that invoke the term 'water' at a time, circa 1750, prior to the discovery of the molecular structure of what is called 'water' on each planet. Suppose that Oscar and Elmer internalize a representation of the form 'water is wet' in 1750. Though absolutely everything in the head of Oscar is in the head of Elmer, and vice versa, Putnam maintains that the extension of the term 'water' is different for each agent. Putnam puts this claim in an extreme and radical form:

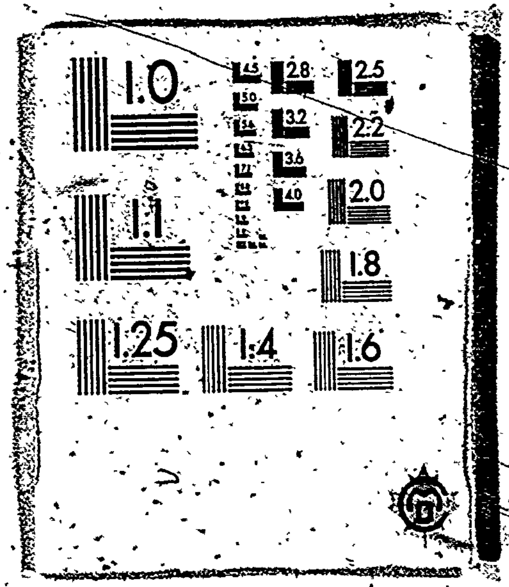
> [Oscar and Elmer] understood the term 'water' differently in 1750 although they were in the same psychological state, and although, given the state of science at the time, it would have taken their scientific communities about fifty years to discover that they understood the term 'water' differently. Thus the extension of the term 'water' (and, in fact, its meaning in the intuitive preanalytic usage of that term) is not a function of the psychological state of the speaker by itself. [Putnam, 1975, p. 224]

Putnam's way of putting his intuition is unduly misleading. Putnam asserts that Oscar and Elmer "understood" the term 'water' differently in 1750. This can hardly be the case if the notion of "understanding" is a genuine psychological concept and, by hypothesis, Oscar and Elmer are complete psychological copies of each other, i.e. if their conceptual schemes are completely equivalent. Furthermore, Putnam appeals to what he claims is the intuitive preanalytic usage of the term 'meaning' in order to motivate the suggestion that the meaning of the kind term 'water' is different for

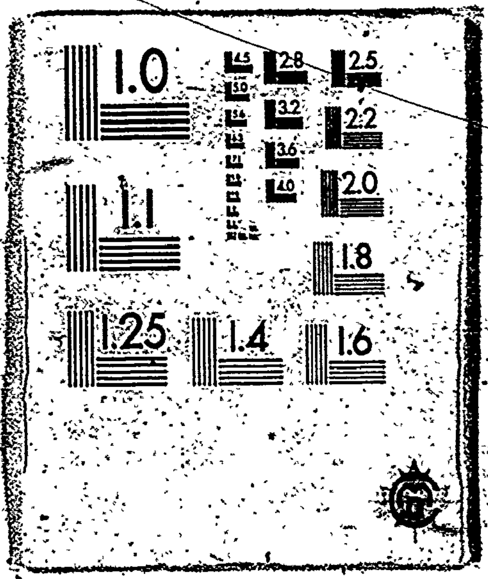
Oscar and Elmer. But, the sense of 'meaning' Putnam has in mind is just extension and it is far from obvious that there is anything preanalytic or intuitive about this usage of the term.

Now, we will shortly suggest an approach to the Twin Earth example, and its relatives, which specifies conditions for the equivalence of the concepts of agents who do not know the extensions of certain terms that they employ. But first, let us consider the degree of generality possessed by Putnam's conclusions that mental states don't fix extensions and that meanings are not in the head. The line of argumentation that Putnam employs is peculiar in one important way. First we are told that, by hypothesis, the extension of the term 'water' on Earth is $H_2O$, or more precisely "the set of all wholes consisting of $H_2O$ molecules", and that the extension of the term 'water' on Twin Earth is XYZ, or more precisely "the set of all wholes consisting of XYZ molecules" [Putnam, 1975, p. 224]. Yet, in characterizing the conclusion of the argument Putnam is prepared to say that "psychological state does not determine extension" [p. 223]. But if this conclusion is correct and fully generalizable, then it would appear that nothing knowable is conveyed by the proposition that "the extension of 'water' on Earth is all wholes consisting of $H_2O$ molecules", or by "water on Earth = $H_2O$". The propositions expressed

3



4 OF/DE 4



by these statements purport to specify the extension of 'water' as it is used on Earth. Moreover, it is crucial to Putnam's argument that he establish a specifiable difference between the extensions of 'water' on Earth and Twin Earth. But, if an agent's total mental or conceptual system cannot determine the extension of any kind term which he employs, then either "water on Earth = $H_2O$" does not fix the extension of 'water' on Earth, or "water on Earth = $H_2O$" is unknowable, perhaps unlearnable. For if Oscar were to learn "water on Earth = $H_2O$", he would learn something which specifies the extension of the Earth term 'water'. After Oscar learns the appropriate formula, the extension for 'water' is fixed by something that has become part of his conceptual system. Hence, it is appropriate that one hesitate to receive Putnam's conclusion as a fully general thesis about the capacity of mental states to fix the extensions of kind terms that agents employ. For unless Putnam is prepared to defend a thesis about the unlearnability of "water on Earth = $H_2O$", he must allow that the inculcation of this proposition in an agent puts the agent in a position to specify the extension of the term 'water' contained in the agent's language. One cannot hold (i) that the extensions of certain terms can be specified by expressions in our language, and (ii) that the conceptual systems of members of our language group are inherently incapable of specifying the extensions of those same terms. Putnam allows himself full recourse to the idea that the extensions of kind terms are specifiable and the theory that "water on Earth = $H_2O$" -- his thought-experiment won't fly without these assumptions. But, the conclusion of the argument is supposed to be that mental states don't determine extensions. What we have endeavoured to point out is that Putnam's argument does not demonstrate an inherent limitation on the capacity of mental states to specify the extensions of terms.

There is much more that we could say in connection with this point. For example, we might point out that it is a common occurrence for a neologism to be introduced whose extension is fixed by conventional means. This happens often in technical circles where some new term is needed for some new invention. Further, we might press Putnam for a full account of the place of *experts* in the sociolinguistic division of labor which he hypothesizes: Putnam asserts that our ability to use terms like 'elm' and 'aluminum' efficiently presupposes that there are some among us who possess "a way of recognizing elm trees and aluminum metal" [1975, p. 227]. An expert is simply someone who possesses criteria for the application of a certain term, e.g. someone who possesses a way to determine if something is an elm tree or if something is a piece of aluminum. According to Putnam's theory of the sociolinguistic division of labor, when there are experts who possess criteria for the application of a given term "the sociolinguistic state of the collective linguistic body...fixes the extension" [1975, p. 229]. Putnam stops short of maintaining that the criteria known to experts fix the extensions of technical terms.. But if the sociolinguistic state of a collective body can fix the extension of a technical term, and average speakers in the collective body merely rely upon what the experts know, then it appears that the appropriate experts must be in command of extension fixing information from time to time. If this is correct, then there are agents, the experts, whose mental states sometimes fix the extensions of technical terms, e.g. 'water', 'elm', 'aluminum'.

These points, together with the observation that Putnam's Twin Earth thought-experiment presupposes the specifiability of the extensions of 'water' on Earth and Twin Earth, strongly suggest that Putnam must allow that a mental state, is the kind of thing that can fix the extension of a term. Now, it is completely possible that Putnam would acknowledge these points; they are not intended as arguments against the claim that the extensions of some terms are sometimes left unspecified by the conceptual scheme of an agent who uses those terms. But, two morals can be drawn from these points. The first is that Putnam's apparently unqualified claim that mental states do not determine the extensions of the terms that they employ must be qualified if it is to be taken seriously. The second is that once Putnam's thesis is put in a qualified form, the problem that the Twin Earth argument poses for theories of the content of mental states is minimized. In particular, when the extensions of the terms occurring in a mental representation are fixed by an agent's conceptual scheme, we are free to evaluate the content of the representation, in part, by reference to its referential truth conditions. Thus, the Twin Earth argument poses no problem for the criterion proposed in the last section, (SC.3), in those cases in which the extensions of referring terms are fixed by the mental states of the agents who use those terms. In such cases a necessary condition for the semantic equivalence of mental states is the identity of their referential truth conditions. Furthermore, the same necessary condition must be imposed for beliefs of the form '_x_ is _F_' where _x_ is replaced by a singular term, and we may retain (SC.3), unchanged, for those cases as well.

However, Putnam's argument illustrates the need for a concept of the equivalence of the contents of mental states applicable to token states which invoke kind terms whose extensions are not specified by the system in which they occur. One approach to this problem consists in maintaining that the difference between the extensions of 'water' on Earth and Twin Earth is completely irrelevant to matters of content: Since Oscar and Elmer share every psychological property, the fact that what is called 'water' on Earth and Twin Earth belong to different natural kinds must be irrelevant to the evaluation of the content of their mental states. The idea is to "factor away" any difference associated with mental states invoking the term 'water' on Earth and Twin Earth [see Pylyshyn, 1980, p. 159]. Putnam holds that factoring away the difference between the extensions of 'water' on Earth and Twin Earth is to factor away the reference of Oscar's and Elmer's beliefs. The reference of a mental state is, however, a crucial constituent of its content -- here we are in agreement with Putnam. Thus, Putnam is diametrically opposed to the maneuver of factoring away differences in extensions. Witness the following passage in which Putnam comments on an example closely analogous to the Twin Earth example:

> Once we decide to put the reference (or rather the difference in reference) aside, and to ask whether [a representation] has the same "content" in the minds of Oscar and Elmer, we have embarked upon an impossible task. Far from making it easier for ourselves to decide whether the representations are synonymous, we have made it impossible. In fact, the first approximation we have to a principle for deciding whether words have the same meaning or not in actual translation practice is to look at the extensions. "Factoring out" differences in extensions will only make a principled decision on when there has been a change in meaning totally impossible. [Putnam, forthcoming]

There are several points to take note of in this passage. First, Putnam ~~apparently~~ accepts the practice, which we adopted in the last section, of evaluating the content of representations, in part, in terms of their referential conditions -- "the first approximation we have to a principle...". Second, Putnam is convinced that there can be no notion of the content of Oscar's and Elmer's beliefs about "water" that does not advert to different extensions for those beliefs -- "Once we decide to put the reference aside...we have embarked upon an impossible task." However, there may be a notion of the contents, or perhaps of the reference, of Oscar's and Elmer's beliefs such that the referential conditions of their beliefs are equivalent. If this is correct, then we needn't factor away the reference of beliefs when ignoring a difference between the natural kinds to which they are related. In order to motivate such a conception of the contents of beliefs, let us first consider Putnam's own predicament. If the contents of Oscar's and Elmer's beliefs can be specified only by reference to different extensions, then their beliefs differ in content. That is, there is a specifiable difference between the mental states of Oscar and Elmer when their mental representations invoke the term 'water'. But, if there is such a difference between the mental states of Oscar and Elmer, then the thesis of the supervenience of mental states upon their physical realizations must be rejected. According to the supervenience thesis, a specifiable difference between mental states requires a specifiable difference between their physical realizations. In this form, the supervenience thesis serves as a fundamental constraint on the construction of theories of mental states. Putnam must abandon the concept of psychological supervenience if he is to insist that the fact that what is called 'water' on Earth and Twin Earth belong to different natural kinds establishes a difference between mental states which invoke the term 'water' on the two planets. Oscar and Elmer are, recall, a pair of physical Doppelgängers. Now, Putnam may pose his problem for a theory of the content of mental states by relating examples, such as the "grug" example to be discussed later, which do not refer to physical Doppelgängers. But, as long as the Twin Earth case provides a coherent example, it can always be used to implicate Putnam in the rejection of the supervenience thesis.

Preservation of the supervenience thesis as a constraint on the construction of theories of mental states requires that we view the mental states of Oscar and Elmer as fully type-identical and, hence, it requires a treatment of the contents of their mental states according to which their respective concepts of "water" are equivalent. Putnam's contention is that no such treatment is possible. But suppose that we cleave religiously to Putnam's conclusion that the psychological state of Oscar, for example, does not determine the extension of 'water'. Doing so, we have a very good reason for ignoring the fact that "water on Earth is $H_2O$" when offering a construal of Oscar's mental states about "water" in 1750. We can put this observation in the following form. If mental states and the conceptual systems in which they occur do not determine extensions, then the extension putatively associated with Oscar's belief that "water is wet" *transcends* Oscar's total conceptual scheme. Thus, we may characterize the notion of the *transcendental extension* of a kind term: The transcendental extension of a kind term '*w*' is formed by all things *x* such that *x* bears the relation *K* (same essence or same kind) to an exemplar of *w*. On this construal, everything that has the molecular structure represented by '$H_2O$' lies within the transcendental extension of 'water' -- as the term is used on earth. Let us grant that every mental state that invokes a kind term in a referential position is associated with a transcendental extension which may not be fixed by the total conceptual scheme of the agent who forms the mental state.

Now, Putnam must hold either that transcendental extensions are knowable or that transcendental extensions are unknowable. This point reiterates our earlier observations concerning the generality accreditable to Putnam's thesis that mental states don't determine extensions. If Putnam claims that transcendental extensions are unknowable, then his doctrine is simply a species of scepticism and, as we suggested earlier, a crucial assumption in the Twin Earth argument is simply unintelligible, e.g. "water on Earth = $H_2O$". Hence, Putnam must allow that the transcendental extensions of kind terms, for at least some such terms, are knowable. In those cases in which agents know the extensions of kind terms which they employ, e.g. Oscar and Elmer in 1750, we may determine the equivalence or non-equivalence of the contents of their beliefs according to the criterion, (SC.3), presented above.

What is required, then, is nothing more nor less than a treatment of those cases in which agents do not know the transcendental extension of a term that they use. This is precisely the case to which all of Putnam's examples speak -- including the Twin Earth, 'grug', elm-beech, aluminum-molybdenum, and arctic grass examples. We suggest the following rule for the type-identity of mental tokens which invoke terms whose transcendental extensions are not fixed by an agent's conceptual system: Token mental states which differ only with respect to their transcendental extensions are type-identical. An important advantage of this principle is that it allows for the preservation of the thesis of the supervenience of mental states upon their physical realizations.

Since we are committed to the type-identity of Oscar's and Elmer's mental states, we are committed to the specification of equivalent concepts of "water" for Oscar and Elmer. In order to fulfill this commitment we may profitably build upon the notion of a transcendental extension. Consider the following analysis of '*w* is *F*' where '*w*' is a kind term:

(∀*x*) (if *x* is *K* to *e*, then x is F).

This construal of '*w* is *F*' says simply that everything which bears the relation *K*, same kind, to an exemplar *e* (that falls under the term '*w*') is *F*. This analysis needn't be considered exhaustive or complete and it is offered only in an attempt to expose more of the structure of '*w* is *F*'. Putnam actually suggests a construal like this when he sketches the logic of natural kind terms [1975, pp. 224-225]. The idea is that given a sample of water on Earth, the extension of 'water' is *formed* by everything that bears the relation *K* to the sample. Further, the extension of the Earth term 'water' is cognitively *fixed* if and only if (i) a sample of water *e* is identified, and (ii) the relevant parameters of the relation *K* are specified.

Consider an agent whose knowledge fails to fix the transcendental extension of the term 'water'. There are still two variables relevant to the determination of the *concept of "water"* possessed by such an agent. One variable concerns the range of items the agent takes to be samples of "water". The other variable concerns the individual's beliefs, or information, about the parameters relevant to the relation K for water. The two variables are interrelated. An agent might, for example, deny that a piece of ice is an exemplar of water because she believes that a parameter relevant to K is liquidity. Given this basic framework, it is easy to say when two agents have the same concept of the substance they refer to by use of the term 'water':

> (SC.4) The concept of a kind w possessed by $S_1$ and $S_2$ is equivalent if and only if (a) every item acceptable to either $S_1$ or $S_2$ as an exemplar falling under 'w' is acceptable to the other, (b) every commitment of either $S_1$ or $S_2$ concerning the relation K, for the items acceptable as exemplars, is a commitment shared by the other, and (c) every coreference assignment internalized by either $S_1$ or $S_2$ for 'w' is internalized by the other.

The content of 'water is wet' for Oscar and Elmer is the same just in case, for a fixed conception of "wetness", Oscar's and Elmer's concepts of "water" satisfy the conditions of (SC.4).[8]

There are many points that we could make in connection with this approach to the equivalence of concepts. First, note that an agent whose knowledge does not fix the extension of 'water' can be thought of, under certain conditions, as formulating a theory of water. Oscar may believe, in 1750, that he can successfully ostend a sample of water and that he possesses a full understanding of the relevant parameters of K for the sample. Unfortunately, Oscar's theory of water in 1750 fails to specify a class coextensive with the transcendental extension of 'water' on Earth. Oscar's theory will define K in such a way that all items in the transcendental extension of 'water' on Twin Earth will be included in the theoretical extension he specifies for his term 'water' -- similarly for Elmer's theory. That is, Oscar's and Elmer's theory of water -- they have the same one -- specify a class of items that includes what is called 'water' on each other's planet. In 1750 their common theory defines the relation K in terms of parameters of similarity that correspond to superficial, or non-essential, properties of samples of "water" on each planet. Nevertheless, Oscar and Elmer may share a complete, though incorrect, theory of "water" because they fix specific, identical values for the two variables which are relevant to the specification of the extension of the term 'water' on each planet; i.e. e and K: Oscar and Elmer accept the same items as exemplars of "water" and they define the similarity relation K in the same way.

This observation is important. It allows us to show that when we assign truth values to certain of Oscar's and

Elmer's beliefs about "water" we cannot advert to the extensions of 'things that are $H_2O$' and 'things that are XYZ' respectively. For, suppose that, with their theory of "water" in hand, Oscar and Elmer predicate a certain property of the class of items specified by the theory. Suppose, in particular, that Oscar and Elmer predicate "kindhood" of this class of items. For example, Oscar formulates the belief that "water is a natural kind". Is Oscar's belief true or false? Following Putnam's course, since "water on Earth is $H_2O$", and the extension of a term is a necessary constituent of its meaning, Oscar's belief that "water is a natural kind" is a true belief! But, clearly Oscar has made a mistake: He includes in the theoretical extension of 'water' certain substances, e.g. those composed of XYZ, that do not share the essence of water on Earth. If a specification of the referential truth conditions of Oscar's 1750s belief that "water is a natural kind" includes the clause "'water' stands for $H_2O$", then Oscar's belief is true. But in 1750 Oscar has not succeeded in specifying a natural kind. In fact, he has succeeded in specifying a class of items that, we now know, does not form a natural kind. The only way to capture Oscar's mistake, and to see that his belief "water is a natural kind" lies within the extension of 'false', is to fix the extension of 'water' according to Oscar's theory.

Notice also that if we allow the theory of water formulated by Oscar and Elmer to specify the extension relevant to the determination of the contents of their thoughts, we can allow for changes in their concepts of water brought about by successive refinements of their theories. For eventually, Oscar's and Elmer's concepts of water will diverge as their sciences converge on theories which specify theoretical extensions coextensive with the transcendental extensions of 'water' on Earth and Twin Earth. This allows for the preservation of the realist doctrine advocated by Putnam when he contends that the extension of the Earth term 'water' never changes. In our terms, the transcendental extension of a kind term never changes -- unless nature is inconstant -- though the theoretical extension associated with the term may change through successive approximations to the truth.

Furthermore, this approach to the contents of mental states which do not fix the transcendental extensions of terms they invoke allows us to occupy a middle ground between a purely subjectivist theory of content and a purely social theory of content. Recall that a subjectivist theory holds that the extension of a term invoked in a belief is determined by <u>the set</u> of an agent's beliefs which invoke that term. We have suggested that an agent's <u>theory</u> associated with a term serves this role. Thus, only an agent's commitments to a class of exemplars and the relation K, rather than his total set of domain-specific beliefs, fix the extension of a kind term relevant to the determination of the content of his belief. There is one qualification that ought to be entered at this point. In some cases an agent will use a term for which he has no full theory. For example, Elmira acquires use of the term 'water', but Elmira is no expert and relies upon Elmer to theorize about water. For cases of this type, we may assume that Elmira intends her use of 'water' to conform to the criteria formulated by the experts. Thus, if she thinks that "Elmer drinks a lot of water", and Elmer has slyly conditioned Elmira to refer to gin as 'water', Elmira's belief is false if the experts of their community don't include samples of gin within the extension of 'water'. This point makes use of Putnam's notion of a sociolinguistic division of labor and points out the need to scrutinize the credentials of the alleged experts upon which one relies -- for one may fall into false beliefs by relying on the wrong individuals as arbiters of the application of a term.

At this point it may be helpful to briefly work through another of Putnam's examples in order to more fully illustrate the approach proposed here. Putnam asks us to imagine a country called Ruritania in which the extension of a common term 'grug' differs in the northern and southern dialects of the country. In the north of Ruritania what is referred to as 'grug' is silver, while in the south of Ruritania what is referred to as 'grug' is aluminum. But what is called 'grug' in each part of Ruritania is used for many of the same purposes. In particular, it is used to make pots and pans. Two children, Oscar and Elmer, grow up in Ruritania, one in the north and one in the south. Putnam contends that for a time the mental representations, or concepts, of "grug" formed by Oscar and Elmer will be indistinguishable. Eventually, however, Oscar's and Elmer's concepts of "grug" will diverge. Oscar will learn that "grug = silver" while Elmer learns that "grug = aluminum". Putnam says that "each of them will know many facts which serve to distinguish silver from aluminum, and "grug" in the South Ruritanian sense from "grug" in the North Ruritanian sense" [Putnam, forthcoming].

The challenge of the Ruritanian example for the advocate of the notion of the equivalence of the contents of mental states is to account for the change in Oscar's and Elmer's concepts. Tactically, Putnam wants to construct a case in which specifying the difference in the concepts of 'grug' possessed by Oscar and Elmer in the final state requires reference to the different extensions of the north and south version of 'grug'. Putnam presses the claim, against Pylyshyn in particular, that since the difference in the final state concepts is specifiable only by reference to the extensions of Oscar's and Elmer's concepts, it will not do to "factor away" such differences in extension. But, we have not advocated the total factoring away of aspects of the reference of mental states. The most natural thing to say about the Ruritanian example on the present proposal is the following. Oscar and Elmer share the same concept of "grug" as long as their mini-theories of "grug", i.e. the values they fix for e and K, are the same. But, as the theoretical extension of 'grug' specified by Oscar begins to converge on "silver", i.e. on the transcendental extension of "grug'-northern version, Oscar's concept begins to diverge from Elmer's concept. At the end of this process, each will internalize different coreference assignments for 'grug', e.g. Oscar believes "grug = silver" and Elmer believes "grug = aluminum". Notice that Oscar's and Elmer's early concepts of "grug" are equivalent but incorrect: When Oscar believes that "they use grug for pots and pans all across Ruritania", his use of the term 'grug' does not conform to the standards set by the northern experts.

A synopsis of the main principles utilized in our treatment of Putnam's problems might be given succinctly as follows -- in no particular order: The transcendental extension of a term 'w' is irrelevant to the content of a mental representation in which '<u>w</u>' occurs unless the theoretical extension associated with '<u>w</u>' converges on the transcendental extension of '<u>w</u>'. Further, the content of a mental representation, <u>M</u>, in which a referring term '<u>w</u>' occurs is determined, in part, by any and all coreference assignments for '<u>w</u>' internalized by the system which forms <u>M</u>. And finally, whether or not the transcendental extension of a term '<u>w</u>' is fixed by an agent's sociolinguistic system, in the typical case the agent intends his usage of '<u>w</u>' to conform to the sociolinguistic standards for the application of '<u>w</u>', to an object or a class, set by his community. Obviously, the status of these three principles may be different. For example, the point about an agent's intentions to conform to the standards of his sociolinguistic group may best be thought of as an empirical hypothesis. The other two points may be variously interpreted as aspects of the analysis of the <u>concept</u> of the content of mental states or, if you prefer, as proposals underwriting a certain program for the explanation and interpretation of mental states. However they are to be understood, we submit that the resources needed to markedly assuage, and hopefully to eliminate, the problems that Putnam poses for the idea of the semantic equivalence of mental states are plausible in their own right. Moreover, with the possible exception of the notion of the transcendental extension of a term, which after all is just Putnam's idea of the <u>extension</u> of a kind term, the resources to which we have appealed in treating Putnam's problems are almost mundane or routine philosophical observations. This, of course, is an absolute boon of the approach advocated here.

Unfortunately, it is not possible to offer a brief and uncomplicated statement of the necessary and sufficient conditions for the semantic equivalence of any arbitrary pair of belief tokens. But, this is to be expected and it should not, in itself, cause us concern. What would be cause for concern is nothing less than a pair of beliefs that resist all attempts of analysis, but this Putnam has failed to provide. Essentially, there are two kinds of cases: First there are those cases in which the (transcendental) extensions of terms are fixed by the conceptual schemes of the agents who invoke those terms. In this class we may include both singular terms, although they are not thought of as possessing "transcendental" extensions, and kind terms whose extensions are known to an agent. Secondly, there are all the rest, i.e. those cases in which an agent employs a term whose transcendental extension is unspecified by his conceptual scheme. For cases of the former type, i.e. where the actual extensions of terms invoked in beliefs are specified, the semantic equivalence of beliefs is determined by (i) the equivalence of the referential truth conditions of the beliefs, and (ii) the equivalence of the coreference assignments internalized for the referential terms invoked in the beliefs. For cases of the latter type, i.e. where transcendental extensions are unspecified, the semantic equivalence of beliefs is determined by (i) the equivalence of the theoretically specified referential truth conditions of the beliefs, and (ii) the equivalence of the coreference assignments internalized by the agents for those terms. In cases of this latter type, we have suggested that the referential truth conditions of a belief might be determined in accordance with the agent's theory of the items falling under the term whose transcendental extension is unspecified. But, it is not absolutely mandatory that we do so in the restricted sense that we can always determine the equivalence of the concepts of w, for an arbitrary w, formed by two agents by evaluating the equivalence or non-equivalence of their theories of w, i.e. according to (SC.4).

## FOOTNOTES - CHAPTER III

1. Ramsey-sentence functionalism is a doctrine due, in one form, to D. Lewis [in Block, 1980, pp. 205-215]. It sometimes goes unnoticed, though it is obvious, in talk of Ramsey-sentence functionalism that the Ramsey-sentence of a theory _T_ is a _functional theory_ only if _T_ functionally defines its theoretical terms, or only if _T_ causally individuates the theoretical entities it names. Thus, all questions concerning causal specifications of mental states are logically prior to any questions concerning Ramsification _per se_. When it comes to the crunch, Lewis does not demand exclusively causal specifications of mental states. Lewis maintains that mental states are to be individuated by reference to all shared platitudes regarding the causal relations of mental states, but then allows the addition of non-causal platitudes such as 'Toothache is a kind of pain'. These latter platitudes are, presumably definitional or analytic. The addition of definitional platitudes, arguably, implicates Lewis in a kind of semantic approach to the individuation of mental states.

2. There are two lines of argumentation for the formality condition constructed by Fodor. One more nearly resembles a theme than an argument and is predicated upon the observation that non-formal properties, i.e. semantic properties, are causally inefficacious and should not be used to distinguish between mental states. We do not, for example, distinguish between the mental states of two agents who believe that "the moon is full" even though their respective beliefs may have different truth values. These considerations are well taken, but they do not show that individuating mental states in a way sensitive to any of their semantic properties will prohibit theory from capturing the causal properties of mental states.

Fodor's other argument has to do with the need to respect the opacity of mental state attributions. Essentially, Fodor observes that the formality condition allows only opaque attributions of mental states and that, since opaque attributions are mandatory, this is a point in its favor. Again, the lesson is well taken, but it does not follow that mental state attributions which respect semantic properties are necessarily non-opaque. It is a simple matter to conjoin semantic criteria to formal criteria for the type-identity of representations and, thus, individuate mental states in a semantically sensitive, yet opaque, fashion. For example, 'Representations are type-identical iff (a) they are formally indistinguishable and (b) they refer to the same object.

3. The supervenience thesis has been treated as a doctrine about the relation of mental states to their physical realizations by J. Kim [1982, pp. 51-70]. Kim takes the supervenience thesis to be a doctrine in need of an argument. However, the considerations to which Kim appeals in support of the thesis are systematically richer than the thesis itself. For example, at one point in an argument for physical supervenience Kim appeals to a version of functionalism [p. 67]. I think that we might as well simply take the physical-supervenience thesis to be a (defeasible) constraint on construction of theories of mental states.

4. An acknowledgement concerning the idea of causal theories of mental state types is in order before launching our attack on the causal-functional construal of mental states. One theory for the reference of names has it that names refer to the dominant causal source of the set of an agent's beliefs implicating the name [G. Evans, 1977]. Suppose that a causal theory for not only the reference of names, but for the reference of predicates were available. One can, it appears, imagine a theory which took the meanings of a class of expressions to be a function of the combinatorial organization of items themselves causally interpreted. Given such a theory, the theorist would be in a position to assign semantic interpretations to at least some beliefs by reference to the causal relations that the appropriate causal theories specify. Though what "the right sort of causal connections" are is largely a mystery, a causal theory of semantic content appears to be at least a logical possibility, insofar as we can envision a restricted language which admits of a causal semantics. The point that we wish to acknowledge is that even if beliefs cannot be individuated in terms of their causal roles, it may still be the case that the contents of beliefs are specifiable by some causal semantic theory. This possibility is neither denied nor assumed by the arguments to follow. But, a salient observation is that the causal roles of beliefs are not constitutive of "the right sort of causal connections" with respect to which the contents of beliefs might be specified--that is, causal roles are not the right sorts of causal connections if the considerations to follow in the text succeed in undermining the possibility of causal-role construals of belief types.

5. Notice that Shoemaker completely overlooks the requirement that observational terms must appear in the Ramsey-sentence of a psychological theory. The way that Shoemaker has it, '$\exists F_1 \ldots F_N [T(F_1 \ldots F_N)]$', not only are all theoretical terms replaced by bound variables, but observational terms are simply dropped. Hence, (1) and (2) cannot be the correct formulations of the respective Ramsey-sentences. I am unable to diagnose the reason for this oversight.

6. (SC.1) has been stated in a simplified form. Technically, the antecedent of the biconditional should be qualified in the following way: 'Token beliefs which differ only with respect to the occurrence of the coreferential terms 'a' and 'b', which occur in referential positions, are type identical...'. This qualification changes nothing important to the discussion and is necessary since we will not type-identify the beliefs that "Bob said that 'a is F'", and "Bob said that 'b is F'".

7. We have relied upon the simplicity of the examples 'a is F' and 'b is F' in order to formulate relatively uncomplicated criteria for the semantic equivalence of extensionally equivalent beliefs. The symbols 'a' and 'b' are, typically, taken to be names in the discussion. However, one may also employ descriptions in mental representations. But two representations which differ only with respect to the occurrence of different descriptions should not be construed as representations which differ only with respect to the occurrence of coreferential terms. Russell apparently held that descriptions have no analysis outside the context of a closed sentence. Thus,

(1) The author of Waverly is Scott.
and,
(2) The author of Guy Mannering is Scott
have different analyses. The analysis for (1) is as follows.

(1*) There is at least one individual who wrote Waverly; and if x wrote Waverly and y wrote Waverly, then x=y; and Scott wrote Waverly.

In (1*) reference is made to the novel Waverly. The analogous analysis for (2) will make reference to the novel Guy Mannering. Thus representations which differ with respect to descriptions which are "satisfied by the same individual" needn't be thought of as representations which differ only with respect to the occurrence of coreferential terms.

A similar move should be made for putatively coreferential "terms" like 'creature with a heart' and 'creature with a kidney'. For example,

(3) Every creature with a heart is F.
(4) Every creature with a kidney is F.

The appropriate analysis for (3) is as follows.

(3*) (∀x) (if x has a heart, then x is F).

The point here is that the analogous analysis for (4) will place the term 'kidney' in a referring position and, hence, (3) and (4) needn't be thought of as representations which differ only with respect to the occurrence of coreferential terms.

8. We have simplified our discussion of the concept of the semantic equivalence of mental states by avoiding the problem of the equivalence of the concepts of properties possessed by cognitive agents. However, it would appear that the approach suggested here to the equivalence of concepts of natural kinds is generalizable to the case of properties.

# BIBLIOGRAPHY

Boden, M.A. [1970] "Intentionality and Physical Systems", Philosophy of Science, V. 37, N. 2, June, pp. 200-214.

__________ [1972] Purposive Explanation in Psychology, Cambridge, Mass.: Harvard University Press.

Brentano, F. [1874] Psychologie vom empirischen Standpunkt.

Burge, T. [1979] "Individualism and the Mental", Midwest Studies in Philosophy, V. 4, pp. 73-121.

Cherniak, C. [1981] "Minimal Rationality", Mind, V. XC, pp. 161-183.

Chisholm, R. [1967] "Intentionality", The Encyclopedia of Philosophy, V. 4, New York, N.Y.: Macmillan.

Chomsky, N. [1965] Aspects of the Theory of Syntax, Cambridge, Mass.: The MIT Press.

Cohen, L.J. [1981] "Can Human Irrationality be Experimentally Demonstrated?", The Behavioral and Brain Sciences, V. 4, N. 3, pp. 317-370.

Davidson, D. [1980] "The Material Mind", Ch. 13, Essays on Actions and Events, Oxford: Oxford University Press.

Dennett, D. [1978] Brainstorms, Montgomery, Vt.: Bradford Books.

__________ [1979] "True Believers: The Intentional Stance and Why It Works", to appear in a volume of the 1979 Herbert Spencer Lectures on Scientific Explanation, Oxford University Press.

__________ [1981a] "Three Kinds of Intentional Psychology", to appear in Reduction, Time, and Reality, Cambridge University Press.

__________ [1981b] "Making Sense of Ourselves", Philosophical Topics, V. 12, No. 1, Spring, pp. 63-81.

Dreyfus, R. [1978] "Cognitivism vs. Hermeneutics", The Behavioral and Brain Sciences, V. 1, No. 2, pp. 233-234.

Evans, G. [1977] "The Causal Theory of Names", in Naming, Necessity and Natural Kinds, Schwartz, S.P., ed., Ithaca, N.Y.: Cornell University Press, pp. 192-215.

Field, H. [1972] "Tarski's Theory of Truth", The Journal of Philosophy, V. LXIX, N. 13, July 13, pp. 347-375.

_________ [1977] "Logic, Meaning, and Conceptual Role", The Journal of Philosophy, V. LXXIV, N. 7, July, pp. 379-409.

_________ [1978] "Mental Representation", Erkenntnis, 13, pp. 9-61.

Fodor, J. [1975] The Language of Thought, New York, N.Y.: Thomas Y. Crowell Company.

_________ [1980] "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", The Behavioral and Brain Sciences, V. 3, N. 1, pp. 63-109.

_________ [1981] Representations, Cambridge, Mass.: Bradford Books/The MIT Press.

_________ [forthcoming] "Cognitive Science and the Twin Earth Problem", to appear in Notre Dame Journal of Formal Logic, April 1982.

_________ [forthcoming b] "Psychosemantics, or: Where Do Truth Conditions Come From?".

Geach, P. and Black, M. [1952] Translations From the Philosophical Writings of Gottlob Frege, third edition, Oxford: Basil Blackwell.

Harman, G. [1982] "Conceptual Role Semantics", Notre Dame Journal of Formal Logic, V. 23, N. 2, April, pp. 242-256.

Haugeland, J. [1978] "The Nature and Plausibility of Cognitivism", The Behavioral and Brain Sciences, V. 1, N. 2, pp. 215-260.

_____________ [forthcoming] "The Mother of Intention", submitted to Cognition and Brain Theory.

Kahneman, D. and Tversky, A. [1982] "The Psychology of Preferences", Scientific American, V. 246, N. 1, January, pp. 160-173.

Kim, J. [1982] "Psychological Supervenience", Philosophical Studies, 41, pp. 51-70.

Lewis, D. [1980] "Psychological and Theoretical Identifications", in Readings in the Philosophy of Psychology, Block, N., ed., Cambridge, Mass.: Harvard University Press, pp. 207-215.

Marras, A. [forthcoming] "Intentionality Revisited", to appear in an issue of Philosophia.

Putnam, H. [1975] Mind, Language, and Reality, Philosophical Papers, Volume 2, Cambridge: Cambridge University Press.

__________ [forthcoming] "Computational Psychology and Interpretation Theory".

Pylyshyn, Z. [1980] "Computation and Cognition: Issues in the Foundations of Cognitive Science", The Behavioral and Brain Sciences, V. 3, N. 1, pp. 111-169.

__________. [forthcoming] Computation and Cognition, Cambridge, Mass.: Bradford Books/The MIT Press.

Quine, W.V.O. [1960] Word and Object, Cambridge, Mass.: The MIT Press.

Searle, J.R. [1967] "Proper Names and Descriptions", The Encyclopedia of Philosophy, V. 6, New York, N.Y.: Macmillan.

Shoemaker, S. [1981] "Some Varieties of Functionalism", Philosophical Topics, V. 12, N. 1, Spring, pp. 93-119.

Simon, H. [1969] The Sciences of The Artificial, Cambridge, Mass.: The MIT Press.

Stich, S. [1980] "Computation Without Representation", The Behavioral and Brain Sciences, V. 3, N. 1, p. 152.

__________ [1981] "Dennett on Intentional Systems", Philosophical Topics, V. 12, N. 1, Spring, pp. 39-62.

__________ [1981b] "Inferential Competence: Right You Are, If You Think You Are", The Behavioral and Brain Sciences, V. 4, N. 3, pp. 353-354.

END

14·01·83

FIN