THREE STUDIES IN NATURALIZED PHILOSOPHICAL PSYCHOLOGY

by

Lawrence Jeffrey Kaye

M.A., Philosophy
University of Wisconsin-Milwaukee
(1984)

B.A., Psychology, Philosophy
University of Wisconsin-Milwaukee
(1982)

Submitted to the Department of
Linguistics and Philosophy
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

at the

Massachusetts Institute of Technology

April, 1990

Signature of Author.....
Department of Linguistics and Philosophy
April, 1990

Certified by.....
Robert C. Stalnaker
Professor, Linguistics and Philosophy
Thesis Supervisor

Accepted by.....
Ned J. Block
Chairman, Department of Linguistics and Philosophy

THREE STUDIES IN NATURALIZED PHILOSOPHICAL PSYCHOLOGY

by

Lawrence Jeffrey Kaye

Submitted to the Department of Linguistics and Philosophy
April, 1990 in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy

ABSTRACT

The dissertation consists of three distinct essays. Each is concerned with the implications that scientific theories of cognition have for philosophical issues.

In the first essay I explore the relationship between common sense belief-desire psychology and computation psychological theories, using belief as a model. I criticize the widely-held view that to have a belief is to explicitly store a representation and defend an alternative explanation which identifies having a belief with being disposed to use an appropriate representation in reasoning and decision processes. I argue that this dispositional account of belief suggests that common sense belief-desire (B-D) concepts and explanations may not figure prominently in scientific explanations of our cognitive make-up, although it is likely that scientific psychology will nonetheless recognize the legitimacy of B-D concepts and explanations. The result is a moderate realist view of common sense psychology, which does not commit us to anything as strong as Fodor's hypothesis of a language of thought. This dispositional account also allows us to understand how holistic aspects of belief-fixation can be realized in a modular cognitive architecture.

In the second essay I recommend that semantics be naturalized to empirical psychological inquiry. That is, I maintain that the descriptive project of specifying how the elements of our languages are related to the world should take into account facts about the psychological states that underlie our understanding and use of language. And the confirmational status of such accounts should be the same as for any other scientific hypothesis. I develop this proposal by sketching a *prima facie* conception of naturalized semantic methodology, which proceeds from a competence theory to postulations of mental states and processes. After defending this proposal against several objections, I turn to the question of whether there is any alternative, non-scientific methodology available to the semanticist. I argue that our pre-scientific knowledge of meaning fails to yield sufficient ingredients for a non-naturalistic semantics. Specifically, a substantial portion of

our knowledge of meaning appears to be non-explicit, and the explication of non-explicit knowledge is a task for empirical psychology.

In the third essay I defend a neo-kantian or anti-realist view of metaphysics, which maintains that the world we ordinarily perceive and theorize about is mind-dependent. This surprising result is based on the fact that our cognitive systems make a substantial innate conceptual contribution to our perceptions. As I show, both research on infants and methodological considerations from computational theories of perception strongly support the existence of a significant innate contribution. I then argue that the conjunction of perceptual nativism and any of the standard accounts of the justification of belief is inconsistent with a metaphysical realist view that claims that we can have knowledge of a mind-independent world. Specifically, accounts of justification all advocate some form of epistemic reliance on perceptions. But, since there is no guarantee that our innate perceptual concepts accurately correspond to a mind-independent world, there is no reason to think that increasing justification, as defined by any of the standard accounts, will bring us any nearer to realist truth--i.e. correspondence to a mind-independent world. Yet, it seems that justification, by its very nature, must leads us nearer to the truth, at least in the long run. The solution is to abandon a metaphysical realist view of truth in favor of a verificationist view, where truth is identified with ideal justification. This, together with the hypothesis of perceptual nativism, implies that the world that we know is in part constituted by the innate contributions of our perceptual systems. Such contributions are, in effect, synthetic *a priori*.

Thesis Supervisor: Dr. Robert Stalnaker

Title: Professor, Department of Linguistics and Philosophy

CONTENTS

PREFACE

I would like to thank Robert Stalnaker, who served as my advisor, for helping me to clarify my views. Thanks also to Ned Block, my other official reader, for providing very useful comments. I am also indebted to Noam Chomsky, who read several drafts of this dissertation, for insightful discussion.

Thanks are also due to Georges Rey and to Stephen White, each of whom served as substitute readers, for their valuable comments and criticism. And thanks to Jim Higginbotham for reading and commenting on the second essay. Finally, thanks to Eric Lormand and to Paul Pietroski who read and commented on the first essay.

A round of thanks is also owed to my fellow MIT philosophy graduate students who provided a positive, supportive atmosphere throughout my stay there. In particular, I would like to acknowledge the many useful discussions I've had with Michael Antony, Eric Lormand and Paul Pietroski.

Finally, a thousand thanks to my wife, Karen, for providing support, both financial and emotional, and for putting up with my nocturnal writing habits.

INTRODUCTION

Each of the following three essays is a self-contained inquiry into a distinct set of issues. There is, though, a common theme throughout--the view that traditionally philosophical questions about the nature of mind, meaning and knowledge will be answered, in part, through the emerging theories of cognitive psychology. Here I will comment briefly on philosophical naturalism and cognitive psychology.

In general, naturalism is the view (or better, the methodological stance) which maintains that there is no first philosophy, no knowledge that can be established in advance of empirical investigation (scientific or otherwise.)[1] Each of my lines of inquiry is consistent with this outlook, although specific naturalization may take a variety of forms. First, there is the issue of how traditional philosophical questions are to be dealt with. It might be maintained that traditional pursuits should be abandoned--that acknowledging that there is no first philosophy should lead us to give up or radically reformulate the target problems in a given area of inquiry, such as epistemology.[2] Alternatively, we might acknowledge the

1. I see no reason for the naturalist to also be scientistic--to reject any non-scientific inquiries. For instance, it seems reasonable to conceive of an empirical ethics, or perhaps a general normative empirical inquiry, that is not part of science *per se*.

2. This is apparently what Quine recommends for epistemology in "Epistemology Naturalized," in *Ontological Relativity and Other Essays*, New York: Columbia University

legitimacy of traditional philosophical questions, purged of any demand for certainty beyond the necessity that attaches to the laws of nature, and seek to answer them with the resources of empirical theory. I take this latter approach to the issues that I discuss--i.e. I think that questions about the nature of belief, meaning, justification, and even reality are perfectly reasonable. But they are not questions that are independent of or prior to scientific study.

Further, different forms of naturalism result from differing expectations about who will answer the relevant questions. On the one hand, we might suppose that the answers will simply fall out of scientific inquiry--that the only job left for the philosopher is that of cataloging the results that science has produced. This outlook suggests the withering away of philosophy as a distinct discipline. On the other hand, it might be that while scientific (or other empirical) theories provide the resources necessary for philosophical inquiries, scientists themselves will be largely unconcerned with the investigation of the issues traditionally addressed by philosophers. The latter possibility seems likely in many cases, for several reasons. First, philosophical questions are often much more abstract than the lines of inquiry that drive scientific research programs--so much more abstract that it

Press, (1969). Specifically, he seems to reject the pursuit of a normative account of justification in favor of a purely descriptive account.

seems reasonable to think that study of philosophical issues must remain somewhat separate from specific research programs. Second, naturalized philosophical issues may involve meta-questions, i.e. in the theory of theories, which suggests separate, albeit empirical, inquiry. Finally, it may simply be that people in philosophy departments investigate certain issues because those issues have been investigated by people in philosophy departments as a matter of historical accident, and the need for training in the relevant literature serves to insure that these issues will, for the most part, continue to be investigated by members of the same departments.[3]

My particular studies fall at various points on the continuum between naturalization to (non-philosophical) research programs and relatively autonomous philosophical inquiry. In the second essay, I recommend that semantics be naturalized to empirical psychological inquiry. That is, I suggest that questions about the meanings of our words will be largely answered by research in cognitive psychology and (cognitive) linguistics, with or without the assistance of philosophers, although current philosophical accounts of meaning may prove invaluable as starting points for the theoretical work. By contrast, I think that it is highly unlikely that the metaphysical and epistemological issues I

3. I assume there is no *a priori* division of subject matters to legislate what a given branch of academics should study.

pursue in the third essay will ever be directly dealt with by (non-philosophical) scientists, for all of the reasons mentioned above, but particularly because of the abstract nature of these questions. Finally, the question of the status of the common sense attitudes vis-a-vis computational, representational states should be answered by empirical psychology in the long-run. However, it seems that the issues are too broad and abstract to expect anything very specific from scientific research in the near future. Thus, the dispositional analysis of belief that I offer in the first essay should be viewed as a tentative theoretical account, which will stand or fall with future developments in psychological research.[4]

I should also point out that while I am not attempting any general defense of naturalism here, these studies do lend some support to the outlook. The naturalist holds that there is no *a priori* methodology that enables us to establish foundational truths in advance of empirical inquiry. Specifically, the naturalist must reject the view that we have knowledge of meaning that enables us to determine (non-trivial) analytic truths from the armchair. In the second essay I argue that our knowledge of meaning is largely non-

4. As will the alternative view that I criticize in that essay, i.e. Fodor's proposed identification of the attitudes with explicit computational states and his language of thought hypothesis.

explicit, which supports this rejection of a pre-scientific philosophical methodology.[5]

And, as I note in that essay, naturalism cannot be defended via *a priori* argument. Instead, all positive support must come from the successful naturalization of philosophical issues; the present studies contribute to this on-going project.

A few words about my conception of cognitive psychology may be helpful as well. Throughout I assume that psychology will, ultimately, produce explanations of our behavior and mental abilities in terms of (relatively) high-level representations and processes characterized as operations on those representations. My assumption rests largely on the present existence of modestly successful theories of this sort, but I also have several other reasons for favoring this conception. First, the research programs that have advocated methodological behaviorism--i.e. explanations in terms of functions from stimuli to behaviors--have ended in failure. The obvious alternative is to seek explanations that postulate inner states, thereby abstracting away from the bewildering mess of stimulus-behavior connections. Further, common sense psychology as well as pre-theoretical reflection suggests that many of the important aspects of our behaviors and abilities

5. This support for naturalism thus differs substantially from the views of Quine, whose naturalism is founded on the rejection of any substantial conception of meaning.

involve relations to states of the world. The obvious resulting supposition is that scientific explanations of these behaviors and abilities will postulate states that somehow correspond to the world. (Although we may discover that the world that we represent is not, in fact, mind-independent, as I argue in the third essay.) Finally, a review of the kinds of questions we want psychology to answer, i.e. questions about knowledge and abstract abilities such as the ability to speak a language, strongly suggests that empirical theories of representational states will postulate fairly abstract, high-level representations. This is to say that we should expect that psychology, to the extent that such a discipline is able to answer questions that we are currently asking, will produce theories that are distinct from (current) neurological theories in the abstractness of the level of explanation.

The approach to theoretical psychology that postulates abstract representational states and processes has been closely associated with the idea that the brain is a computer. While I think that the computer metaphor is generally a good thing--i.e. it provides us with a picture of how the brain can achieve various abilities--it is important to note that the general cognitivist approach need not imply rigid formalism. Specifically, while there is nothing wrong with the search for highly formal, computational explanations--they flourish throughout the natural sciences--it may turn out that we are rather "sloppy" computers. That is, our representational

processes may resist exact, mathematical specifications, and the content of many of our representations may be vague or fuzzy. Ultimately, such matters reflect how many and what sorts of generalizations we will be able to establish in theoretical psychology. I merely assume that we will be able to establish a substantial number--that we are not so "sloppy" as to resist all explanations--but this hardly implies that all or even most processes can be characterized in a mathematically precise manner.

Finally, it is worth noting that the conception of cognition that I rely on throughout these essays is officially neutral on one standard philosophical issue, the question of the ontology of mental states, and on one recent philosophical issue, the question of how to analyze the notion of representation or content which (apparently) underlies theories of cognition. Surely, some sort of materialism is ultimately appropriate to answer the general ontological issue--dualism seems explanatorily untenable--but this leaves open the question of whether mental states and attributes are type-reducible to biological or physical states and attributes, or merely token-reducible. As I have suggested above, we should expect to find relative autonomy for psychological research in the near future--as for the ultimate ontological status of mental states, I have no predictions. Nor do I have anything to offer concerning the issue of the notion of "representation" that underlies cognitive theories--

if there is, indeed, a single notion here. However, as I stress in the first essay, we should not automatically assume that our common sense means of determining the content of mental states must be preserved in the theories and methods of cognitive psychology--the latter will, no doubt, revise and improve upon the former. Most of all, the postulations and explanations of common sense psychology should not be mistaken for the postulations and explanations of a mature scientific psychology. This may mean that we will have to wait for cognitive theories to develop before we are able to answer ontological questions about mental states. But this is what we should expect once we abandon the presumption of an *a priori* philosophical methodology.

BELIEF, COMPUTATION AND COGNITIVE ARCHITECTURE

On the one hand, we have cognitive psychology, which seeks to explain behavior in terms of symbolic representations and causal, computational transformations of these representations. We are beginning to see the emergence of modestly successful theories within this information-processing paradigm. On the other hand, we have common sense belief-desire psychology--a set of concepts, ascription criteria, explicit and implicit generalizations and explanations that appear to characterize states with propositional content. These concepts, ascriptions, principles and explanations are very tightly woven into our daily lives and self-conceptions, and this appears to be a good thing, for belief-desire psychology embodies a highly successful set of (apparently) causal explanations of behavior.

What we want for the enterprise of cognitive psychology are successful theories, particularly theories that successful explain our higher cognitive abilities--those that are most abstract and furthest removed from sensory input. What we want for our understanding of common sense psychology is an explanation of the nature of belief-desire states, in particular, an account of how states with propositional content can be the causal determiners of behavior. Towards satisfying both these needs, Jerry Fodor has championed a

union of belief-desire and cognitive ontologies and explanations:

> The trick is to combine the postulation of mental representations with the 'computer metaphor.' Computers show us how to connect semantical with causal powers for *symbols*. So, if having a propositional attitude involves tokening a symbol, then we can get some leverage on connecting semantical properties with causal ones for *thoughts*.[1]

The idea is to identify states such as belief and desire with the explicit presence or activation (i.e. "tokening") of representations, and explicate belief-desire explanations in terms of computational transformations of these representations. This yields a scientific realism about belief-desire states, i.e. the expectation that many or most of the explanations of common sense psychology will be included in--and thus preserved by--a mature cognitive psychology. This alliance promises a set of successful explanations for cognitive psychology in an area (higher cognitive processes) where we are currently lacking them. And the proposed identification of belief-desire states with symbolic states promises to show us how states with propositional content could be efficacious--i.e. if semantics

1. Fodor (1987), p. 18. It may be tempting to read Fodor as saying that it is only common sense belief-desire states that have representational content. But that is evidently false, for many existing cognitive theories postulate representational states unknown to common sense. And Fodor would apparently agree that the set of representational states is larger than the set of (common sense) propositional attitude states, since he champions theories of the perceptual systems that postulate representational states unknown to common sense (or introspection)--see Fodor (1983).

mirrors syntax, then since we can see how syntax can be causally efficacious, we can also see how states with content cause behavior, even if content itself isn't the cause of behavior.

These are certainly welcome consequences. However, this viewpoint has some unpalatable implications too. First, it appears that if we think of our central cognitive systems as collections of explicit belief-desire states, then, owing to the holistic properties of common sense psychology, we must admit that the prospects of finding strict, computational accounts of these systems, in the information-processing paradigm, are rather dim.[2] Second, ascriptions of attitudes to pre-verbal children and animals implies, given this outlook, that they have an inner high-level language similar to our natural languages. But their apparent lack of any external language or language skills suggests that this is implausible. Third, if it is these explicit belief-desire states which underlie concept acquisition, then it seems that we can only acquire concepts that we are already able to extensionally formulate in our inner language. But, since few concepts reduce to simpler ones, it appears that most of our concepts must be innate, a consequence that is generally regarded as highly counter-intuitive, and most likely false.

2. See Fodor (1983), 101 ff.

Thus, there is substantial motivation for questioning Fodor's proposed form of scientific realism for belief-desire states. The crucial point of Fodor's position is the proposed identity of common sense states with explicit informational states. If the position is sound, it must turn out that we can isolate classes of explicit computational states that can plausibly be identified as common sense state types, e.g. belief, desire, etc. In the first section, I examine the computational status of (the propositional attitude state type) belief. As I will show, belief is better analyzed as a dispositional computational state rather than as an explicit computational state. This account can be generalized to the view that common sense belief-desire states are typically dispositional rather than explicit or activated cognitive states, which is to say that Fodor's envisioned union of cognitivism and common sense psychology is mistaken.

As I then proceed to show, the alternative, dispositional view of common sense belief-desire states embodies a plausible form of moderate realism for these states, i.e. a viewpoint which avoids the undesirable implications of Fodor's account. First, I will argue that the dispositional view allows us to acknowledge certain worries that have lead to eliminativist and instrumentalist views of belief-desire psychology, but without adopting these implausible alternatives. I will then show that the dispositional view of belief-desire states allows us to understand how holistic central system functions

can co-exist with a modular or otherwise non-holistic computational architecture. Finally, I will examine the commitments of the dispositional view of belief-desire states. I will argue that we need not, as Fodor maintains, be lead to the postulation of a language(s) of thought for all believers, or to the claim that most concepts are innate. In sum, I shall be presenting and defending what I believe to be a more plausible alternative to Fodor's, and any other, account of the cognitive make-up of belief-desire states.

I. Specifying Belief

What is it to have a belief? I.e., what sort of a state is a belief state? The approach to this question that I will pursue is that since belief is a psychological state, we should expect scientific psychology to eventually tell us what beliefs are, just as, e.g., we might expect that questions of the form "what is it for something to be X?", where X appears to be a specification of some category of chemical substance, say water, will be answered by chemistry. What I will be seeking is, in Cummins' terms, a **property instantiation** explanation--an analysis the property of belief in terms of concepts from (cognitive) psychology.[3]

Current wisdom has it that cognitive psychology is both computational and representational, and so we should expect to

3. See Cummins (1983), Chapter 1, especially pp. 14ff.

find a computational, representational explanation of (the property of) belief. Specifically, Fodor has suggested that having the belief that *p* is a matter of being in a computational relation to a mental representation that means *p*.[4] This divides the property specification project into two parts, namely (1) specify the appropriate computational relation for belief and specify the representations that are so related and (2) explain what it is for mental representations to mean *p* (in general.)[5] While much of the discussion of the relationship of belief-desire psychology to scientific psychology has focused on the content of that-clauses or content of the representations underlying the attitudes (2), I will instead examine the computational relation which we might expect to find for belief (1).

4. See the introduction to Fodor (1981) as well as "Propositional Attitudes," in that volume.

5. A standard assumption of this view about the representations which are hypothesized to underlie attitudes, seemingly too obvious to state, is that it appears that different types of attitudes, e.g. belief, desire, fear, hope, etc. are formed via combinations of the same types of tokens in different computational relations to cognition. E.g. it appears that the difference between hoping that abortion will be outlawed and fearing that abortion will be outlawed is entirely one of computational role--the that-clauses appear to have identical meanings in such cases. So, the obvious hypothesis is that a given representation (typed by content) can stand in different computational roles to achieve the various attitude types. I will make this assumption throughout.

The more or less standard computational explanation of belief begins with the idea of **storage**.[6] More formally, the hypothesis is that having the belief that *p* is a matter of storing a representation that means *p*. This cannot be the full account because it is likely that storage will also be involved in computational explications of other attitudes, notable memory. Thus, belief must be a matter of storing appropriate representations "in the right way," where belief storage is distinguished from other sorts of storage via functional role.[7]

In the rest of this section I will criticize the storage model of belief and introduce and defend the alternative view that the property of belief should be identified with dispositions to use a representation in explicit processes--a

6. Adherents of this view include Lycan (1988) chapters 1 and 3, Block (1990) pp. 271-4 and Field (1978) pp. 80-84. Field suggests a general dispositional account of the sort I will defend below, but then opts for the storage and disposition-to-infer view as an elaboration of the former view.

It is also worth noting that the storage model is perpetuated to some extent by talk of "belief-boxes." This term, as used by Schiffer and by Fodor is intended to stand for whatever computational relation belief turns out to be. However, talk of "boxes" invokes the idea of explicit storage.

Finally, as noted at the outset, Fodor is generally committed to the identification of common sense states with explicit cognitive states, and although he has not, to my knowledge, ever explicitly advocated the doctrine, some version of the storage view would seem to be the likely candidate for belief.

7. See Lycan (1988), pp. 6-7.

view which makes no commitment at all to explicit storage for any beliefs.

An obvious and difficult problem for the storage view is that it appears that we can have infinitely many beliefs, but only store finitely many representations (assuming that the storage view is otherwise correct.) Or at least, it seems that the number of beliefs a person has can easily exceed the amount of storage space that it is reasonable to postulate for explicitly stored (high-level) mental sentences.[8] Thus, the storage view is typically supplemented with the further hypothesis that there are certain "explicit" beliefs that are representations that are explicitly stored, with the remainder of our beliefs being "implicit", i.e. not explicitly stored but related appropriately to those beliefs that are explicitly stored, where having the implicit belief that *p* is a matter of being disposed to produce (e.g. infer) a token that means *p* from stored tokens, via an "extrapolator-deducer" as Dennett calls it.[9]

However, this account does not stand up under scrutiny, as Lycan and others have pointed out.[10] The difficulty is that

8. Intuitions differ widely on this matter. Also, as I shall discuss below, a possible move here is to opt for realism about only a select core of beliefs.

9. See "Brain Writing and Mind Reading", in Dennett (1978), where he explicitly proposes an "extrapolator-deducer" but ultimately rejects a realist account of the attitudes.

10. Lycan (1988), Chapter 3, which includes mention of others who have noted these problems, including Dennett, *op. cit.* Lycan sometimes refers to these as "tacit" beliefs, but I

we also want to allow that people sometimes produce new beliefs from old ones, and it is not apparent how we are to distinguish these acquisitions from implicit beliefs. For instance, we frequently infer things from what we already believe. And we think that such inferences produce new beliefs. However, it seems that the hypothesized extrapolator-deducer might follow similar inference patterns. What, then, is to distinguish implicit beliefs from those we are disposed to acquired? There are several possible answers, though none of them succeeds.

It might be claimed that conscious processing separates explicitly acquired from implicit belief; i.e. it might be claimed that the extrapolation which results in implicit belief being made explicit is all unconscious, whereas all belief change is the result of conscious thought. But this will plainly not do. We allow for unconscious belief change--especially in cases of vast, gradual revision. For instance, many people are trained as children to believe one or another religious ideology, and many of them abandon these beliefs later in life, but it is implausible to claim that in all such cases, all of these belief changes were carried out in a conscious manner. E.g., such people haven't consciously

prefer the term 'implicit' since others, notably Chomsky, use 'tacit' in a way that suggests explicit, non-conscious representation.

reconsidered **everything** that they were explicitly taught as children.

Also, even when belief change is partially conscious, much of the inference process itself is not conscious. Our conscious thoughts typically express only a few crucial premises in what is often a fairly complex series of inferences. Yet, it is often plausible to attribute belief to the suppressed premises and sub-inferences used in reasoning. For instance, I may arrive at home, see my wife's coat and unconsciously infer to the conscious thought that she is home. Presumably, I believe that if her coat is here then she's home, and have used this belief in my unconscious inference. But my belief in the conclusion is surely belief acquisition rather than extrapolation.

Perhaps it might be suggested that when reasoning from conscious premises is occurring, then belief change is occurring. But this overlooks the fact that extrapolation might occur in the midst of reasoning. For instance if you assert "if I'm a conservative then pigs can fly" I might extrapolate the (implicit) belief that pigs cannot fly, and then infer that you want me to believe that you're not a conservative. Clearly I haven't acquired the belief that pigs can't fly in such a case. So in general, admitting the existence of unconscious or partially unconscious reasoning--as we surely must--undermines the claim that being conscious

is what distinguishes belief change from the accessing of implicit beliefs.

Another *prima facie* candidate is a temporal criterion. Much of our belief change results from long chains of reasoning, whereas candidate implicit beliefs, such as "elephants don't wear pajamas", bring immediate assent. However, a bit of reflection shows that this will not do. For we also frequently acquire new beliefs through immediate assent. Sometimes, a suggestion that has never occurred to us before can seem immediately plausible. And sometimes we draw inferences quite rapidly as well. For instance, a violation of Grice's conversational maxim of relevance can lead us to immediately infer that the speaker did not approve of the previous remark. And there is no basis for thinking that such cases of rapid assent or inference are any slower than the (apparent) inference involved in (apparent) implicit beliefs. Thus, any choice of a temporal criterion for distinguishing implicit beliefs from beliefs we are disposed to acquire would appear to inappropriately classify many cases that are very obviously newly acquired beliefs as implicit beliefs and vice versa, so no such criterion is acceptable.

One final suggestion might be that there are simply two different processes in cognition, one which extrapolates implicit beliefs and one which produces new beliefs from stored beliefs. The problem, though, is that it is not apparent how we could identify such processes as being either

implicit belief "actualizers" or new belief producers. E.g. suppose that we discover that there are 23 distinct inference-extrapolation process which operate on a stored set of beliefs. Which of these would be the implicit belief processes and which the new belief processes? It seems that we need some criterion, i.e. some conceptual distinction between implicit belief and dispositions to acquire belief, which is more or less prior to empirical theories of inference and extrapolation. But this is what we are lacking.

Thus, the extrapolator-deducer view fails to yield an adequate distinction between implicit belief and dispositions to acquire beliefs. This leaves us with the other more or less standard means of attempting to characterize implicit belief, i.e. in terms of disposition to judge. Thus, it might be hypothesized that having the implicit belief that *p* is being disposed to (inwardly) judge that *p* upon entertaining the occurrent thought that *p*. However, this faces the same sorts of problems as the extrapolator/deducer view, since some inner judgments signal the acquisition of new beliefs while the rest affirm old ones. Consider the following counterexamples that Lycan has presented against the dispositional view:

> 1. *The opinionated people.* They are Peircians, in that they abhor being agnostic on any subject, but not Peircian enough, in that in them the "irritation of doubt" triggers not inquiry but snap judgments. On many occasions, at least, when they entertain a proposition for the first time, they immediately affirm the proposition or deny it, depending on what else is going on in their global psychology at the time. Thus, at a time *t* our subjects have countless dispositions to judge--determined by their global

> psychology--but we would not count these as antecedently existing beliefs, however [implicit].[11]

Of course, this is mere fiction. However, it is easy enough to find cases from everyday life. Thus, consider a second counterexample that Lycan borrows from Audi:

> 2. *The excited raconteur.* He is regaling his dinner companions with a voluble account of some startling incident, waving has arms and talking too loudly. If he were simply to entertain the proposition that he was talking too loudly, he would instantly realize that it is true. But not having entertained the proposition, he does not already know or believe it in any sense.[12]

Perhaps, in an attempt to save the disposition to judgement view, it might be suggested that the dispositions must be of the right sort, i.e. they must result from the more or less evidential function of the explicit beliefs. However, this will not do either. For the following possibility seems plausible enough. Consider a scientist who has never thought of some theory *T*, even though this theory would explain a lot of relevant data and coheres wonderfully with the other theories she believes. One day someone else suggests *T* to her and she immediately sees many of its appealing features, relative to her other beliefs, and so her immediate reaction is to say "*T* must be true, I wonder why I never thought of that..." Thus, it would seem that the disposition to affirm *p*

11. Lycan (1988), p. 58.

12. Lycan (1988), p. 67.

is not the way to go in distinguishing implicit belief and dispositions to acquire belief.

A line of response to the problem of accounting for implicit belief is to simply give up on implicit belief and opt for realism for only an explicit "core" of beliefs--i.e. those that involve explicit storage.[13] The proposal assumes that science will, in general, show us what the real extensions of our common sense natural kind-concepts are. And since we apparently cannot computationally account for implicit belief, the next best option would seem to be to adopt a computational account of belief which turns out to only include explicit beliefs. I will now examine this line.

An initial problem that arises for the core view is how to account for apparent implicit beliefs. Thus, it seems rather counter-intuitive to flatly deny that, e.g. I believe that there are no anteaters in the room or that 10,013 is the successor of 10,012 (assuming these candidate beliefs would not make the explicit core--I shall take up the issue of what's in the core next.) A slightly more reasonable move, suggested by Audi[14], is to claim that in such cases, while we

13. See Block (1990), p. 271. Lycan (1988) too would apparently hold this view if, as he suggests, there is no plausible account of tacit beliefs. And Audi (1982) also proposes this sort of view, although he does not seem to view his proposal as scientific revisionism. Also see Fodor (1987), pp. 20-21 for a general advocation of the core view for all attitudes.

14. Audi (1988).

do not have these (apparently implicit) beliefs, we are disposed to acquire them upon forming the appropriate propositions in thoughts. While this is better, it is still not completely palatable. Consider related cases where I genuinely did not know. E.g. until I enter and look around a room for the first time, I have no beliefs about what's in it (expectations, maybe, but no beliefs, at least in some cases.) And until I calculate, I do not believe that 78*67=5226. The former cases appear substantially different from this. Thus, in the cases of the propositions that there are no anteaters in the room or that 10,013 is the successor of 10,012, I would like to say I **know** that these things are true even before I formulate the thoughts. It was not as though I had no attitude, no opinion at all on such matters until I contemplated them. This is not, I think a fatal problem for the core beliefs proposal because one can, with revisionism, always bite the bullet and opt to throw out certain intuitions. But a view that advocates too much revisionism itself becomes implausible, and this is a lot of revision--so this problem is a strike against the view all the same.

A further issue concerns which (common sense) beliefs are supposed to be in the core. One possible view is that it is roughly those propositions that have been affirmed in consciousness and are still stored. While this is not a position that advocates of the core hypothesis typically advance, this does seem to be the basis on which the

distinction is actually drawn. Thus, why should it be that the belief that dogs are animals is part of the core, whereas the beliefs that 10,013 is the successor of 10,012 and that there are no anteaters in the room are not part of the core? Presumably, the decisions about how to classify such candidate beliefs are based on our knowledge of what we have and haven't ever consciously thought about. It is likely that we all have at some point consciously affirmed that dogs are animals and highly unlikely that we have ever consciously entertained, let alone affirmed the other two propositions. Or consider Lycan's discussion of why his wife's beliefs that she is less than 18 feet tall and that 10,329>10,328 are implicit:

> There is no plausible sense in which these things are represented explicitly within her at this very moment, much less hooked up with the other relevant concepts in even a quiescent way. In particular, she never episodically *judges* that she is less that 18 feet tall, or the like.[15]

I take it that an "episodic judgement" is a conscious judgement. Or consider Audi's rejection of the claim that he believes before entertaining a certain thought:

> But is it at all likely that my belief that the sun is more than 100.542 miles away, was formed before I entertained the proposition, given that (for instance) I never perceived, inferred, or introspected it, nor experienced anything in which it figured in any special way?[16]

15. Lycan (1988), pp. 55-56.

16. Lycan (1988), p. 117.

Again, it would seem that what is doing the work here is knowledge of what has and hasn't been consciously entertained. So it is worth examining the proposal that one's core beliefs are roughly those affirmative conscious judgments that are explicitly stored.

This view faces two severe difficulties. One is that there are lots of successful belief-desire explanations whose attributed beliefs apparently do not correspond to consciously affirmed thoughts, and it seems arbitrary to maintain that such cases are not genuine beliefs. For example, consider perceptual beliefs. In our daily trafficking with the world we perceive the location and attributes of countless objects, yet we explicitly pass judgement on only a few of these cases. Thus, we do not typically walk around thinking "there's a blue book about 8" by 5" by 2" lying towards the corner nearest me of a wooden brown-lacquered table..." All the same, we can usually volunteer such information, if need be. But on the occurrent core belief proposal, most such states would not count as perceptual belief. However, we typically explain a person's actions with some of the information in perceptual states, e.g. "he saw the table and stepped around it," and it seems unacceptable to deem such explanations no good. Certainly, they appear to be as successful as any other belief-desire explanations.

Perhaps this type of case can be remedied by broadening the class of conscious states beyond occurrent judgments to

all conscious states that get explicit stored. Thus, it may be that much of our perceiving is carried out in terms of iconic states, and it is these which we frequently consult in finding our way about. While judgement or occurrent thought does not occur in such states, we nevertheless (apparently) consciously entertain them, (i.e. we "receive" perceptions) and apparently also sometimes store them, so perhaps occurrent core belief could be expanded to include such (apparently) non-sentential conscious states.

However, it does not appear that a similar strategy is available for other cases. For example, we typically do not consciously entertain Gricean attitudes when we communicate. E.g. we do not think "I want her to come to believe that I am uttering S to get her to believe that I uttered S in order to get her to believe that I believe that *p*..." Yet, again, such explanations offer a fairly successful explanation of how we appear to understand one another's speech acts when we communicate. And the occurrent core view apparently must reject the postulation of such attitudes and the related account of communication. Or consider that many of our ordinary actions apparently rely on beliefs about the world. We turn the faucet because we believe that this will produce water, and we turn doorknobs because we believe that this will open doors. But surely, few of us have consciously affirmed such mundane beliefs at any recent point in our lives--e.g. I suspect that we generally can't remember when, if ever, we

explicitly affirmed that turning the doorknob will open the door. But on the occurrent core view, this is to say that we have little basis for attributing these beliefs to ourselves. And this is quite implausible. For, again, such explanations are the bread and butter of belief-desire explanations. I.e. it would seem that if we are going to allow that there are beliefs at all, we should allow that there are such basic action-guiding beliefs.

A second major problem for the conscious core proposal is that it simply seems unlikely that many of our consciously affirmed thoughts are actually explicitly stored. If a representation is explicitly stored, we can typically reproduce it, more or less at will, or with appropriate cues. This is true of "memorized" sentences, including quotes, sayings, slogans and poetry. Thus, it is reasonable to hypothesize that explicit storage underlies such memorizations. However, reproduction is not at all characteristic of occurrent thoughts. Rather, it seems that our thoughts are rarely exactly the same as previous thoughts, and if they are closely related to previous thoughts at all, they are variations or modifications of what came before. Thus, we are typically very bad at recalling exactly what we were thinking after thoughts have ceased to be occurrent. E.g., I can tell you the **general content** of what I was thinking a few minutes ago and perhaps reconstruct a few significant points, but I have no idea exactly what my inner

utterings were--we are rarely able to quote ourselves a few minutes after a thought is gone. In fact, an important function of physical transcriptions is to preserve a concrete formulation of thoughts, something we are usually unable to do for large numbers of thoughts. For instance, you probably would want to allow that I believe virtually all of these sentences, as I write and re-read this paper, yet at any moment I am lucky if I can recall even a single sentence of this essay word for word. All this would suggest that we typically do not explicitly store our occurrent affirmations, which is to say that the occurrent core view is mistaken.[17]

There is an alternative means of formulating the core belief hypothesis, namely in terms of representations that are explicitly stored and potentially active in cognition. As Lycan puts it:

> A paradigm case of this would be one in which a previously tokened representation is now stored quiescently in long-term memory. The stored formula is accessible to various executive agencies and can be hauled out on cue, resulting in a new judgement or tokening bearing the same computational shape.[18]

What is important about the stored representations is that they can produce appropriate effects in cognition and

17. This is not to say that no explicit storage lies behind our thoughts and judgments, just that it probably isn't storage of conscious thoughts themselves.

18. Lycan (1988), p. 56. This is also presumably the view Fodor would want to maintain, as the quote on p. 1 above suggests.

ultimately behavior, i.e. that they have the causal role of (the common sense notion of) beliefs. As Block suggests:

> we home in on cases in which our beliefs **cause** us to do something (say, throw a ball or change our mind) and cases in which beliefs are caused by something (as when perception of a rhinoceros causes us to believe that there is a rhinoceros in the vicinity). So the protoscientific concept of [core] belief is the concept of a **causally active** belief.[19]

However, this explicit storage/causal role version of the core hypothesis avoids the implausibility of the conscious core hypothesis only if the two are relatively non-co-extensive. Thus, if is also hypothesized that most of the explicitly stored representations in cognition correspond to conscious or potentially conscious states, then the active representation hypothesis faces the same problems we have just noted for the occurrent hypothesis.

On the other hand, at present we have very little idea what explicit stored representations are behind the operations of cognition. So the explicit representation hypothesis gives us very little to go on. E.g. it does not tell us which apparent beliefs will turn out to be actual--"in the core." Nor do we have much evidence, beyond consciousness, as to whether or not the explicitly stored representations in cognition will correspond in any way to the beliefs that

19. Block (1990). I take it that "causally active" means potentially causally efficacious rather than causally activated--for otherwise, under the proposal, most of us would probably have just a handful of beliefs at any given moment, and would sometimes have no beliefs at all, but this absurdly revisionary.

common sense attributes to individuals. Thus, while the hypothesis is not implausible, it is not particularly plausible either, since we really have no relevant evidence with which to evaluate it.

And further, there is a lingering implausibility. The explicit core hypothesis has got to be somewhat revisionary, if not radically revisionary in regard to what beliefs we have, and, as I noted above, such revision does not come easy--all things being equal, a non-revisionary account of some concept is preferable over a deeply revisionary account.

As far as I can see, this exhausts the available options for an account of belief as explicit storage, and thus exhausts the options for the identification of belief with an explicit cognitive state.[20] So the explicit core/causal role view appears to be the only reasonable form of the storage hypothesis. However, as I will now argue, there is a more plausible alternative to this view, an alternative which abandons an identification of belief with explicit states in favor of an identification with implicit or dispositional states. As I will show, it is the causal role which does all the work for the hypothesis we are considering--the idea of

20. Could belief be something other than storage? The problem is that storage seems like the only informational state that has any real over-lap with the apparent extension of belief. I.e. all you really get in computers are storage and computations, and since belief appears to be an enduring rather than a momentary state, only the former seems like a reasonable candidate.

storage is in fact superfluous. But, since it is the appeal to storage that motivates the core hypothesis, this means that we should abandon the core storage view in favor of a "pure" causal role account of belief.

The storage/causal role specification proposes that having the belief that *p* is a matter of storing a representation that means *p* where this representation has an appropriate causal role in terms of its relations to other representations, states and behavior. The latter clause is required since there will presumably be other attitudes whose computational explication involves storage, notably memory. For instance, it is possible to remember something but not believe it. E.g. I remember the first verse of the Bible, but I do not believe it.[21] So belief must at least be a matter of storage with appropriate additional relations to other elements of cognition, ones that merely remembered representations do not share.

21. Perhaps it might be argued that there is belief involved in such cases after all, e.g. that I believe that 'In the beginning...' is the first verse of the Bible. However, it seems possible to remember a quote without remembering the source or context. In fact, this seems to be a common occurrence. In any case, we can see that this reply will not do on other grounds. It is an intelligible empirical hypothesis that we remember quotes via explicitly storing them. Thus, suppose that I remember the first verse of the Bible by explicitly storing an English token which expresses that verse. But, if belief is nothing but explicit storage, then I must believe this as well since I have, by hypothesis, explicitly stored a representation with the appropriate meaning. But, (by hypothesis and in fact) I don't believe it, so we need to specify something further, in addition to storage, which enables us to distinguish belief from memory.

Here it is worth noting that although this hypothesis is presented as the claim that there is a "belief box" in cognition, where this is understood as the claim that all beliefs are stored in a unit which bears an appropriate functional role to the rest of cognition, it is unlikely that there will actually be a storage unit dedicated solely to belief. This is because we apparently remember all, or most of what we believe. But it is unlikely that there is such massive duplication in cognition, i.e. where for most beliefs that *p*, there is one representation that means *p* that is literally stored in a belief box and another that means *p* that is stored in a distinct memory box. A much more plausible version of the storage account postulates one memory/belief container, where individual representations are classified (as either remembered, believed or both) by their individual causal roles.

It will useful for my purposes to make the causal role a bit more vivid. Two features suggest themselves, more or less from traditional functionalism. First, beliefs are the basis for our reasoning and theorizing. If we believe something, then we are willing to draw conclusions from this belief and to test other potential beliefs against this belief. Second, beliefs often lead to actions, when combined in an appropriate way with action-driving states. Specifically, we use our beliefs to reason out courses of action which will lead us toward our goals, and sometimes act on them. Thus, the

relevant functional role for the belief that *p* is roughly to be disposed to use *p* as a basis for theoretical and practical reasoning, where the latter sometimes leads to action. Or, to put it more traditionally, beliefs are those states which combine with other beliefs to yield new beliefs, and combine with desires to produces intentions, actions, or further desires.[22]

I will now argue that, in fact, it is this causal, functional[23] role alone, and not storage, which is doing all the work in the storage/causal role model. Consider first a case where storage is present but the causal role is removed, i.e. a case of failure to immediately update. Suppose we indeed have a storehouse of representations that have the causal roles appropriate to belief. It is unlikely that all these explicitly stored representations are updated every time changes in belief occur. The most obvious reason is simply

22. This type of specification differs from traditional functionalism in that no attempt is being made to give a functional specification of the representation or its content. That is, only the state type belief is being identified with a functional role (assuming that storage is also ultimately functionally specified.) By contrast, the traditional functionalist identifies each belief that *p* with a distinct functional role. Also, as I have discussed at the outset, and will consider again below, the cognitivist assumes that these common sense characterizations can ultimately be made more computational.

23. I use "causal role" and "functional role" as synonyms here, although there may be a difference. I.e., some traditional functionalists may wish to deny the causal efficacy of belief, but I am not concerned with such views here.

that the "core" of explicitly stored representations would have to be quite large and it is much too computationally expensive to be feasible to update all of them after every few cognitive operations. It follows, then, that there will sometimes, perhaps often, be explicitly stored representations that are in need in of up-dating. But in such a case the explicit presence of the representation will apparently be irrelevant to belief attribution. Let us assume, for the sake of example, that explicitly stored sentences underlie some beliefs. Suppose that I once believed the Christian ideology, but now have rejected it. Yet, at a point in time not too long after these major belief revisions have occurred, I might still have some explicitly stored tokens corresponding to my former beliefs. Suppose, for instance, that I have explicit stored a token of 'Jesus changed water into wine.' Owing to my belief changes, I am now not disposed to act on this representation--upon (eventually) accessing it, I will, let us say, erase it. But I have not had occasion to access it for years. During this dormant period, do I still believe that Jesus changed water into wine? Someone transfixed with the explicit storage view of belief might want to claim that I do, but consider that, in such a case, I would have no dispositions to act on this belief or express this proposition--my behavioral dispositions would be indistinguishable from someone who had never explicitly stored this proposition. Nor would any conscious judgement or thought

reflect this belief. Thus, it is plausible to claim that in such a case, I would no longer have this belief even though I have explicit stored a token which has the appropriate meaning. In other words, the continuing presence of the causal role is required for the presence of the belief.

The previous case suggests that the causal role is a necessary and important part of the storage/causal role model of (core) belief. Now, however, I will argue that causal role alone is sufficient for belief. To see this, consider two people, Romulus and Remus, who share a set of causal roles that are characteristic of belief. Romulus' causal roles are associated with explicitly stored tokens in just the way that the core storage plus causal role model predicts. So, e.g. he has the causal role characteristic of belief associated with some stored representation *p*, and all this leads to the attribution, on the part of himself and others, of the belief that *p*. Remus, on the other hand has, for every causal role and stored representation of Romulus, the same causal role associated with a different set of stored representations and the disposition to produce the (semantic) type of representation that Romulus has stored. So, for the causal role associated with *p* in Romulus, Remus might have stored a representation that means "if *q* then *p*" and another that means *q*, and have these associated with the disposition to infer a representation that means *p* from these other two representations on just the occasions when a representation

that means *p* is active in Romulus. (We needn't suppose that all this production from stored representations on Remus part is deductive, but that's an easy way to simplify the example.) So, in effect, for all of Romulus' "explicit" beliefs, Remus has causally equivalent implicit beliefs.

Now, it would seem that the storage plus causal role model of core belief must make the dubious claim that while Romulus has a belief that *p* for every stored representation and associated appropriate causal role, Remus in fact has no beliefs at all.[24] But this is wildly implausible. Consider that all their behaviors and dispositions to behavior are identical, as well as all of their cognitive states beyond those involving the storage aspect of the belief causal roles. Exactly the same propositional attitude explanations will seem true of them, and the same belief ascriptions will seem to apply equally well. Moreover, they themselves will (independently of a commitment to the storage view and knowledge of what they store) attribute exactly the same beliefs to themselves.

24. Note that the view would not attribute beliefs corresponding to Remus' explicitly stored representations either, since they lack the appropriate causal role. E.g. when representation meaning "if *p* then *q*" becomes active in Romulus, one also becomes active in Remus, though not as a result of the stored token that means "if *p* then *q*," but rather as a result of extrapolation from some other stored representations, e.g. one meaning "*r* or if *p* then *q*" and another meaning "not *r*." And, as we have just seen in the preceding case, and in considerations of memory, storage alone is not sufficient for the attribution of belief.

Does the storage theorist have any basis for arguing that only Romulus has beliefs, other than a brute appeal to the storage model? It would seem not, and thus it would seem that a more plausible view than the core storage plus causal role account is the hypothesis that it is the (appropriate) causal role alone that should be identified with belief, particularly given the problems we have noted with the core view--i.e. its failure to explain apparent implicit beliefs in a satisfying way, and the lack of any evidence for supporting the hypothesis that there are in fact stored representations in us corresponding to many of our ordinarily attributed beliefs. As we have seen, the storage view requires that we are able to identify a belief's causal role anyway, and since causal role by itself appears to come closer to being co-extensive with our common sense criteria for belief attribution and explanation, it would seem that the latter is a more appropriate scientific explication of belief. It is to this approach that I will now turn.

The immediate problem is to provide a specification of the causal role that does not rely on the notion of storage. Recall that our specification was that to have belief that *p* is to store a representation that means *p* and to be disposed to use this representation as a basis for theoretical and practical reasoning. As our previous example has shown us, what we want is for such dispositions to be realized when it matters, i.e. in situations when a token that means *p* is

activated in the system. This suggests a conditional, dispositional formulation of the causal role:

> the belief that p =$_{df}$ the disposition to use a representation that means p in theoretical reasoning and as a basis for action and practical reasoning when such a representation is explicitly formulated.

If this account is to succeed in place of the core belief view, it must deal with the issue with which we began, namely how to distinguish the disposition to acquire belief from dispositional belief. I will first present some clarifications of this "pure" causal role view, and then proceed to address the problem concerning dispositions to acquire beliefs.

It should be noted that, while I have been using common sense terminology, e.g. "theoretical reasoning," the definition, in keeping with the computational approach, is intended to define belief dispositionally in terms of non-dispositional cognitive processes. As such, the definition is best understood as a sketch of a more exact and formal definition which will be possible when we have a fuller, well-confirmed theory of cognitive processes. Perhaps a better immediate formulation would be:

> the belief that p =$_{df}$ the disposition to use a representation that means p as input to processes which use it as a basis for inference and explanation and as input to processes which use is as a basis for planning,

deciding and acting, when such a representation is explicitly formulated.

I take it that it is a fairly safe bet that we do have explicit processes of explanation, inference, decision, and planning, or at least that we have explicit processes underlying the states we ordinarily identify with these terms.[25] The significant feature of the definition, however, which will loom important throughout the rest of the essay, is that formulation doesn't identify belief with explicit states, but with dispositions for the use of representations in explicit processes.[26]

Also note that the "explicitly formulated" clause must imply some sort of causal role itself in order to distinguish this view from one on which "explicit storage" replaces explicit formulation. I take it that the representation in question must be "on-line", that is, available to most or all of the processes that could potentially use it as input. Perhaps there is one central "buffer" which allows for such

25. If not, then my specification will be falsified. It is, after all, offered as an empirical property specification as part of very abstract cognitive psychology.

26. Note that from here on in, when I speak of this as a "functionalist" or "dispositional" account, I mean this special sense of functions or dispositions to use in explicit cognitive processes, not the traditional sense of functions or dispositions to (observable) behaviors.

Also note that the account says nothing about the output of reasoning or decision processes. So, e.g., I am not claiming that we always do what we believe to be the best action.

availability.[27] Or perhaps the necessary availability is a more complicated matter, e.g. of a controlling unit feeding the representation to a number of appropriate locations. In any case, I shall rely on the intuitive notion of being "on-line," where required, while assuming that this part of the definition will also be subject to suitable modifications as cognitive theory develops.

It is also worth noting that this definition does not imply anything about conscious or self-access to the explicit formulations of representations and processes. It appears that many of the explicit representations and computational processes characterized by cognitive theories are sub-conscious and this means that many of the processes and representations which are relevant to this definition will be outside the access of consciousness. The proposed definition allows for sub-conscious causal roles and therefore allows for sub-conscious belief.[28]

One final point of clarification concerning the causal role specification concerns the need for both a clause citing theoretical reasoning processes and a clause citing practical

27. See Baars (1988) for the suggestion that such a buffer is what we ordinarily call conscious thought.

28. I take it that in cases of sub-conscious belief attribution the postulated beliefs do play a role in inference, planning and action. For instance, in Freudian theory, the primary roles of repressed beliefs is to interact with repressed desires (or repressing desires) to produce other, conscious beliefs and desires and to determine the content of dreams.

reasoning and action guiding processes. Clearly, the latter clause alone is not sufficient, for this would not distinguish belief from desires, intentions and the like which also enter into practical reasoning and guide action. However, it might seem that the first clause alone could suffice for our definition, since we need not always act on our beliefs. But this would not allow us to distinguish hypothetical reasoning from belief. Thus, we are sometimes disposed to use hypotheses in reasoning, and to deduce further consequences from these consequences and so on. What seems to ultimately distinguish belief from a supposition is that we are prepared to act on the former, to bet on it--as they emphasize in decision theory--but not on the latter. So it would seem that both a theoretical reasoning and a practical reasoning and action guiding clause are required for our functional specification of belief.[29]

Consideration of apparent cases where the two criteria diverge further supports this specification. Suppose we have someone who accepts a certain proposition as a basis for various inferences, but who, in situations where the belief appears relevant, completely fails to act on it. Thus, suppose an individual is willing to assert with conviction that racial

29. There are cases in which we are prepared to act on what we don't believe--e.g. when we want to act as though something is the case. But in such cases we are be prepared to act on *p* only in a limited way--we wouldn't want to make all decisions as though it were the case.

differences should not affect an individual's opportunities in society, and who, in fact, draws inferences based on this claim, but who consistently fails to hire job candidates of certain races even when their qualifications are obviously superior to any of their competitors. Were the disposition to use an explicit representation in theoretical reasoning alone a sufficient condition for belief, then we should have no trouble ascribing belief in such cases. However, we often do withhold belief ascriptions in such cases--failing to act on a professed belief makes us skeptical about the actual possession of the belief.[30] I suggest that our hesitation to ascribe belief here is evidence enough to demonstrate that both dispositions to theoretical reasoning and to practical reasoning and action are required for belief. And our analysis suggests that such cases **should** be difficult--it is not as though they are the same as when there in no belief at all. What we might expect, then, is a shift from the language of belief to other conceptions which distinguish these two aspects, and this is apparently what we find. We describe individuals of the sort envisioned as having accepted the proposition "in theory but not in practice." Or we might say they have "become convinced" of the claim but have "failed to

30. There are some interesting complications here. For instance, we may sometimes sincerely express an apparent belief that in fact is not something that we would either be willing to reason from nor act upon. Rey (1988) explores such cases a bit. I discuss his view along with Stich's evaluation of such cases in the Appendix (below.)

apply it." And significantly, we do not say things such as "they **believe** it in theory but not in practice"--this locution sounds odd in just the way that we should expect if belief requires dispositions to both "theory and practice."

Having clarified the computational functional role specification, I now turn to the problem of distinguishing previously held beliefs from dispositions to acquire belief. Given that having a belief is being disposed to use an appropriate explicit representation, if formulated, in theoretical reasoning and decision-making processes, the disposition to acquire a belief becomes a second-order disposition, namely the disposition to become disposed to use an appropriate representation, if formulated, in reasoning and decision processes. That is, belief is defined as a certain first order disposition *B*, and the disposition to acquire a belief is the disposition to acquire *B*.

The counterexamples that I presented above to the dispositional definition of implicit belief are supposed to show that in some cases an apparent disposition to acquire belief also satisfies *B*. To review, we have the opinionated people who are disposed to "immediately affirm" or deny any entertained proposition depending on the psychological context. We have the excited raconteur, who is talking too loudly. Were he to "entertain the proposition that he was talking too loudly, he would instantly realize that it is true." But he does not already believe it. And we have the

scientist who has never thought of some theory *T*, although it is so coherent with her other beliefs that when it is proposed to her she immediately affirms its truth. These cases show that believing that *p* cannot be a matter of being disposed to judge affirmatively that *p*, since in each case the disposition is present when, intuitively, the person has not yet acquired the belief--in each case they do not acquire it until they actually make the judgement.

Will these cases also serve as counterexamples to my proposed dispositional, causal role definition of belief? It is not apparent that they do, since it is not apparent that in the described cases the appropriate dispositions to use a representation in reasoning processes exist prior to the drawing of the judgement. What characterizes these cases is both an introspective immediacy (the subjects would "instantly" or "immediately" affirm the proposition) and a justificatory minimality--the affirmation requires but a single step of reasoning given the background beliefs or available evidence. However, this does not show that no change in dispositions for inputs to reasoning and action processes occur when the proposition is entertained. Consider that in the cases presented it is as reasonable to suppose that the judgement affects the causal role of the representation as it is to suppose this in cases of longer, more explicitly reasoning to belief change. That is, it is reasonable to hypothesize that the causal effect of the judgement is to

alter the person's dispositions to reasoning and decision input. Presumably, the opinionated people are not prepared to reason from or act on a representation meaning *p* or not *p* until the psychological event occurs which causes them to leap to the affirmation or denial of *p*. And likewise, the raconteur's and the scientist's judgments cause appropriate changes in their dispositions to reasoning and action. If they do not, then it seems more appropriate to characterize these as cases of affirming something that is already believed. Thus, if the scientist already accepts **every** consequence (deductive, inductive and explanatory) of the theory, then it is no longer intuitively obvious that she does not already believe the theory. And if the raconteur's entertaining the thought that he is talking too loudly does not change his dispositions to act on this representation then it seems more accurate to describe this thought as leading him to act on a belief that he already has rather than as changing his belief. E.g. suppose that the raconteur had been explicitly asked a few minutes earlier to speak louder by someone who is now gone, so that he now no longer needs to bellow, although in his excitement he does. Here, it is no longer clear that the entertaining of the thought changes his beliefs--we are more inclined to say that he has temporarily suppressed his belief that he is speaking loudly, or we may drop the notion of belief and turn instead to an explanation of his attention and access to stored information.

Perhaps there are cases like this where an explicit formulation does not change the functional role. But then it is not apparent that belief changes. For instance, consider a case of temporary forgetting that is apparently a failure of access. Suppose I believe some fact, say that Kant was from Königsberg, and that I have achieved this belief by explicit storing an appropriate token. But, suppose further that on a certain occasion, when asked, I cannot immediately recall this. However, I have not forgotten it permanently--when someone suggests that Kant may have been from Königsberg, I immediately affirm this--"ah, of course, Königsberg." It is plausible to assume that the failure was purely one of access--that I could not find the appropriate explicitly stored token when I first wanted to. Here it is apparent that the following conditional has remained true of me all along, if a representation that means that Kant was from Königsberg would be explicit formulated--i.e. available or on-line to central or appropriate processors, I would be disposed to use this representations in reasoning or decision processes. Thus, my suggested account classifies this as a case of continued belief. However, I do not think that this is inappropriate. If the mere suggestion of the right answer leads to my immediate recognition, we will probably want to ascribe continued belief, although belief seems an odd or inappropriate notion to use in characterizing this case--instead it seems more appropriate to just stick with an explanation of failure of

memory or access. My account suggests that the problem here is that the causal role for the belief apparently cannot come into play since the appropriate representation cannot be formulated or activated. Thus, we typically neither ascribe belief or non-belief in such a case, but speak instead of (a failure of) memory.

I suggest, then, that the dispositional, functional role account does not appear susceptible to the kinds of counterexamples that plague definitions of (implicit) belief that are formulated in terms of dispositions to judge. However, there may be a great number of cases where there is no distinguishing belief from dispositions to acquire belief. The problems arise in cases where the acceptability of an explicit representation for use by reasoning and decision processes is complicated, so that there will only be a certain probability that the representation will be used by the processes. Thus, suppose that a given reasoning processor (or set of processors) is disposed to use the output of another processor, an "evaluator," that takes a representation as input and uses a set of rules or heuristics to attempt to derive the representation from a given data base. If it succeeds, the representation is fed to the reasoners, if it fails, the representation is discarded. For some representations, the evaluative procedure may be highly complicated, and may depend on what other representations have recently been activated, or may depend probabilisticly on

which evaluation procedures are applied. For cases where the outcome is very doubtful, it does not seem appropriate to say that the reasoning processor is already disposed to use a given outcome. And where the outcome is very certain, it does seem that the disposition already exists. But notice that there is ro obvious dividing line between a likely outcome and an unlikely one. And this would mean that there would be no determinate point at which dispositions to use a representation if formulated could be distinguished from changes in dispositions to use. Unfortunately, on the suggested account of belief, this is to say that in such a circumstance there would be no exact dividing line between previously held beliefs and dispositions to acquire new beliefs.

Is the this an acceptable consequence of an analysis of belief? It appears that our intuitions about various cases support not only the possibility but the existence of such indeterminacy. Consider mathematical beliefs. With mathematical truths that are extremely simple to compute, e.g. 1000+3=1003 or 19 is the successor of 18, we are inclined to say that we already believe them. And when they are very difficult to compute, e.g. 77 is the square root of 5929, we are inclined to say the we don't believe them until we compute them, or unless we recall the result of the computation. But notice that there is no exact point in increasing computational difficulty at which previous belief ceases and

acquired belief begins. Thus, 85+33=118 is probably a borderline case, since, while the computation is easy, it is not so easy that we are guaranteed to get it right. Similar considerations hold for tautologies. We are willing to say that we believe that cows are cows, but probably not the truth-functionally valid proposition that either it's raining and grass is not green or snow is white and it is not raining or either snow is not white or grass is green, and it is indeterminate as to whether we already believe less difficult tautologies or not.

Another bit of support for the claim that in many cases belief and the disposition to acquire belief are indistinguishable comes from the fact that often there appears to be no determinate time at which a given belief begins or ceases to be. This is particularly true of very general beliefs. For instance, most readers probably believe that the methodologies characteristic of the "analytic" approach to philosophy are generally superior to the methodologies by characteristic of the "hermeneutic" approach to philosophy (or vice-versa.) But it is also likely that in most cases, there was no particular event, e.g. no conscious judgement, which marked the onset of this belief. Rather, it is likely that most individuals acquired this belief by gradually acquiring preferences for one tradition's literature over the other's, where there was no particular point in the study of the viewpoints that marked the onset of the belief.

Thus, it appears that we will have to give in to some extent to the worries that plague implicit belief accounts and allow for a certain amount of indeterminacy concerning the exact determination of which beliefs an individual has. Does this implied indeterminacy does show us that something is wrong with our account of belief? I think not. We set out to characterize a certain property and our investigation has revealed that it has indeterminate cases. No *a priori* dictum forces us to find that all properties are determinate in all cases. Nor is this the same sort of trouble we raised at the outset for the storage model of belief, for there is no reason to think that the dispositional model of belief misclassifies what appear to be clear cases of belief or dispositions to acquire a belief. Rather, as I have suggested, the model appears to reveal an indeterminacy in our ordinary conception of belief. Nor, finally, does this show that anything is wrong with the notion of belief itself. Many cases of belief and change in belief remain unproblematically distinct, which means that the concept is useful in ordinary descriptions and explanations of our psychology.[31]

31. The notion of belief may not seem indeterminate, since it is always possible to decide on a given proposition by consciously evaluating it. But this does not show that we can tell if the outcome of such an evaluation constitutes belief-change or not.

II. Other Attitudes, Explicitness and Causation

My analysis of belief suggests that there will be on the one hand dispositional, and on the other more explicit cognitive states, since belief, at least, is a disposition to use an explicit representation if formulated. In addition to the notion of belief, we need an explanation of what sorts of representations and processes are responsible for producing these dispositions, e.g. what representations are explicitly present or stored and how are they manipulated? Analogously, if we have a substance that is soluble, we still need an explanation of what makes it soluble, i.e. what in its chemical make-up produces the disposition to dissolve? Thus, we might suppose that cognitive psychology will postulate states that are different than belief in that they are states that require the explicit presence of representations.

Perhaps it might be suggested that other common sense attitudes will fulfill this role of non-dispositional explanation. However, it is fairly clear that the other most prominent attitude, viz. desire, is much like belief, and it is reasonable to think that it too will have a dispositional, computational specification. Specifically, it may be possible to have infinitely many desires (e.g. I want to live for more than 10 more years, more than 10.1 more years, more than 10.11 more years, etc.) And it also seems that we could have functional role twins that would exhibit the same desires as far as common sense ascriptions are concerned, despite radical

differences in storage. Attitudes such as hopes and fears may seem more occurrent, but we must be careful to distinguish feelings and attitudes. Feeling hopeful or feeling fearful is an occurrent state, but it is not propositional, any more that feeling tense or joyful is. E.g. I may feel fearful but not know what I am afraid of. Fearing that *p* or hoping that *p* are attitudes, but these again seem dispositional--while our thoughts or utterances may **express** such attitudes, they are not identical with them.

In general, any of the attitudes that last over long periods of time and which do not always have determinate beginnings and endings would seem to be good candidates for the dispositional model. This leaves us with a few occurrent attitude types, viz. occurrent thought, recognition, immediate intention, and several related states. But it is not as though we have a vast set of common sense explanations which deal only with these states. Nor is there any obvious means of transforming dispositional attitude explanations (e.g. those involving belief and desire) into explanations that cite only occurrent states. E.g. If I did A because I wanted *p* and believed that doing A was the best means to achieve *p*, then it is not always true and often false that I occurrently thought "doing A is the best means to achieve *p*." Thus, it seems that while we might acknowledge that there are some explicitly represented attitudes that play a role in cognition, we will have to look beyond the ordinary attitude types for a

characterization of most non-dispositional elements of cognitive architecture.

Notice that this is not to say that there is something bad or wrong about belief-desire states. All that I am claiming at the moment is that we should expect cognitive explanations to concern states that are different from most common sense states in that they are more explicit or occurrent and less dispositional. So this claim should bring no dissent from someone who is a belief-desire realist--who thinks that there really are beliefs and desires. However, this does effectively undermine the view that Fodor espouses of the attitudes, quoted at the outset, namely that attitudes are typically "tokenings" of symbols. The view I have just outlined suggests rather that most common sense attitudes are dispositions to token symbols, where the actual tokenings are not states that are normally characterized by common sense psychology. Let us then consider the general motivations and support of Fodor's view.

Fodor's advocation of belief-desire psychology is based on its explanatory success. As he puts it:

> Commonsense psychology works so well it disappears. It's like those mythical Rolls Royce cars whose engines are sealed when they leave the factory; only it's better because it isn't mythical. Someone I don't know phones me at my office in New York from--as it might be--Arizona. "Would you like to lecture here next Tuesday?" are the words that he utters. 'Yes, thank you. I'll be at your airport on the 3 p.m. flight' are the words that I reply. That's all that happens, but it's more than enough; the rest of the burden of predicting behavior--of bridging the gap between utterances and actions--is routinely

> taken up by theory. And the theory works so well that several days later and several thousand miles away, there I am at the airport, and there he is to meet me.[32]

Perhaps Fodor might attempt to defend the idea that belief-desire states are typically explicit cognitive states by invoking this argument from explanatory success. Suppose we grant that common sense belief-desire psychology has a high degree of success in terms of predicting and "coordinating" behavior. It might be argued that this stunning success shows that cognitive psychology will formulate explanations primarily in terms of common sense attitude states. And from this, together with the assumption that cognitive psychology will generally or typically offer explanations which postulate causal sequences of explicitly tokened representations, which is derived independently from the computer metaphor, it follows that most common sense attitudes must be states that are explicit tokenings of representations.

The fault with this line of reasoning concerns the move from belief-desire psychology's explanatory success to the claim that cognitive psychology will explain primarily in terms of belief-desire states. This assumes that the realm of facts that cognitive psychology explains consists largely of facts concerning the prediction of behavior, particularly the coordination of behavior with utterances. However, it appears that there is much else that psychology should explain, as

32. Fodor (1987), p. 3.

Churchland forcibly notes ('FP' is common sense belief-desire psychology):

> As examples of central and important mental phenomena that remain largely or wholly mysterious within the framework of FP, consider the nature and dynamics of mental illness, the faculty of creative imagination, or the ground of intelligence differences between individuals. Consider our utter ignorance of the nature and psychological functions of sleep, that curious state in which a third of one's life is spent. Reflect on the common ability to catch an outfield fly ball on the run, or hit a moving car with a snowball. Consider the internal construction of a 3-D visual image from subtle differences in the 2-D array of stimulations in our respective retinas. Consider the rich variety of perceptual illusions, visual and otherwise. Or consider the miracle of memory, with its lightning capacity for relevant retrieval. On these and many other mental phenomena, FP sheds negligible light.[33]

We need not, however, draw the conclusion that Churchland does, namely that we should abandon common sense belief-desire psychology in favor of alternative (e.g. neurological) accounts.[34] A reasonable reply here on the part of the belief-desire/cognitive realist is that belief-desire psychology only explains a certain range of facts. Those that Churchland cites might plausibly be explained by representational states and computations other than those of common sense belief-desire psychology. And, in fact, various computational/representational theories have been developed to

33. Churchland (1981), p. 73.

34. I briefly criticize Churchland's eliminativist position in the next section.

explain many of these types of facts, including theories of vision and perception, memory, and motor control.

This suggests, however, that cognitive theory may offer explanations which postulate quite a wide range of states in addition to belief-desire states. Specifically, it seems plausible to suppose that abilities and specific states (perhaps including belief-desire states) will be explained in terms of explicit representational states that are unknown to common sense. And it is consistent with this assumption to hold that most common sense states are dispositional. Thus, the explanatory success of belief-desire psychology does not show us that common sense states must be mostly explicit, since it is success for a limited range of explananda. There is no reason to think, and some good reasons against thinking, that cognitive theory will be little more than a cleaned up version of common sense belief-desire psychology.

Another main motivation of Fodor's in holding the view that belief-desire states are explicit cognitive states is the need the explicate the causal efficacy of the propositional attitudes. As we have seen at the outset, the causal role of symbols allows us to see how attitudes can cause behavior if they are explicit, symbolic states. Moreover, Fodor would apparently insist that this is the **only** way we can explicate the attitudes' causal efficacy. As he puts it, "no intentional

causation without explicit representation."[35] Is there any basis for this insistence on explicit representation?

Fodor's support for this view appears to be the more general claim that only explicit, occurrent states can be causes:

> *Qua dispositional*, attitudes play no causal role in *actual* mental processes; only occurrent attitudes--for that matter, only occurrent *anythings*--are actual causes.[36]

The reasoning would appear to be that since causation requires explicit causes, intentional causation must require explicit representation.

However, this position stands or falls with the explicitness or dispositionality of attitude concepts--what we have investigated above in the case of belief. To see this, first consider that it appears that there can be true causal **explanations** which cite dispositional rather than explicit states. For instance, consider the assertion that the glass broke because it was brittle. There is no apparent reason to deny that this is a causal explanation. For instance, the appropriate counterfactual seems true if the former statement is, i.e. had the glass not been brittle, it would not have broken.

Further, consider that by anyone's estimates, and by Fodor's own admission, there will be many ordinary belief-

35. Fodor (1987), p. 25.

36. Fodor (1987), p. 22.

desire explanations which will be apparently true seemingly causal explanations, but which will concern dispositional attitudes rather than occurrent ones, since the number of attitudes that figure in ordinary explanations appear to outrun the number of high-level occurrent states we can expect to find in cognition. But there is no reason for claiming that such explanations are not causal. In particular, it seems that this (postulated) class will not exhibit any less explanatory success than the (postulated) class concerning core cases, since success is pretty much uniform throughout (apparently) true attitude explanations. And there is no obvious sub-class of ordinary attitude explanations whose members seem any less causal than all other attitude explanations. Thus, it would be purely arbitrary to deny the causal status of dispositional attitude explanations.

However, we can grant Fodor that in cases of dispositional causation, there would always seem to be an additional, explicit/occurrent cause present. As noted in the opening of this section, it seems that a second, non-dispositional explanation always underlies a dispositional explanation. For instance, if a substance broke because it was brittle, then there is an explanation of the breakage which cites the structure of the physical substance. Yet, the dispositional explanation may be more useful for at least two reasons. First, we may be interested in very abstract features of the substance. E.g. brittleness or solubility are qualities

that a large number of otherwise physically and chemically different substances share. And second, we may not know the structural details of the substance in question, but we may still be able to attribute at least one abstract feature of the substance.

If we grant that in cases of dispositional attitudes, there is another, explicit underlying cause, then Fodor's position amounts to the claim that the state that is the underlying cause must have the same content as attributed to the dispositional attitude. But there is simply no basis for this claim, particularly when we note that on Fodor's view, it is not the content that is the actual cause of the behavior, but rather the syntactic properties of the psychological state. If content is, in this sense epiphenomenal with regard to behavior, then why insist that the occurrent, causal state must have this content? Here, I think the only reason is that it is Fodor's hope that belief-desire explanations will turn out to be not merely true, but at the very heart of the theories of cognitive psychology. If the explicit causal states underlying behavior do not turn out to have the content assigned by the relevant belief-desire explanations, then the latter might be relegated to the back-burner of explanatory psychology, as strictly-speaking true, but uninteresting explanations in comparison to the explanations which cite the

underlying, explicit (e.g. informational) states.[37] But hopes do not provide justification--i.e. once we see that most common sense attitudes including belief and desire are dispositional states, there is no longer any basis for the claim that all common sense attitude causation requires explicit representation.

Therefore, we cannot have the union of computationalism and common sense psychology that Fodor envisages. If the dispositional view is correct, then belief-desire explanations are true because there are other explicit "tokenings" of representations which interact appropriately. On some occasions, these tokenings may be of representations with the same content as the associated attitudes. But it may also be that on many occasions the tokenings and the attitudes do not correspond in this way. Thus, I suggest that we do not really, as yet, have any clear idea of the nature of the explicit states that drive cognition. And determination of these states is the difficult task of cognitive psychology. Until we know what the explicit states of cognition are like and how they interact, we will not know exactly what makes common sense explanations true. But this should not be particularly surprising. In other areas of science, the explication, vindication, and, where necessary, reform and rejection of common sense explanations has proven to be a long and

37. See Fodor (1987), pp. 23-24 for expression of something similar to this worry.

difficult matter. It would be truly amazing if this did not happen in psychology as well.

III. Implications

I will now explore three implications of this dispositional picture of the nature of belief-desire states. The first concerns how we are to respond to the claims that have provided the basis for non-realist views of the attitudes. I will argue that, on the view I have just presented, we can accept that there are apparent features of the attitudes that make attitude notions inappropriate for scientific theories, without going as far as eliminativism or instrumentalism with regard to the attitudes. Second, I will examine Fodor's claims concerning the holistic and non-modular nature of the central systems. I will argue that the dispositional view shows how we might simultaneously have epistemically holistic attitude properties, and computationally modular elements at the heart of cognition. Third, I will examine the commitments behind a realist view of the attitudes, specifically Fodor's claim that realism with regard to the attitudes leads to the hypothesis of an innate language of thought. I will argue that while it is likely that adult humans frequently think in an internal version of their spoken language, there is no reason to postulate an internal language for all instances of propositional attitudes, in humans, animals, etc. I will also examine Fodor's argument for

the innateness of most concepts and show that the present account of the attitudes allows us to see how most concepts might be both partially innate and partially acquired.

A. The Status of Belief-Desire Psychology

So far I have been working within scientific realist assumptions regarding belief-desire states. That is, I have been assuming that some scientific, computational account can be given of belief-desire states. And this assumption appears to be at least partially vindicated by the dispositional account of belief that I have presented above. With this account in hand, I will now critically examine non-scientific-realist views of belief-desire psychology.

There are three main alternatives to scientific realism concerning the attitudes that are to be found in the literature. They are: 1) **eliminativism**. The eliminativist holds that there are no beliefs, no properties corresponding to commonsense attitude psychology's postulated states, and envisions a scientific psychology which will operate with completely different notions. 2) **Explanatory dualism**. This view maintains that there are states corresponding to common sense belief-desire ascriptions and explanations, but that the investigation of such states must be completely independent from the investigation of the states that scientific psychology concerns itself with. 3) **Instrumentalism**. This view is like the previous two in holding that scientific psychology

will not be concerned with the states belief-desire psychology postulates, but attempts a subtle line between realism and eliminativism as far as the ontology of belief-desire states are concerned, claiming that while there really are no such states, the postulation of them proves useful in ordinary explanations.

A majority of the arguments for these views concern the issue of content. The view I have developed does not involve any new claims about content *per se*. However, as I shall now suggest, it appears that there has been a failure to distinguish characteristic features of attitude states and means to knowledge of them from the content the states appear to have. When we make such distinctions, the case against scientific realism for the attitudes, as far as content is concerned, appears quite weak.

First, I think it is fairly clear that complete eliminativism with regard to representations is implausible. Examinations of connectionist models--which have been offered as an alternative to attitude explanations, e.g. by Churchland--suggest that these theorists are actually committed to some notion of content.[38] And it seems that a complete eliminativism is somewhat incoherent--what is the eliminativist doing? Surely not **asserting** a negative **theory about** representations--these are, after all, concepts that are

38. See Fodor and Pylyshyn (1988).

part of the representational realist framework. Thus, the full thesis of representational eliminativism would seem to constitute a *reductio* of the view.[39] So eliminativism with regard to content seems highly implausible, if not downright incoherent.

It has also been argued that psychological explanations involving states with content are inappropriate for science because of certain features of such states. The two most notably problematic features are context dependence and normativity. It has been argued that what attitudes a person can be said to have depends on the physical or social environment that they are situated in.[40] And some have taken this to show that scientific psychology cannot involve explanations which postulate states with content.[41] Also, it has been maintained, e.g. by Quine and by Davidson, that what beliefs and desires a person has is a matter of how we interpret their behavior. And interpretation is thought to involve an essentially normative element--we have to see others as rational by our lights. This normative element,

39. See Baker (1987) Chapters 6-7 for development of this argument.

40. Most notably by Putnam (1975) and Burge (1979), (1986).

41. Most notably, Stich (1983).

Davidson and others claim, is unacceptable for scientific theories.[42]

We must distinguish two very different claims that might be made here. First, let us distinguish the representational states that theories in scientific psychology ascribe from common sense representational states, i.e. the propositional attitudes. It might be claimed that all representational states, common sense or otherwise, have the (apparently unscientific) features of context dependence and normativity. This, however, is a difficult claim to defend. Consideration of the ascription practices and intuitions of common sense belief-desire psychology is not necessarily relevant to such alternative representational states unless it can be shown that common sense states necessarily share appropriate features with the representational states that scientific psychology will concern itself with. But given that we may not as yet have discovered the appropriate representational states for scientific psychology, it is difficult to see how this could be accomplished. Surely, nothing in the literature provides any basis for such a claim.

42. See Essays 11-13 in Davidson (1980). Also see Putnam (1988) Chapter 1. Such authors also argue that belief-desire psychology is unscientifically "holistic" and "indeterminate." Indeed, these properties are typically explained in a highly inter-related way. What I have to say about normativity in what follows, which is mostly a strategic point, applies equally well to holism and indeterminacy.

Moreover, the positive basis for the claim that psychology will cite representational states as part of theoretical explanation is simply the existence of actual, successful theories which do so. And since there are at least some such theories that claim at least modest success, it would seem that either context dependence or normativity are scientifically acceptable features of states, or psychological theory concerns representational states that do not have these features. Thus, it would seem that the burden of proof is on the non-realist to show that any theories that appear to cite representational states are either no good or are not actually representational. And again, such a case is hardly forthcoming.

A weaker claim on the part of the non-realist might be that the content attributed by common sense psychology is unacceptable for scientific theorizing. Thus, such a position might maintain realism or agnosticism about non-common sense representational theories, which rejecting the scientific acceptability of the ordinary attitudes. However, there are two further problems with such a position. First, it might be the case that while ordinary ascription practices have certain apparently "unscientific" features, such as context dependence or normativity, it is possible to theorize about such states in ways which dispense with such unscientific features. Thus, just because we sometimes use normative means to determine which attitudes someone has, it does not follow that those

individuals' attitudes must be ascribed by such means. Perhaps attitudes are like most other natural kinds, in that the common sense means of identifying their instances and determining their features will give way to alternative scientific theories which do not fully incorporate the ordinary means of identification. (E.g. think of the difference in methods for identification of substances--water, salt, etc.--between common sense and scientific chemistry.) Thus, no amount of consideration of ordinary ascription practices shows that no such reform is possible. Indeed, it is possible that a property instantiation explanation of the sort I have sketched for belief above will be part of such alternative means of specification.

A second problem for the line that common sense attitudes have features that make them unacceptable for scientific psychology concerns the role that such states will play in psychology. As the dispositional account suggests, belief-desire states may be very abstract in comparison with explicit cognitive states. It may turn out as a matter of practical necessity that the only reasonable means of determining the presence of such states is through ordinary, e.g. normative or context-dependent, methods. But this practical limitation does not show that no in principle cognitive computational account of belief-desire states can be given. And it seems clear that the opponent of scientific realism must rule out not only practically feasible specifications of belief-desire states,

but also in principle specifications that are not practically feasible.

Perhaps explanations involving belief-desire states will not prove generally feasible for scientific interests. Perhaps most "strict" psychological laws will be formulated in terms that refer to explicit representations rather than the more dispositional representations of belief-desire psychology. This would not affect scientific realism about belief-desire states one bit. As I understand it, scientific realism about belief-desire states merely requires the very weak claim that **whatever** sorts of notions (e.g. representational, computational) prove acceptable for scientific psychology, reductive property specifications of belief-desire states can be given in those notions. Scientific realism, therefore makes no claim about the explanatory role of belief-desire explanations in scientific psychology. Specifically, there is no claim about such concepts being paradigm psychological explanations. Thus, the fact that belief-desire explanations may "feel" unscientific because of context dependence, normativity, etc. does not in and of itself provide any basis for views which oppose scientific realism about the attitudes. What must be shown, to repeat, is something much stronger, that no in principle scientific psychological account of belief-desire states can be developed. This is a very difficult claim to defend for **any** property, and I do not see that the considerations of context dependence, normativity,

etc. have done much to support it in the case of the attitudes.

The fact that non-scientific-realism about belief-desire states is such a difficult position to defend may have been obscured by the fact that some of the most noted belief-desire scientific realists, Fodor in particular, not only think that psychology can recognize belief-desire states but that belief-desire explanations will be at the core of scientific psychology--or, e.g. that scientific psychology is to be virtually **identified** with belief-desire explanation, that many or most scientific psychological theories will resemble common sense generalizations. That is a much stronger position which does seem susceptible to the detection of apparently non-scientific features in common sense psychological concepts and explanations. But the present point is that there is no reason to go to the lengths of non-realist views, e.g. eliminativism or instrumentalism in order to avoid this stronger position. It is perfectly reasonable to maintain that while belief-desire states will be accounted for by scientific psychology, most of the "strict" laws and the best explanations of scientific psychology will not involve common sense belief-desire notions.

I conclude this sub-section with some specific criticisms of each alternative to scientific realism.

Eliminativism would seem to be an over-reaction to the intuition that common sense belief-desire concepts will not

play a central role in scientific psychology. We can grant this, yet nothing about the non-existence of beliefs and desires follows, just as nothing about the non-existence of common sense observational properties such as "warm" or "heavy" follows immediately from the mere fact that physics does not use them in causal explanations. Further, belief and desire are not clearly theoretic notions, like ether and no phlogiston, which were developed as part of science.[43] Rather they are deeply ingrained, perhaps innate, common sense notions that most people attain in the course of normal development. Given that belief-desire explanations achieve certain successes, at least in ordinary applications, it seems extremely unlikely that we will ever dispense with them completely, if indeed we can.

The explanatory dualist, most notably Davidson, holds that there are two separate forms of explanation, viz. common sense attitude psychology and cognitive psychology[44], but claims that the two are methodologically disparate. This idea is expressed fairly clearly by Putnam, who writes:

> To have a description of how a system of representations works in functionalist terms is one thing; to have an *interpretation* of that system of representations is quite another thing.

43. See Clark (1978) for defense of the view that belief-desire psychology is not a theory, any more than common sense biology or physics are theories.

44. Or, in Davidson's case, brain science--he apparently refuses to acknowledge the existence of non-common sense psychology, which makes it difficult for the cognitivist to engage with his writings.

> The difference between functionalist psychology and interpretation theory is in part due to this: functionalist psychology treats the human mind as a computer. It seeks to state the rules of computation. The rules of computation have the property that although their interaction may be complicated and global, their action at any particular time is local. The machine, as it might be, moves a digit from one address to another address in obedience to a particular instruction, or to finitely many instruction, and on the basis of a finite amount of data. Interpretation is never local in this sense. A translation scheme, however well it works on a finite amount of the corpus, may always have to be modified on the basis of additional text.[45]

This in and of itself need not be particularly problematic. The further claim that the explanatory dualist makes, though, is that the ontology that common sense psychology concerns itself with, i.e. beliefs, desires, etc., cannot be dealt with in any way by cognitive methodology.[46] That is, the view is that cognitive psychology will never be able to explain what beliefs are, nor have anything to say about what beliefs we in fact have. As with eliminativism, this seems like an over-reaction, in this case to the apparent divergence in methodologies. An alternative possibility is this: It might be that when we assign attitudes holisticly, normatively or whatever, we are getting things wrong--we are only approximating the actual attitudes that we have. It may be

45. Putnam (1983), p. 150. While Putnam does not explicitly endorse the thesis that I have title "explanatory dualism," this seems to be one of the implications he draws from his explications of the differences between interpretation and computation.

46. Putnam (1983), chapter 8, hedges on this a bit. See pp. 150-154.

that no perfection of interpretation methodology will ever "mechanize" it, but that may be because we need to alter methodologies and take a wholly computational approach to explaining the attitudes. Unless one assumes that the attitudes cannot be characterized except by ordinary interpretation practices, which begs the question, then this sort of scientific approach seems like an open possibility. Nor would this mean that we would adopt a computational approach in our ordinary interactions. It may be that our normal sources of evidence to one another's psychologies (including our own) are so limited that a normative, holistic means of ascription is the only practical means of approximating correct characterizations of our psychologies. Further, it may be that while we will never able to formalize interpretation theory, we will gradually become able to answer various questions in semantics, epistemology and the like through the substitution of computational explanations for explanations which rely solely on common sense intuitions and ascriptions practices.

In general, the main reason for thinking that attitude explanations and cognitive theory will not be completely unrelated is that the both attempt explanations of some of the same phenomena, namely behavior, and do so with concepts at approximately the same level of abstraction (compared to, say, neural concepts.) Since it is difficult to find other areas in which completely ontologically unrelated explanations of the

same phenomena coexist, where the explanations are at about the same level of abstraction, it is reasonable to suppose that the states characterized by attitude psychology and cognitive theory will likewise not be unrelated.

Similar problems plague an instrumental view of the attitudes. Dennett holds[47] that there are three distinct sorts of explanatory "stances", the physical stance, the design stance and the intentional stance. Each of these is understood as a certain sort of explanatory strategy concerning a given system:

> The physical stance[:] If you want to predict the behavior of a system, determine its physical constitution 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...
>
> Sometimes...it is more effective to switch from the physical stance to what I call the design stance, where one ignores the actual details of the physical

47. While Dennett has portrayed himself as instrumentalist for many years, he has recently withdrawn this view and has moved towards an explanatory dualism. See "Instrumentalism Reconsidered" in Dennett (1987). Here I shall discuss his (earlier) instrumentalist views.
While Dennett was a pioneer in the development of the view that psychology is "sub-personal", and has always stressed the significance of non-attitude cognitive explanations, he has equally resisted the idea that the attitudes may be explained by features of sub-personal psychology. This seems to be because he assumes that explicit storage and manipulation models are the only plausible candidates for realism about the attitudes. For instance, in Dennett (1987) on p. 70, (4), after making the point that beliefs are not to be identified with whatever explicit representations we possess, he seems to draw the conclusion that there are no beliefs, rather than (what I consider to be) the obvious alternative, namely that the property of belief is to be identified with some other sub-personal property or properties, e.g. dispositions to certain explicit states.

> 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...
>
> Sometimes even the design stance is practically inaccessible, and then there is yet another stance or strategy one can adopt: 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, given the same considerations, and finally you predict that this rational agent will act to further its goal in light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many--but not all--instances yield a decision about what the agent ought to do; that is what you predict the agent will do.[48]

While this is surely an excellent basis for a theory of the nature of various sorts of explanation, particularly as regards systems' explanations, it is not clear why this is an answer to questions about the ontology of the attitudes. Are the ascriptions made under such stances **true**? If not, in virtue of what **features of the world** are they successful? Instrumentalists commit themselves to a negative answer to the first question, but then an answer to the second becomes difficult. I.e. what, if not beliefs and desires, explain the success of belief-desire explanations? The only plausible answer I can see to this for the instrumentalist is the following application of the notion of a "stance":

> when operating within a given stance, the ground rules of attribution are completely characterized **within** the stance. But it's a mistake to think that

48. Dennett (1987), pp. 16-7.

> you can answer ontological questions outside of the stance. When you're in a stance you're in a different "world." Thus, the question in dispute has neither a positive nor a negative answer from the general physical framework. It's an unaskable question.

But this line is surely mistaken. When we adopt the design stance the **ontology** doesn't change. Auto mechanics don't literally spend their days in a different world. When someone successfully explains something using the design stance, we understands the predicates he uses as referring to things in the physical world. Carburetors and engines (and tables and chairs, for that matter) exist just as surely as do hunks of metal. Ultimately, we may want to tell an instantiation story--functionalism looks like it will succeed for most reasonably developed design-types. Thus, a carburetor is a functional type which is potentially instantiated in systems meeting the **physical** parameters of "engine theory." And similarly for functional kinds terms in biology. Thus, it seems that there are independent ontological answers to questions about he existence of design-types outside of the stance. In general, this is because explanation is only a part of our knowledge. We can know of things and conceive of them apart from particular explanations. So, the instrumentalist owes us an answer to the ontological questions surrounding the attitudes.

I conclude that a moderate scientific realism, which claims that scientific psychological theories of computation and representation will ultimately explain the nature of

belief-desire states, and may, but need not, include common sense belief-desire explanations, stands as the most plausible view of the relationship of common sense belief-desire psychology to scientific psychology.

B. Modularity and Cognitive Architecture

A second implication of the cognitively dispositional view of the attitudes concerns the question of cognitive architecture--what kinds of states and processes we will find in cognition. Fodor has suggested that cognition consists of central systems which embody attitude states and input systems consisting of informationally isolated "modules."[49] The input systems are conceived of as collections of units or modules which process perceptual input and produce tentative perceptual representations, including parsed utterances. Specifically, he characterizes modules as **informationally encapsulated**. That is, the information inside them--other than their output--is not typically available to the rest of cognition, they have a limited amount of information, i.e. rules for processing the input--and do not draw on other outside information. Other notable features of most modules are that 2) they are domain specific--i.e. they operate in only a specific, limited area of knowledge, 3) their operation is mandatory, 4) there is limited central access to their

49. In Fodor (1983). Page numbers in the rest of this subsection refer to this work.

contents, 5) they are (very) fast processors, 6) they have relatively simple inputs, 7) they are associated with fixed neural architecture, and they have 8) characteristic and specific breakdown patterns and 9) characteristic pace and sequencing.

The central systems are conceived of as not being divided into modules, but rather as being a more or less homogenous collection of states that are similar, if not identical to the common sense attitudes. The central systems are viewed as informationally unencapsulated, and also as not having the (nine just-listed) properties thought to be characteristic of modules. So the model has it that input gets processes by the modules and it is their output which is used by the central systems in producing belief-desire states, which ultimately lead to actions in the usual manner (e.g. beliefs and desires combine to cause actions.)

Fodor's thesis might simply be that input systems are the only cognitive units that share all of these (nine) features. With this I have no quarrel. But he also seems to be arguing something stronger, namely that the central systems will not be divided into any sort of units at all, specifically not informationally encapsulated units. He suggests that the holistic features of belief-desire states make the central systems poor candidates for the type of divide-and-conquer explanatory strategy that appears successful for modular systems. What I will now argue is that once we recognize that

many common sense states will be dispositional states of cognitive architecture rather than explicit states, it is possible to see how the central systems might have both holistic properties and explicit, informationally encapsulated units. That is, I shall argue that once we adopt the dispositional view of the attitudes, we no longer have any reason to reject the idea that the central systems might contain a substantial number of modules.

Before I consider Fodor's arguments against central system modularity, I want to motivate the idea that there may be informationally encapsulated non-perceptual units in cognition. Fodor claims that "there is practically no direct evidence, pro or con, on the question whether central systems are modular" (p. 104.) However, there is some significant, *prima facie* indirect evidence. Following Chomsky, we should note that in general it seems plausible to postulate a domain specific module when most of the following cluster of properties are fulfilled for a given knowledge domain: The knowledge in question is fairly readily distinguished from other knowledge, the computations underlying this knowledge are fairly specialized (compared to e.g. general reasoning), explanation of the acquisition of the knowledge requires postulation of domain specific innate principles, principles that we are not consciously aware of and possession or lack thereof of this knowledge seems independent of general intelligence. That is, it would seem that when most or all of

these conditions are satisfied, the postulation of a module for approximately the domain in question is a good explanation of these facts.

For example, linguistic theory postulates a universal grammar to accounts for humans' ability to readily acquire a language (at a certain age) with only minimal exposure to data. The attained knowledge of a grammar is thought to be highly specialized non-conscious knowledge of the structures of sentences. A partial, plausible explanation of the cognitive realization of this knowledge is that we have a "language module" which develops a representation of a grammar through the setting of "hard-wired" parameters--i.e. through setting switches or selecting features within a fixed set of options. In such a case, the background set of possibilities determined by all possible parameter settings forms a limit on knowledge of the domain, or of the domain itself, and the initial settings plus means for determining the settings forms a universal knowledge of the domain. Thus, we may postulate that the fixed set of options in the language module constitutes the domain of possible (spoken) languages, and the initial settings in the language module, and the means whereby parameters are fixed, constitute a universal grammar or universal knowledge of (certain abstract aspects of) language. So it appears that the postulation of a language module is a plausible hypothesis about cognitive architecture that

provides a (partial) explanation of some of the significant facts about our knowledge of language.[50]

Plausible candidates for domain specific central modules include language (or a set of linguistic sub-modules including syntax, a lexicon, phonetics and perhaps also pragmatics),[51] musical abilities,[52] mathematical knowledge and ethical knowledge. Research in these areas has, to some extent, revealed distinctive sets of principles or competencies, typically too complex to be acquired in a general way with the small amount of training many people receive.[53] So it would

50. Chomsky has presented these views in various places, beginning with Chomsky (1965), Chapter 1. See also Chomsky (1986) and see Chomsky (1980) including a discussion of the modularity of cognitive architecture ("mental organs," as he calls them) in chapter 1.

51. As Chomsky has argued, it seems that language is not merely a "peripheral" input system as Fodor suggests, but is rather a more central module. This is because it appears that knowledge of language must be used both in input and output, and perhaps in thought as well. See Chomsky (1986) p. 14 fn. 10. Fodor tends to minimalize the role of a grammar in central thought since he also hypothesizes a(n innate) language of thought. A more reasonable alternative as I will suggest below, is that when human representations are linguistic, it is because they make use of the same knowledge of language which is used in comprehending and producing (external) speech acts. This would, in turn, suggest that the language module is centralized.

52. Note that while there is clearly an input element to our musical abilities, musical imagery is central. E.g. the claim that Beethoven's input systems composed his symphonies is ridiculous.

53. For the idea of competence principles for language in general, and for an introduction to the thriving research program in syntactic competence, see Chomsky (1986). For an initial attempt at a (partial) semantic competence theory, see Jackendoff (1983). On competence in musical abilities, see Lehrdahl and Jackendoff (1983). On competence in mathematical

seem to be reasonable, *prima facie*, to postulate centralized modules corresponding to these areas of knowledge.[54]

Why, then, couldn't there be informationally encapsulated units in the central systems as well? Fodor argues against central system modularity through an analogy between belief fixation and confirmation theory. He distinguishes two apparent features of confirmation which he suggests are characteristic of processes of belief fixation as well, viz. they are **isotropic**--any knowledge may apply to a given problem--and they are **Quinean**--a change in any one representation may potentially affect any other representation. He then argues that these holistic features imply a lack of informational encapsulation, and thus a lack of modularity:

> When we discussed input systems, we thought of them as mechanisms for projecting and confirming hypotheses. And we remarked that, viewed that way, the informational encapsulation of such systems is tantamount to a constraint on the confirmation metrics that they employ; the confirmation metric of an encapsulated system is allowed to "look at" only a certain restricted class of data in determining

knowledge, in children, see Gelman and Gallistel (1978/1986), Gelman, Meck, and Merkin (1986) and Greeno, Riley and Gelman (1984). And on the idea of an ethical knowledge module, see Rawls (1971), pp. 46-48. Kohlberg's work might also be viewed as a competence theory for ethics, although he does not explicitly endorse this view. See Kohlberg (1969) and (1981).

54. See Cam (1988) for the argument that split-brain research and facts about our self-attribution provide support for central system modularity. I agree, although I think that this evidence at most supports the postulation of one particular module, of (roughly) self-attribution.
I discuss this evidence briefly in the appendix in regard to Stich's and Rey's treatment of it.

> which hypothesis to accept. ...encapsulation implies constraints upon the access of intramodular processes to extramodular information sources. Whereas, by contrast, isotropy is by definition the property that a system has when it can look at anything it knows about in the course of determining the confirmation level of hypotheses...
>
> ...Quinean confirmation metrics are ipso facto sensitive to global properties of belief systems. Now, an informationally encapsulated system **could**, strictly speaking, nevertheless be Quinean. Simplicity, for example, could constrain confirmation even in a system which computes its simplicity scores over some arbitrary selected subset of beliefs. But this is mere niggling about the letter. In spirit, global criteria for the evaluation of hypotheses comport most naturally with isotropic principles for the relevance of evidence. Indeed, it is only on the assumption that the selection of evidence is isotropic that considerations of simplicity are **rational** determinants of belief. It is epistemically interesting that H & T is a simpler theory than -H & T where H is a hypothesis to be evaluated and T is the rest of what one believes. But there is no interest in the analogous consideration where T is some **arbitrarily delimited** subset of one's beliefs. Where relevance is non-isotropic, assessments of relative simplicity can be gerrymandered to favor any hypothesis one likes. This is one of the reasons why the operation of (by assumption informationally encapsulated) input systems should not be identified with the fixation of belief; not, at least, by those who wish to view the fixation of perceptual belief as by and large a rational process.(pp. 110-111.)

I do not think that these arguments show that there are no informationally encapsulated modules in the central systems. To see this, suppose you have a number of encapsulated modules connected by an expert system. The expert system must test a hypothesis using the modules. How does it deal with isotropy--the need to possibly access anything it knows? Simple, it feeds the hypothesis to one module after another--this way the relevant (by hypothesis) information will eventually be

reached. And how does it deal with Quineanism--the fact that a change in one representation may potentially effect any other representation? Again, simply by making the output of a given module generally available to others. That way what happens inside one may potentially effect any representation in any other module.[55] So it would seem that the isotropic and Quinean nature of hypothesis confirmation does not show that the central systems do not consist partly or even largely of informationally encapsulated units.

Now, I have not given a plausible model of hypothesis testing involving encapsulated modules, but that cannot be the demand, since that is tantamount to asking for a fairly complete theory of the central systems, and this we haven't got yet. And until we do, any sort of skepticism is certainly possible. But what seems to drive Fodor's arguments here is not general skepticism so much as the pursuit of a specific sort of identification, viz. finding a module that, on its own, will test hypotheses, or, as the second quoted paragraph clearly shows, a module that can be identified as the "belief

55. In the second quoted paragraph, Fodor appears to claim that simplicity could not be computed with anything short of an entire belief-set. But this is dubious, given that our typical belief-sets are infinite or at least out-run any capacities that our processors seem capable of handling, yet we can compute simplicity. If the claim is that no encapsulated computation could be relevant to the determination of simplicity, then this requires a lot of support that is not forthcoming. I.e. we are really not sure what simplicity is, to say nothing of knowing how to compute it in a model of cognitive processing.

fixation module." All his considerations show, then, is that it is unlikely that you will find **local** identifications of belief-desire processes with cognitive processes. But this is unsurprising, once you abandon an explicit storage model of belief in favor of a dispositional model. Just as there is no explicit state corresponding to a given belief, so there is no explicit process corresponding to belief fixation. Rather, to fixate a belief is to go through any number of "sub-personal," potentially modular processes, as long as you get the right result, i.e. a disposition to use an explicit representation, if formulated, in inference and decision.

Note that it is also plausible to think that we have some processes that operate on very broad data bases, since, e.g., we are able to reason about just about anything. But this is not incompatible with there also being a lot of modular, specialized units in the central systems.[56] So it would seem that consideration of belief fixation does not show that the central systems are non-modular, but only that belief and belief-fixation cannot be directly identified with such modules or their contents. But that is perfectly consistent with the dispositional view of belief I have defended above. Thus, a dispositional account of the attitudes allows us to

56. An interesting proposal which integrates these ideas is Baars' (1988) functional model of consciousness which identifies conscious states with a generalized workspace that takes inputs from various, competing (or cooperating) modules and makes its output available to all or various modules.

see how a number of separate units might interact to jointly achieve holistic belief properties, and, *ipso facto*, beliefs.

Fodor presents two other arguments against central system modularity. The first is that (what he calls) the frame problem in AI,[57] viz. "which of my beliefs ought I to reconsider given the possible consequences of my action,"(p. 114) shows the inability of "local", i.e. modular theory to deal with the central systems' features. As he puts it:

> If we assume that central processes are Quinean and isotropic, then we ought to predict that certain kinds of problems will emerge when we try to construct psychological theories which simulate such processes or otherwise explain them; specifically, we should predict problems that involve the characterization of nonlocal computational mechanisms. By contrast, such problems should not loom large for theories of psychological modules.(p. 117)

Again, this argument seems to rest on a faulty inference from properties of processes to properties of individual processors. If we grant that many (though, no doubt, not all, perhaps not even a majority) of central system processes are holistic (Quinean and isotropic), it does not follow that any given central system processor must have these features as well. As we have just seen, it seems perfectly possible for non-holistic modules to collectively produce holistic properties. The fact that central systems (apparently) have

57. There is little agreement about what the frame problem is, let alone how to solve it if it exists. See Pylyshyn (ed.) (1987) which includes a criticism of Fodor's view of the frame problem by Hayes, along with a response from Fodor.

holistic properties that the input systems lack may be due either to the fact that there are some non-modular processors in the central systems or to the fact that while both systems ar largely modular, the central systems are organized differently than the central systems. In either case, we may still have a lot of informationally encapsulated modules in the central systems, perhaps as many as we do in the input systems.

The organizational approach to the (apparent) holistic features of the central systems is consistent with a heuristic approach to the (apparent) frame problem. This approach involves abandoning the pursuit of a principled solution to this problem, and instead attempting to design local systems that collectively accomplish appropriate solutions "heuristically," i.e. by reconsidering appropriate beliefs in most situations where reconsideration is required. This doesn't produce a **principled** solution to the frame problem, but rather concedes that it is not solvable within belief-desire theory, i.e. by allowing that you can't formalize principles of belief revision in terms of strict laws using the concepts of belief-desire psychology. I.e., there is no principled set of beliefs that we check for revision when deciding what to revise. Rather, we just check some set or other in various circumstances. Perhaps this is because belief revision is not a matter of the operation of a single processor, but is rather accomplished by various processors

and modules for various knowledge domains. While this means that we are sometimes inconsistent and perhaps much less than ideal rational beings, we normally manage to get by.

This solution is perhaps not what one would have hoped for, but it seems otherwise unproblematic unless you insist, as Fodor appears to, that most psychological explanations, particularly of the central systems, must be presented in terms of more or less law-like belief-desire statements. Once we acknowledge that the explanation of cognition may involve a lot of notions other than common sense belief-desire notion, and that the latter may not even play a significant role in "strict" psychological theory, then the frame problem is transformed from a problem of principle into a design problem.

Fodor's final argument against central system modularity concerns the lack of neural identifications of central system structures:

> Roughly, standing restrictions on information flow imply the option of hardwiring. If, in the extreme case, system B is required to take note of information from system A and is allowed to take note of information from nowhere else, you might as well build your brain with a permanent connection from A to B. It is, in short, reasonable to expect biases in the distribution of information to mental processes to show up as structural biases in neural architecture.
>
> ...in Quinean/isotropic systems, it may be *unstable, instantaneous* connectivity that counts. Instead of hardwiring, you get a connectivity that changes from moment to moment as dictated by the interaction between the program that is being executed and the structure of the task in hand. The moral would seem to be that computational isotropy comports naturally with neural isotropy in much the same way that

> informational encapsulation comports naturally with the elaboration of neural hardwiring.
>
> ...there are no content-specific central processes for the performance of which correspondingly specific neural structures have been identified. Everything we now know is compatible with the claim that central problem-solving is subserved by equipotential neural mechanisms. This is precisely what you would expect if you assume that the central cognitive processes are largely Quinean and isotropic.(117-119)

There are at least four problems with this line of argument. First, there isn't, in fact, a lot of well-supported evidence about the full neurological instantiation of **any** informational processes, except perhaps sensory "transducers." Thus, what we now know is compatible with almost any view of psychological processes--witness the recent excitement over connecitonism.

Moreover, part of the problem is that we really have very few settled views as to the make-up of cognitive architecture at this point, much less any detailed idea of the nature of specific processors, particularly in the central systems. Thus, when neuro-physiologists make tentative central system identity claims, they are usually very primitive by cognitivist standards--e.g. identifications with "memory" (with no specification of informational units or memory structures) or "intention." But this is precisely what we should expect given that most inter-scientific identities come about through independent development of the two characterizations to be identified. That is, it is extremely unlikely that neural scientists will not only produce mind

(cognition)-brain identity claims but also articulate the terms or descriptions for the cognitive side of the identity statement. So, again, it seems way, way too early in the game for arguments from known neurological structures to carry any weight.

Third, and perhaps most significantly, there might be reasons to "build" a brain with non-hardwired modules. E.g., suppose a given module is not neurally hardwired, but is rather "programmed" or "assembled" on each occasion it needs to operate. One reason to avoid hardwiring is in order to maximize diversity of operations. So instead of committing all your memory to specific functions, you leave a lot of RAM, even though what you load into RAM is often something that could have been effectively hardwired. Thus, it is not apparent that evolution would choose to hardwire modules.

Finally, it is not apparent why we should expect that our brains were "built" according to rationally optimal or near optimal design. If I were building an ideal humanoid, I wouldn't give it human retinas or knees, but it doesn't follow that we don't have those often less than optimally functioning structures. I thus see little hope for an argument against central system modularity based on current neuro-physiological evidence.

I conclude that once we recognize that common sense attitudes and attitude processes may not be explicit cognitive states and processors, we will see that the holism of belief-

fixation does not provide evidence against central system modularity.

I will close this section with a comment about the relationship of common sense attitudes and cognitive architecture. Despite common assumptions to the contrary, it is likely that modules will not themselves have or contain attitudes, particularly not beliefs and desires. To see this, simply apply my property specification of belief and note that it is unlikely that any modules themselves contain decision or reasoning units, though they may sometimes contribute to these processes. Thus, we cannot say that, e.g. the language module has beliefs about the syntax of language. The inclination to make such a claim appears to be the mistaken identification of belief with explicit storage. While it is quite likely that modules do contain (and store) explicit representations, and perhaps in some cases explicit representations corresponding in content to actual belief states, such states are not themselves beliefs. Rather, it is the relation of the modules and their representations to theorizing and decision processes which makes it true that the individual has certain beliefs. Thus, while modules may be largely responsible for the dispositions that are certain beliefs, it is a mistake to think of their explicit representations as themselves being beliefs.

C. The Commitments of a Computational Account of The Attitudes

I will now turn to the question of what commitments a computational view of the attitudes must make. Fodor has argued that a computational account should hypothesize a language of thought, and that this language must be innate. He does not, I think, claim that a computational account of the attitudes **implies** the innate LOT hypothesis, but only that it is very strongly supported by computationalism--the most viable, and perhaps only plausible hypothesis. In this section I will examine these claims, and argue that there are more reasonable alternatives that are suggested by a dispositional account of belief and the attitudes.

1. The Language of Thought

First, consider whether a language of though is required or strongly suggested by a computational account of belief. Fodor argues that the systematicity of our belief (i.e. attitude) capacities show that the representations responsible for beliefs must be language like. E.g., if someone is capable of believing that John loves Mary, then they are also capable of believing that Mary loves John and that John is loved by Mary, etc. On the other hand, anyone not capable of having any one of these beliefs is not capable of having any of them. This systematic aspect of belief-competence suggests that the representations underlying belief have a combinatory structure like that of language, e.g. that the representation underlying

the belief that cows eat grass is composed of the concepts COW, EAT and GRASS.[58]

Further, the fine-grainedness of our beliefs suggests that the representations underlying them are linguistic. For instance, the belief that Mark Twain wrote about life on the Mississippi and the belief that Samuel Clemens wrote about life on the Mississippi are different, one could have one and not the other (e.g. someone who did not know that Samuel Clemens was Mark Twain.) This complexity suggests the need for a highly intricate symbolic medium that is capable of handling such distinctions, and language-like mediums look to be the obvious candidates.[59]

Recall my definition of the belief that *p* as the disposition to use an explicit representation that means *p*, when formulated, in theoretical reasoning and decision making processes. If we also suppose that the inputs to reasoning and decision making processes are often occurrent thoughts that are explicitly formulated internal tokens in natural languages (i.e. the ones an individual speaks), then it is plausible to suppose that in us, beliefs are typically realized by the ability to explicitly formulate such internal tokens of natural languages. Thus, I could have either the belief that the Mark Twain wrote about life on the Mississippi or the

58. See Fodor (1987), pp. 147 ff.

59. I borrow this point from Georges Rey.

distinct belief that Samuel Clemens wrote about life on the Mississippi since I am able to have the occurrent thoughts 'Mark Twain wrote about life on the Mississippi' and 'Samuel Clemens wrote about life on the Mississippi.' Thus, we can admit languages of thought of sorts, i.e. internal tokenings of sentences of the languages we speak. But we have seen no reason, as yet, to think that such occurrent thoughts are not formulated in the languages we speak.

Fodor makes the following objection to this view:

> The obvious refutation of the claim that natural languages are the medium of thought is that there are nonverbal organisms that think. ...Considered action, concept learning, and perceptual integration--are familiar achievements of infrahuman organisms and preverbal children. ...But the representational systems of preverbal and infrahuman organisms surely cannot be natural languages. So either we abandon such preverbal and infrahuman psychology as we have so far pieced together, or we admit that some thinking, at least, isn't done in English.[60]

However, notice that the claim that natural languages are the medium of thought is ambiguous between two assertions, one that the property of thought is in part constituted by natural language representations, and the claim that natural language representations **sometimes or typically** are the medium of thought in natural language speakers. Thus, the former type of claim is that:

> thinking *p* is a matter of bearing some computational relation to a representation in a natural language that means *p*

60. Fodor (1975), p. 56.

whereas the latter is that:

> thinking *p* is a matter of bearing some computational relation to a representation that means *p*

together with the claim that:

> the relevant representations in humans are typically natural language tokens

The latter view says nothing about infants' and animals' thoughts except for the general claim that their thoughts must involve representations with appropriate content. And it is only the latter claim that need be advocated by a computationalist view of the attitudes. Thus, it may be that in some organisms one type of representation plays this role, while in other organisms another type of representation does the job. Specifically, it is plausible to claim that in verbal humans, it is typically the capacity to entertain occurrent thoughts in a language one speaks that fulfills this portion of the explication, while holding that in other organisms and pre-verbal children, the capacity to entertain other types of representations is what fulfills it. The latter cases are no counterexample since it is not being claimed that belief necessarily includes an ability to have natural language representations, but rather that this ability contingently (typically) fulfills that role in us.

We have seen that the reasons for thinking that the representations which underlie belief in us are language-like are the systematicity and fine-grainedness of our belief contents. However, it is not apparent that the beliefs of

animals and the very young exhibit similar characteristics. Could an infant be capable of believing that her mother loves him but not be capable of believing that she loves her mother? This seems to be at least possible--we don't know that it is implausible in the way we do for adult humans. Nor is it apparent that, e.g., a dog could believe that his master is home but not believe that the person who feeds him is home, assuming dogs can have both beliefs. In general, it is difficult to provide very definitive claims about what classes of attitudes infants and animals are capable of having. But it is not unreasonable to hypothesize that the representational abilities that underlie belief in preverbal children and animals involve other representational types, e.g. iconic or schematic representations, rather than linguistic representations, or at least representations in a linguistic code much cruder and more primitive than our natural languages, even though this implies much less systematicity and fine-grainedness than adult humans' beliefs exhibit.[61]

61. Fodor and Pylyshyn (1988), in arguing for the semantic compositionality of infrahuman thoughts, assert that "the organism that can perceive (hence learn) that *aRb* can generally perceive (/learn) that *bRa*." (p. 44) However, note that a) some systematicity does not a language make--it must be shown that much more detailed or sophisticated systematicity underlies animal though/perception in order to support the LOT hypothesis and b) the quoted claim itself is not beyond dispute--in general we know relatively little about animal perception or thought. As Block (1990), p. 277 points out, most animal learning studies were conducted by behaviorists who, as a matter of principle, were completely insensitive to claims about animals' representational capacities or thought contents.

Fodor also offers an objection to the claim that images could be the primary representational medium of thought. This arguments might be taken as lending some support to the more general claim that no organism could have non-linguistic representations without also having linguistic representations, which would undermine the present claim about infant and animal thinkers. His argument is essentially that resemblance, or looking-like, is too vague and indeterminate a notion to capture the content we want for mental representations. For instance, in Wittgenstein's example, a picture of a man walking up the stairs will look exactly like a man walking down the stairs. And, as Fodor notes, the problem is even more exaggerated. A picture of John could represent that John is tall or not fat, or not green, or a human, etc. He suggests that what must determine the content of an image is an associated description. E.g. an image of a triangle serves as a representation of triangles in general if that is how we describe it.[62] This is might be taken to imply that images, or icons, could not be the sole means of representation for an organism. A language of thought would be required in addition to determine the content of the images.

What I think these considerations show, however, is not that images require associated descriptions, but rather that **resemblance** cannot be what constitutes representation. As

62. See Fodor (1975), pp. 178 ff.

Cummins has argued, there are (both historically and currently) three principle views of the nature of representation, i.e. the similarity view, the co-variance view and the functional role view.[63] The type of difficulties just mentioned are, as Cummins states, traditional and apparently fatal problems for the view that a metal representation represents its referent in virtue of its similarity or resemblance to the referent. Specifically, it is hard to see how a token could represent abstract objects or properties on the similarity view.[64] For instance, no image of a triangle will resemble (the class of) all triangles, yet that is something we can represent. So the similarity view appears to fail. However, the two prominent alternatives each allow that there could be images, or more generally, non-linguistic representations, which represent without the presence of linguistic representations.

First consider the co-variance view, which in its crudest form, is that (semantic type) *p* represents (type) X in virtue of *p*'s being present exactly when *X*'s are present (e.g. suppose all and only X's cause *p*'s to be present in the system.)[65] Could a mental image represent, say, the fact that

63. See Cummins (1989). I count functional role and Cummins' interpretational view as the same for the purposes of the present point.

64. See Cummins (1989), pp. 33-34.

65. Fodor holds a counterfactually sophisticated version of this view--see Fodor (1987) chapter 4 and (forthcoming).

someone is walking up stairs without the presence of an associated description? There is no apparent reason why not--just let that image be present or "tokened" just when someone is walking upstairs, and according to the co-variance view, the image has the content that someone is walking up stairs.

Similar considerations apply to the functional role view. On that account, crudely speaking again, a given representation has its meaning in virtue of its function (or, e.g. "use") within cognition. So, a given image might constitute a generic triangle in virtue of being the image that is accessed when reference to a generic triangle is required. It is easy to imagine devices that would use images as representations, e.g. storing and reading them, but which used no language at all.

Thus, it seems that viable theories of representation allow for the possibility of representation in the absence of language, and thus for the possibility of non-linguistic attitudes. Note that the possibility of representation without language does not show that if we have, e.g. images, then their content must be specifiable in the complete absence of language. Language is apparently the dominant representational system in us. Thus, the easiest way to specify the content of an image, for verbal humans, is to use a description. And, perhaps, it is even true that all our non-linguistic representations get their content in verbal humans through association with linguistic representations. Still, the point

remains that current accounts of representation allow for representations in the absence of language, so the hypothesis that our linguistic thoughts occur in natural languages whereas non-linguistic organisms' thoughts occur in other mediums remains plausible.

What, then, is required in order to have attitudes? The answer to this will only come as we develop computational specifications of all of the attitudes--for now we can note the commitments required for belief. The dispositional account of belief requires that there be reasoning and decision-making processes that sometimes take explicit representations as inputs. Thus, thermostats do not have beliefs, since they have no such representations or processes, although belief-desire explanations apply instrumentally to them. On the other hand, it is plausible to suppose that a fair range of animals have both, taking "reasoning" liberally to include any sort of problem solving mechanism. Note that just where the presence of beliefs ends in the hierarchy of species may be indeterminate, since it may well be indeterminate as to when animals, or computers for that matter, cease to have representational reasoning or decision-making mechanisms. But I assume there is no problem here--most people would insist that most of the higher mammals have beliefs, but as for birds and bees and PCs, that is anybody's guess.

2. Nativism

Another area where Fodor has endorsed a strong commitment for a computational account of the attitudes is that of concept acquisition, where he argues that all concepts an organism can acquire must either be innate or reduce definitionally to innate concepts. Since it appears that few of our concepts so reduce, it follows that most of our concepts must be innate.

Fodor's argument for concept nativism involves the claim that to learn a language, one must learn a truth definition which uses a predicate co-extensive with the predicate to be learned. E.g. to learn the meaning of the predicate *P*, one has to learn that ''*P*x' is true iff x is *G*' is true for all substitution instances.[66] But this requires already understanding *G*. This is not to say that *G* has to be a simple predicate. The traditional empiricist model of concept learning suggests that such co-extensive predicates can be produced though associations of simpler predicates. But, as Fodor has often stressed, this traditional account seems wildly implausible. Few if any predicates semantically reduce to other predicates. For instance, most ordinary kind-concepts such as "cat" or "chair" appear to lack definitions.[67] It

66. Fodor (1975), p. 80.

67. See Fodor, Garrett, Walker, and Parkes (1980). Note that this claim might be contested. As the solution I shall offer shortly suggests (but does not imply), there may be extremely complicated definitions for many terms.

follows, Fodor argues, that most concepts must be simple, and thus innate.[68]

The most common reaction to this radical concept nativism is utter disbelief--most people find this view extremely implausible. However, it is not always immediately obvious what this reaction rests on. I will briefly consider several apparent reasons for recoiling from radical nativism. First, there are our experiences in concept attainment. For at least some concepts, particularly artifactual and scientific concepts, it appears that more than minimal exposure to instances is required for concept attainment. Rather, certain training, often rigorous training is required in order to master concepts--consider "quark" (or "electron") or "carburetor." In such cases, we seem to be explicitly taught everything that we need to know to "grasp" a concept--i.e. the extension of the concept, many or most of its (conceptual) entailments, its theoretic role, and the like.

A second reason for rejecting radical concept nativism is that we conceive of ourselves as conceptually creative beings. Thus, we invented trumpets and compact discs, and it is standardly assumed that this invention included the invention of the concept. And although we don't think we invented electrons or quarks, we think that we are creative theorizers

68. "Whatever is not *definable* must be innate," Fodor Garrett, Walker, and Parkes (1980), p. 313. Also see Fodor (1981), chapter 10, e.g. p. 292.

who devised these notions in order to produce better theories. Moreover, we like to think that we can go on inventing concepts, artifacts and theories more or less indefinitely, or at least we see no reason to think we are reaching the limits of our conceptual resources.

This leads to a third reason against radical concept nativism, namely that it is reasonable to think that our species as a whole has already attained more concepts than any individual could. Thus, natural science has long passed the point where any single individual could acquire detailed knowledge of all branches--or even of more than a half dozen or so. While this may be solely due to the lack of memory space for facts, it seems reasonable to think that if an individual could do nothing but acquire concepts the vast spread of concepts in natural and social science and art and literature would far exceed anyone's capacity. Nor is it reasonable to claim that we are all born with only subsets of all human concepts--there is no evidence of anyone ever experiencing a specific inability to acquire select concepts, although more general concept attainment problems (e.g. failure to attain highly abstract concepts) abound. And, again, we seem to be increasing the species-wide conceptual inventory.

Radical concept nativism seems implausible for the first and third reasons and highly unpalatable for the second reason. We are therefore in need of an alternative to Fodor's

proposal. The alternative that I will now sketch balks at the move from undefinable or primitive to innate. Why, after all, couldn't we have primitive, acquired concepts? Fodor's assertion to the contrary appears to rest on the claim that nothing but other concepts can "produce" a concept. E.g. if someone acquires *G*, then this is because *G* is built out of sub-concepts. This seems to assume that acquisition must occur at the level of explanation that such concepts participate in, i.e. common sense belief-desire psychology. Thus, consider part of Fodor's presentation by example of the standard (empiricist) hypothesis formation model of concept learning:

> So, for example, what goes on in your head in the experimental situation we've imagined might be something like this: You make your first guess--the green and triangular card is flurg [the concept to be acquired]--at random...Since, as it turns out, that guess was right, you have evidence for any of a range of hypotheses...You pick one and you try it...[69]

This does not appear to be much of an explanation. What we want to know is not how the subject comes to believe that the experimenter wants her to learn "flurg," but rather how the subject comes in contact with "flurg" in the first place. As Fodor correctly points out:

> What has happened is that the Empiricist story recruits what is really a theory of the fixation of belief to do double duty as a theory of the attainment of concepts. This strategy doesn't work, and the strain shows at all sorts of places. For example, it's surely clear that any normal adult would have acquired such workaday concepts as GREEN OR SQUARE long before he encountered a concept-

69. Fodor (1981), p. 268.

> learning experiment; hence, achieving criterion in such experiments *couldn't*, in the general case, require that any concept be acquired in the course of performing the experimental task. What the subject would have learned in the case described above, for example, is not GREEN OR SQUARE, but only the fact that the experimenter has decided to call thing that are green or square "flurg" for the duration of the run--a fact that is interesting only because it controls the distribution of the rewards.[70]

Fodor's solution, though, amounts to accepting the empiricists' explanatory framework, pointing out that it doesn't work in most cases, and then throwing up his hands and opting for innateness in all cases where definition fails. A more plausible alternative would seem to be to seek an acquisition account elsewhere.

Where else can we look? Our previous considerations suggest that there is much more to cognition than just beliefs and desires. Specifically, we have seen that it is likely that many attitudes are dispositions that are instantiated or realized by a variety of "subpersonal" (e.g. modular) representations and processes. Since concepts are standardly thought of as constituents of common sense attitudes, this suggests the possibility that concepts may supervene on a collection of lower-level representations and processes. And this will lead us away from radical nativism if what is primitive at the belief-desire level of explanation could be compositional at the computational-modular level of

70. Fodor (1981), p. 270.

explanation. Some of the representations and processes which constitute the attainment of a given concept could then be innate while others are acquired. It would follow that concept themselves are neither entirely innate nor acquired but a joint product of innate representations and structures and acquired representations and structures--partially innate and partially acquired, in effect.

Now, I cannot offer a genuine theory of how concepts can be reduced to other, less abstract cognitive states. This is something that we can only hope will happen over a long period as we begin to learn exactly what representations and states cognition contains. However, I think the rudimentary beginnings of such an account are already to be had from owing to various bits of the research on concepts over the past few decades. Consider for example the concept "cat."[71] Suppose that the psychological realization of this concept is a result of both a (perceptual) prototype used in recognizing cats--i.e. a set of features[72] used in detecting cats, together with a medium-sized-self-moving-thing-detection module, as well as a set of more abstract, categorical features in a lexical entry

71. I am controversially assuming that the meaning of 'cat' in most ideolects is somewhat vague and differs from the scientific meaning. For a view which takes into account Putnam-Kripke essentialism, see Rey (1983) and (1985). For criticism of the essentialist view, see Unger (1983).

72. For the next few paragraphs, in keeping with the psychological literature on concepts, I will refer to representations of features of concepts as "features."

module, e.g. a list such as +living thing, +object (rather than mass term), +animal, etc. Think of the role of the prototype and detection unit as that of picking out part of the extension--i.e. things which must be cats. That is, suppose that the features used for such detections are treated as sufficient but not necessary. The categorical features might be thought of as serving to qualify this partial extension, by limiting the possible set (e.g. no things not capable of life are cats.) Finally, suppose that indeterminate cases are matched to the kind whose observable feature set they are nearest to, if any. E.g. three-legged, tailless cats still have more observable features of prototypical cats than e.g. prototypical dogs do, so three-legged cats are included in the extension of "cats". This is, **very** roughly, the line of thinking of some prototype (or combined prototype-definitional) theorists in experimental work on concepts and categorization. Perhaps this is wrong in all the details, and perhaps even in the generalities but the point is to show that there is a reasonable line of investigation here to be pursued.

The suggested account does not yield any definition that decomposes "cat" into simpler concepts, since the recognitional features are not essential and the essential features radically underdetermine the extension. It does allow for a rather complex description of the extension of the term, and perhaps this is a definition of sorts, but it not obvious

that such complicated extension specifications are not to be had.

And we can see how this collection of representations and processes might be partially innate and partially acquired. Thus, suppose (as some vision theorists already do) that there is an innate module for detecting self-moving medium-sized living things and suppose that the features for cats are selected from some innate feature set in this module during the first few encounters with cats. As for the categorical features, we might expect many of them to be innate. However, some might be acquired through use of prototypes and related features. E.g. if "animal" is acquired, it is probably as a result of association with a set of recognitional procedures toether with the functional position of this feature within the categorical network. Finally, there is no reason to think that all reductions of concepts must be closely tied to perception. Some abstract concepts such as mathematical and logical notions might be partially acquired as a result of the expansion and development of certain specialized abilities, e.g. innate counting abilities or reasoning skills.

The meaning of 'cat' is primitive from the point of view of natural languages since there are no other natural language concepts that this concept reduces to. But this primitive concept might be realized in us via a composition out of various representations and processes at the sub-personal, sub-natural language level of explanation. The moral is that

most of our concepts might be primitive yet non-innate, since what is primitive at one level of explanation is compositional at a lower level, and acquisition can be explained at the lower level.

What I suggest, then, is that while Fodor is correct in declaring the bankruptcy of traditional empiricist accounts of concept acquisition, the solution is not radical nativism. Instead, we need to pursue alternative accounts which reduce concepts to less abstract cognitive states and processes and then explain how some of these states and processes can be acquired.

It should be noted that a cognitive architecture will nonetheless be up to its ears in innateness. First, it is plausible to think that virtually all faculty and modular divisions are innate, and it is also plausible to think that there must be innate very low-level, "machine" languages for carrying out the appropriate computations. Further, there is a strong case for thinking that many of the sub-personal states that underlie concepts are innate. The case for such nativism is simply Chomsky's good-old poverty of the stimulus argument: if you have a mental state that is widely present, but you cannot find any shared experience or (very) common training that would account for its acquisition, it is most likely innate. Since we uniformly attain a wide variety of common sense concepts with little or no training, including common sense observational concepts, common sense kinds, and common

sense psychological concepts, it is reasonable to postulate that most of these concepts are partially or even wholly innate. But at the same time, we can postulate that most artifactual and scientific concepts (and perhaps also sociological concepts) are partially or wholly acquired.

IV. Conclusion

I have provided a dispositional, computational account of belief and of the attitudes in general and I have used this view to try to develop a new picture of how we should view the attitudes in relation to scientific psychology. In closing, I will draw out a moral that has been implicit in the past few sections. I want to suggest that the view I have developed may have serious implications for the use of common sense psychology in philosophical investigations, particularly in examinations of the nature and plausibility of scientific psychology. Much philosophical inquiry into the foundations of cognitive theory does not concern actual theories formulated by psychologists, but instead relies on common sense belief-desire psychology. Sometimes this is because no suitable, well-developed, well-confirmed theories are as yet available, and sometimes because common sense is simply more accessible and convenient. This substitution is unproblematic if we can be assured that common sense states are representative of cognitive states in general. However, the contrast I have outlined between dispositional common sense states and

explicit cognitive states suggests that common sense states may indeed have features that are uncharacteristic of many other cognitive states. For instance, it may be that while common sense belief-desire explanations and concepts are holistic, indeterminate or context-relative, due to the abstract and dispositional nature of concepts such as belief, the concepts and explanations of scientific cognitive theory will not generally exhibit these characteristics. My investigations regarding the indeterminacy of belief-onset, at the end of the first section, lend some support to this conjecture.

If this is correct, then this raises obvious problems for critics of cognitive psychology who have focused on common sense explanations. For instance, a plausible line of response to Davidson's claim that psychological notions are "heteronomic"--i.e. they cannot be sharpened into strict, scientific laws[73]--is to allow that while common sense notions such as belief and desire are indeed heteronomic, alternative representational, computational notions such as "storage", "activation" and "computation" are "homonomic"--i.e. they can be sharpened into strict laws. Thus, apparent failings of common sense belief-desire explanations and ascriptions are not necessarily failures of cognitive theory in general. In fact, it seems coherent to acknowledge all sorts of problems

73. See Davidson (1980), p. 219.

with common sense belief-desire concepts and explanations and yet be an avid computationalist who sees a bright future for scientific theories that explain behavior in terms of computations over representations.

Of course, any apparent negative features of common sense psychological concepts and explanations may also be features of alternative cognitive concepts and explanations. My point, though, is that the latter must be evaluated in their own right. There is a tendency in philosophy to think of cognitive psychology as the science of belief. The view that I have presented suggests that this is misleading since the states that common sense belief-desire psychology is concerned with may be much more abstract and less explicit than the states that theories of computation and representation are primarily concerned with. So, it would be prudent for us to turn from the armchair to the textbook and the laboratory when pursuing questions about the nature of psychology.

Appendix: Stich and Rey on Dividing Belief

In the first section I defended the view that the property of believing that *p* is being disposed to use an explicit representation that means *p*, when formulated, in theoretical reasoning and decision making processes. In this appendix I will consider two alternative views of the property of belief, views that are both inspired by Nisbett and Wilson's experimental work involving inappropriate attributions and explanations of one's own beliefs and desires.[74]

The following sort of finding motivates the views in question: In an experimental situation, subjects were presented with identical items, e.g. stockings, and asked to select the best-quality item and explain why they had chosen it. There was an overwhelming preference for the right-most item, though virtually no subjects cited the position of the item as any part of the basis for their selection. Moreover, when questioned about a possible positional effect, the subjects' strongly and sincerely denied that the position of the items had influenced their decisions in any way. This evidence produces the following problem for belief(-desire) explanations. As far as the subjects' behavior is concerned, they are clearly operating with a belief such as "the right-most item is better" (or, e.g. a desire to select the right-

74. Nisbett and Wilson (1977).

most item.) Yet, their sincere verbal reports reflect the absence of this belief (or desire.) Thus, it seems that people in such situations both possess and do not posses certain beliefs (or other attitudes.)

Wilson has claimed that such studies, along with various other data, suggest that there may be two very separate systems underlying these phenomena, one system which controls behavior, and a separate, isolated system which attempts explanations of the individual's behavior not through accessing actual cognitive states, but through the application of prior generalizations to salient environmental influences.[75]

This possibility is most strikingly illustrated by the following result from research on a split brain patient, whose behavior on separate sides of his body was the result of the independent operations of his severed hemispheres. The patient presented unified "rationalizations" of his informationally disjoint actions. For instance, in a task where he was shown a chicken claw with only his right eye (and thus only his left hemisphere receives this information) and snow with his left eye (and right hemisphere) pointed to a picture of chicken with his right hand (LH control) and a shovel with his left hand (RH control.) When asked to explain his action (the speech center is in the LH), he said "I saw a claw and I picked the chicken and you have to clean out the chicken shed

75. See Stich (1983), p. 236 for presentation of Wilson's views.

with a shovel."[76] The experimenters claim that this result was uniform and straightforward:

> In trial after trial, we saw this kind of response. The left hemisphere could easily and accurately identify why it had picked the answer, and then subsequently, and without batting an eye, it would incorporate the right hemisphere's response into the framework. While we knew exactly why the right hemisphere had made its choice, the left hemisphere could merely guess. Yet, the left did not offer its suggestion in a guessing vein but rather a statement of fact as to why that card had been picked.
>
> These varied observations on [the subject] offer us the opportunity to consider whether we were not observing a basic mental mechanism common to us all. We feel that the conscious verbal self is not always privy to the origin of our actions, and when it observes the person behaving for unknown reasons, it attributes cause to the actions as if it knows but in fact it does not. It is as if the verbal self looks out and sees what the person is doing, and from that knowledge interprets a reality.[77]

While it is possible that the surgery (quite minor, as brain surgery goes) impaired a single, unified system, thereby distorting its functioning, it is more likely that many of our own explanations or our own behavior are similar in that they are not based on the introspective observation of actual psychological states, but are rather guesses by an informationally isolated explanatory system.

What is of interest to us here is the reaction to such a view. Suppose, for the rest of this appendix at least, that there actually are two separate cognitive sub-systems, one of which constitutes explicitly stored information for the

76. Gazzaniga and LeDoux (1978), p. 148.

77. Gazzaniga and LeDoux (1978), pp. 148-150.

purpose of actions, and another which produces explanations of the actions, but which has no access to the former's data base. Stich argues that such a result would show that there are no such things as beliefs:

> It is a fundamental tenet of folk psychology that *the very same* state which underlies the sincere assertion of 'p' also may lead to a variety of nonverbal behaviors...In those cases in which our verbal subsystem leads us to say 'p' and our nonverbal subsystem leads us to behave as though we believed some incompatible proposition, there will simply be no saying which we believe. Even in the (presumably more common) case where the two subsystems agree, there is no saying which state is the belief that p. If we really do have separate verbal and nonverbal cognitive storage systems, the functional economy of the mind postulated by folk theory is quite radically mistaken. And under those circumstances I am strongly inclined to think that the right thing to say is that *there are no such things as beliefs*.[78]

I agree with Stich that there is no saying whether or not there is belief in the cited cases, but the judgement that this shows that there are really no such things as beliefs is inappropriate.[79] Consider Stich's claims about what folk

78. Stich (1983), p. 231. Stich's main point in the section the quote is taken from is not that there are no such thing as beliefs, but rather that it is up to scientific psychology to determine whether or not there are any beliefs, and here I agree completely. However, Stich appears to overlook the fact that science often tells us that a certain (apparent) property is much different than we took it to be with our common sense explanations.

79. It is notable that Nisbett and Wilson do not themselves take an eliminativist position toward the attitudes, although they do suggest that their evidence shows that belief-desire explanations constitute an "a priori, causal theory" Nisbett and Wilson (1977), pp. 248 ff. And it is presumably one that is wrong in many cases. However, the "correct" explanations that Nisbett and Wilson offer of subjects' behavior are often couched in terms of attitude

psychology is committed to. When he says that folk psychology is committed to the belief that *p* being the state which *underlies* the sincere assertion that *p*, he may mean one of two things, namely either that the belief that *p* is the normal cause of the sincere assertion that *p*, or the much stronger claim that the belief that *p* is always, necessarily the cause of the sincere assertion that *p*. The difference between them is that in the former case, the explanation of typical sincere assertions as being caused by belief is not undermined by a few exceptions, e.g. a few cases where we cannot say if there is belief present or not and must switch to alternative concepts and explanations. However, the latter, stronger view would seem to be undermined by even one exception. Now, Stich does not distinguish these alternatives, but it seems charitable not to saddle him with the latter view since we do seem to allow for special exceptions in the belief-expression connection. For instance, suppose that I intend to sincerely express my beliefs but fail to express myself correctly, substituting one word for another. E.g. Suppose I say "Boston is south of New York" when I know full well it is to the north. Have I changed my belief here? The obvious explanation is simply that in this case the normal causal connection between belief and sincere expression has gone a bit astray. Thus, it would seem quite dubious to claim that we always

concepts.

(e.g. necessarily) believe what we sincerely express. But as I have pointed out, this weaker view does not show that because in the cases where expression and action conflict there is no such thing as belief, it merely shows that in some cases we cannot apply the concept and thus much switch to an alternative explanation, set of concepts, etc. The fact that a concept is not determinate in a range of cases does not show that it never applies or cannot be used to explain anything.[80]

What leads Stich to eliminativist conclusions in the quoted passage, then? A close reading suggests that he has smuggled an explicit storage view of belief into the argument. Note that on my dispositional view, it is fine to say that there are two separate sorts of explicit storage underling dispositions to use explicit representations in reasoning or decision-making, or **any number** of separate storage units for that matter. The state that "really is belief" is not an explicit storage state but the dispositions--to cognitive systems' use of representations when formulated--that such storage states produce.[81] Nor is there any basis for claiming

80. Although I am inclined to think that the indeterminate cases support the main sentiment of Stich's view, namely that belief-desire concepts--the concepts of belief and desire in particular--are not (completely) suitable to scientific psychology.

81. As evidence of Stich's assumption, consider his summary of the mental sentence view of belief, "to have a belief is to have a sentence token *inscribed in the brain* in such a way that it exhibits the causal interaction appropriate to beliefs," Stich (1983), p. 74, my emphasis.

that "folk theory" is committed to an explicit storage model of belief.[82] I would suggest, rather, that this is a common false picture of the property of belief, much like the mind's eye (or camera in the head) model of mental imagery. Thus, eliminativism with regard to the property of belief seems an unwarranted reaction to the "dual systems" hypothesis.

Rey has offered a much stronger response to Stich's treatment of the dual systems view. He argues that we can preserve attitude explanation in the problem cases by dividing belief--i.e. replacing the notion of belief with two others, the notion of avowed beliefs (and desires) where, roughly, "a person avowedly believes that *p* if she would sincerely and decidedly assert *p* if asked"[83] and the notion of central beliefs, which are supposed to do approximately what "ordinary" beliefs were supposed to do. He sees this as a way of making the dual systems hypothesis consistent with attitude explanation:

> Taking seriously the "two sets of books" Stich fears we keep, we could regard a person as a computer having two sets of addresses: the "central" set, a set of address that contains the contents of attitudes that enter through [certain unspecified computational relations] into instances of practical reasoning that largely determine one's acts in the manner Fodor described; and the "avowal" set, a set of special addresses that specifically provides the contents, accessed by [certain other unspecified computational relations], that serve as the basis

82. See Double (1985), who reports that in a survey, his introductory (philosophy) students failed to share Stich's eliminativist intuitions in cases of multiple subsystems.

83. Rey (1988), p. 278.

> for sincere assertions and other functions in which one is to be taken at one's word (oaths, promises, examinations.)[84]

Rey also maintains that this splitting of belief will allow for successful explanations of the phenomena of weakness of the will and of self deception. For instance, in a case of self-deception, a person may avowed prefer that not *p* (e.g. they stop smoking), yet they may centrally prefer that *p* (e.g. they continue smoking) where the latter state continues to drive their behavior despite their contrary avowal, and despite the avowal that not *p* (not smoking) is to be preferred over *p* (smoking.)[85]

While it is an interesting attempt, I do not think Rey's account succeeds in incorporating the dual systems hypothesis in belief-desire explanation. I shall discuss three main problems that his view faces.

First, and foremost, as we have considered at length in the first section, belief cannot be identified with storage. Thus, belief cannot be split into two separate storage units, so Rey's account is not acceptable as far as his rough sketch of the appropriate computational relations is concerned. However, it is not apparent that suitable alternative relations can be found that really do divide belief, given my dispositional account of that attitude. The obvious route is

84. Rey (1988), p. 278

85. See Rey (1988), pp. 281-2. The example is mine.

to identify the avowal that *p* with the disposition to sincerely express *p*, and the central belief to use and explicit representation that *p*, if formulated, in decision making processes. But this will not capture all of the explanatory force of the dual systems account, for the point there is precisely that both systems are potentially capable of driving behavior. For instance, suppose I am a subject in an experiment who has just exhibited a preference for the right-most item and am talking with the experimenter, avowing that I have no such preference. I might reason "he says I've done something that I clearly haven't, what's wrong with him? Maybe I should try to leave..." and take action. So we need something like a dual disposition to express and to act, and another disposition to act (for central beliefs.) But dispositions cannot be split in this way--either I do or do not have a disposition to use a given representation, if formulated, in reasoning processes. This cannot be "split" into two contrary dispositions. Instead, if we are going to talk about multiple influences, it seems that we must talk of multiple influencing factors, e.g. two separate cognitive subsystems. Thus, once we reject a storage model of belief it is no longer apparent that the computational property can be split into two other properties which allow us to use the dual systems explanations within common sense attitude psychology.

A second problem for Rey's view is that while belief and desire are central notions in attitude explanations and

ascriptions, which are themselves quite complex, he does not provide much help in explaining how the two new attitudes are to fit into these explanations and ascriptions. For instance, consider the fear that *p* together with the desire that not *p*. Suppose someone fears that *p*--will they avowedly desire that not *p* or centrally desire that not *p*? Similarly, knowledge is typically thought to require belief, but will it require avowal, central belief, both, or neither? Further, what are we to say when someone sincerely utters something--how are we to tell if they centrally believe it, avow it, or both? How are we to decide when an avowal explanation of action is appropriate and when a central explanation applies? I.e. when will the practical syllogism apply with avowal, and when will it apply with central belief, and when with avowed or central desires? Unless we have answers to these and many other such questions, it is not apparent that the avowal-central distinction can be integrated into the standard common sense attitude explanatory framework without inhibiting its usefulness in explanations and ascriptions.

In fact, it is probably true that part of the reason that belief and desire are such successful notions is that they cut across a lot of different cognitive systems allowing us to provide very general explanations for a system which appears to have a lot of isolated, different units and functions (see my discussion of modularity in the central systems in section III.B.) Thus, it seems reasonable to remain skeptical about

the possibility of dividing belief until it is actually accomplished.

The third problem is that it is fairly clear that there are more distinctions to be drawn, more than two sources of input for decision processes. But this will only compound the previous problems. For instance, it is apparent that the avowal-central distinction is not enough to account for the wide range of "split attitude" cases. As Rey notes:

> Some reasoning processes may involve only [avowal]-like operations, as when one passes through a piece of reasoning "merely intellectually," and so comes to avow thing that one doesn't centrally believe.[86]

Now consider two cases of smokers who are contemplating quitting. Both assert that they want to quit. But one only "avows" this superficially, although deep down she really likes smoking, likes the image of herself as a smoker, etc. The other person really wants to quit through and through--he has studied the effects of smoking, has had friends die from lung cancer, etc. But, unfortunately, (and this is a well documented phenomenon), he is still unable to overcome the addiction. He requires outside assistance, e.g. a behavior modification program. Aware of his failure, he "avows" "I guess I really don't want to quit." If we identify the former case and this expressed desire as avowals, as Rey's quote would suggest, then the weakness of the will in the latter

86. Rey (1988), p. 279.

case would seem to require a further distinction, and it is not apparent what that would be, in terms of attitudes.

Or consider the following three-tiered case of *akrasia* (more or less.) I offer to try the hosts' cake. However, I am firmly of the belief that sugar is not good for me, and so on deeper consideration resolve to not take any when it is served. Yet, it is chocolate, which I love, so I give into my craving and take a piece. But I am in fact too full already and cannot actually bring myself to eat it, although it looks and smells delicious. We can isolate four separate influences on decision and action here, all of which seem to require separate treatment for the same reasons that lie behind the avowal-central distinction. But, once again, it seems easiest to abandon the attitude framework and instead merely speak of different subsystems and processes, each operating on an explicit set of representations--e.g. a "reasoner" (perhaps with a "Gricean" data-base for reasoning about social situations), a stored set of previously determined "goals," processes that attempt to satisfy "tastes," and a unit which computes and attempts to satisfy "bodily needs." As the scare quotes suggest, such units may correspond approximately to the notions of common sense. Yet, we should not confuse this with common sense explanation itself. The optimal procedure seems to be to move from explanations which cite the dispositional attitudes, including belief, to explanations which postulate

processes and modules which involve explicit storage and computations.

Thus, I suggest that Rey's attempt at preserving such attitude explanations in cases where inconsistency in the determination of belief and desire seem to occur does not succeed, and we appear to be better off pursuing non-belief-desire explanations in these cases.[87]

87. Note that my main objection is really the first one, over the fact that belief or belief-substitutes should not be identified with explicit storage. If the avowal/central belief account succeeds in explaining cases of *akrasia* and self-deception within the attitude framework, then that is fine with me. But that in and of itself has no bearing on the nature of belief states.

BIBLIOGRAPHY

Audi, R. (1982) "Believing and Affirming." *Mind*, 91, pp. 115-120.

Baker, L. R. (1987) *Saving Belief.* Princeton University Press.

Baars, B. (1988) *A Cognitive Theory of Consciousness*, Cambridge University Press.

Block, N. (1990) "The Computer Model of the Mind." In D. Osherson and E. Smith (eds.), *Thinking, An Invitation to Cognitive Science, Vol. 3*, Cambridge MA: MIT Press.

Burge, T. (1979) "Individualism and the Mental." In P. French, T. Uehling and H. Wettstein (eds.), *Contemporary Perspectives in the Philosophy of Language*, Minneapolis: University of Minnesota Press, pp. 73-121.

Burge, T. (1986) "Individualism and Psychology." *The Philosophical Review, XCV*, No.1, pp. 3-45.

Cam, P. (1988) "Modularity, Rationality and Higher Cognition " *Philosophical Studies 53*, pp. 279-294.

Chomsky, N. (1965) *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1980) *Rules and Representations.* New York: Columbia University Press.

Chomsky, N. (1986) *Knowledge of Language.* New York: Praeger.

Churchland, P. (1981) "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy*, LXXVIII no. 2, pp. 67-90.

Clark, A. (1978) "From Folk Psychology to Naive Psychology." *Cognitive Science*, 11, pp. 139-54.

Cummins, R. (1983) *The Nature of Psychological Explanation.* Cambridge, MA: MIT Press/Bradford.

Cummins, R. (1989) *Meaning and Mental Representation.* Cambridge MA: MIT Press/Bradford.

Davidson, D. (1980) *Essays on Action and Events.* Oxford University Press.

Dennett, D. (1978) *Brainstorms*. Cambridge, MA: MIT Press/Bradford.

Dennett, D. (1987) *The Intentional Stance*. Cambridge, MA: MIT Press/Bradford.

Double, R. (1985) "The Case Against the Case Against Belief." *Mind*, 94?, pp. 420-430.

Field, H. (1978) "Mental Representation." *Erkenntnis 13*, 9-61. Reprinted in N. Block (ed.), *Readings in the Philosophy of Psychology, Vol. 2*, Cambridge MA: Harvard University Press, (1981), pp. 78-114. [Page numbers refer to Block (ed.)]

Fodor, J. A. (1975) *The Language of Thought*. New York: Crowell.

Fodor, J. A. (1981) *Representations*. Cambridge, MA: MIT Press/Bradford.

Fodor, J. A. (1983) *Modularity of Mind*. Cambridge, MA: MIT Press/Bradford.

Fodor, J. A. (1987) *Psychosemantics*. Cambridge, MA: MIT Press/Bradford.

Fodor, J. A. (forthcoming) *A Theory of Content* Cambridge, MA: MIT Press/Bradford.

Fodor, J. A., M. Garrett, E. Walker, and C. Parkes (1980) "Against Definitions." *Cognition, 8,* pp. 1-105.

Fodor J. A. and Z. Pylyshyn (1988) "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition, 28*, pp. 3-71.

Gazzaniga, M. and J. LeDoux (1978) *The Integrated Mind*. New York: Plenum Press.

Gelman, R. and C. R. Gallistel (1978/1986) *The Child's Understanding of Number* (2nd ed.) Cambridge MA: Harvard University Press.

Gelman, R., E. Meck, and S. Merkin C. (1986) "Conceptual Competence in The Numerical Domain." *Cognitive Development, 1*, pp. 1-29.

Greeno, J., S. Riley and R. Gelman (1984) "Conceptual Competence and Children's Counting." *Cognitive Psychology, 16*, pp. 94-143.

Jackendoff, R. (1983) *Semantics and Cognition*. Cambridge MA: MIT Press.

Kohlberg, L. (1969) "Stage and Sequence: The Cognitive-Developmental Approach to Socialization." In D. A. Goslin (ed.), *Handbook of Socialization Theory and Research*, Chicago: Rand McNally.

Kohlberg, L. (1981) *The Philosophy of Moral Development*. San Francisco: Harper and Row.

Lehrdahl, F. and R. Jackendoff (1983) *A Generative Theory of Tonal Music*.Cambridge MA: MIT Press.

Lycan, W. (1988) *Judgement and Justification*. Cambridge University Press.

Nisbett, R. and Wilson, T. (1977) "Telling More than We Can Know: Verbal Reports on Mental Processes", *Psychological Review, 84*, pp. 231-259.

Putnam, H. (1975) "The Meaning of 'Meaning'." In K. Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in tne Philosophy of Science, VII, Minneapolis: University of Minnesota Press. Reprinted in Putnam, *Mind, Language and Reality, Philosophical Papers Vol. 2*, Cambridge University Press, (1975) pp. 215-271.

Putnam, H. (1983) "Computational Psychology and Interpretation Theory." In Putnam, *Realism and Reason, Philosophical Papers Vol. 3*, Cambridge University Press, pp. 139-154.

Putnam, H. (1988) *Representation and Reality*. Cambridge: MIT Press.

Pylyshyn, Z. (ed.) (1987) *The Robot's Dilemma*. Ablex: Norwood, New Jersey.

Rawls, J. (1971) *A Theory of Justice*. Cambridge MA: Belknap/Harvard University Press.

Rey, G. (1983) "Concepts and Stereotypes." *Cognition, 15*, pp. 237-262.

Rey, G. (1985) "Concepts and Conceptions: A Reply to Smith, Medin and Rips." *Cognition, 19*, pp. 297-303.

Rey, G. (1988) "Toward a Computational Account of *Akrasia* and Self-Deception." In B. McLaughlin and A. Rorty (eds.), *Perspectives on Self-Deception*, University of California Press.

Stich, S. (1983) *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press/Bradford.

Unger, P. (1983) "The Causal Theory of Reference." *Philosophical Studies, 43,* pp. 1-45.

SEMANTICS NATURALIZED

In this essay I develop and defend the view that a theory of meaning for natural languages should be naturalized to scientific psychological inquiry.[1] This is a methodological claim, namely that the study of the relationship between languages and the world should be conducted as a scientific investigation. Meaning relations appear to be determined by the psychological states, whatever they might be, that underlie our comprehension, production and use of language. I maintain that we must study these psychological states in order to determine the nature of the semantic relations.

1. For other advocations of naturalism in semantics, see Putnam (1970) and Devitt and Sterelny (1987). A notable difference from the present view is that neither of these works advocates the specific naturalization that I defend here, namely the naturalization to cognitive psychology. And neither work presents the type of positive case I attempt to develop here, but merely recommends the methodology. Devitt and Sterelny's book is useful in that, as a survey text, it provides a naturalist evaluation of a number of different philosophical approaches to semantics. Note that I do not follow them in endorsing a causal approach to meaning and reference.

Also note that Chomsky has advanced the view that the study of language is best conceived as the scientific study of the language faculty--for instance, see Chomsky (1980) and (1986). The present work is, for the most part, a defense of his scientific approach for the study of meaning. However, I wish to leave certain questions open as far as the general thesis of semantic naturalism is concerned, i.e. if there a specific language faculty or if our knowledge of language distributed throughout cognition, if what we attribute as knowledge of language (always) involves explicit representation of the attributed content, and whether or not languages are best conceived as abstractions out of cognitive states.

In other words, we should think of "meaning" as an apparent natural kind, to be dealt with in the way that all other natural kinds are, i.e. by developing scientific theories which reveal the underlying nature of the kind, often by revising or rejecting much of the pre-scientific lore about the natural kind in question. This is not to say that all philosophical accounts of meaning to date are false or misguided. However, the view I am suggesting does imply that past and present philosophical accounts should be viewed as empirical theories, whose validity is determined via the usual means of scientific confirmation, i.e. explanatory success--specifically, more success than competing theories--goodness of fit with the (apparent) data and coherence with related theories.

The naturalist is, of course, unable to offer any *a priori* or conceptual defense of his position. Ultimately, naturalism is proven correct when successful scientific theories are developed which succeed in explaining the phenomenon at issue. This seems to leave the naturalist in the awkward position of being unable to defend the approach until it has finally proven successful. But such success may be a long time in coming--certainly, cognitive psychology is only

in its infancy.[2] However, there is a means of defending naturalism prior to the emergence of successful theories. First, naturalism can be shown to be *prima facie* plausible, in advance of actual, successful theories. That is, it can be argued that a naturalistic conception of semantics seems reasonable--that one could reasonably hope to produce a substantial theory of meaning using naturalistic methods. Second, the epistemic basis for alternative non-naturalistic methodologies can be called into question. Specifically, it seems that if a non-naturalistic semantic method is legitimate, then it must be true that we have non-empirical knowledge of meaning, or at least knowledge of meaning that is prior to and independent from the naturalizing scientific methodology, i.e. empirical psychology. But it can be argued that we do not appear to possess such knowledge, thus leaving the naturalistic account as the only plausible semantic methodology, even though the final proof of the position will not come until we have actually developed successful empirical theories.

In what follows, I will develop a defense of naturalism along the lines that I have just sketched. In the first section, I will clarify the notions of a theory of meaning and

2. Note that the present essay is not an attempt at demonstrating that current psychological theories form the basis for a successful naturalization of semantics--I do not think that they do. Nor is it any sort of survey of state-of-the-art cognitive semantic theories.

of naturalism--i.e. I will provide a clearer specification of what it is that is getting naturalized, and what the implications of naturalism are. In the process I will answer several objections to the general idea of naturalized semantics. I will then set out and a *prima facie* conception of semantics naturalized to cognitive psychology in the second section, and defend this view against several further objections. With this conception in place, I will turn to the main defense of naturalism--in the third section I will argue that our pre-scientific explicit knowledge of meaning fails to yield sufficient ingredients for a non-naturalistic theory of meaning. If this is correct, then the naturalistic approach is left as the only viable candidate for semantic inquiry.

I. What is a Theory of Meaning?

I am recommending that semantics be naturalized to scientific psychological inquiry, but current philosophy provides us with a variety of different notions of what meaning is--e.g. sense, reference, truth conditions, functions ranging over possible worlds, logical forms, verification conditions--as well as a number of methodological and explanatory frameworks--e.g. interpretation theory, specification of direct reference, model theory, definitional analysis, causal chains of reference, theories of use, conceptual roles. It is not obvious that all of these conceptions and all of these methodologies add up to a single,

uniform point of view (that could be titled "semantics.") What is it, then, that a naturalized semantics is trying to explain, and which of these methodologies should it adopt?

I am not going to attempt to decide between these accounts, nor am I going to attempt to assemble all (or some) of them into a single outlook. Instead, I am going to work with a more generalized notion of a theory of meaning that, I claim, most of these approaches fall under. Specifically, I will understand "semantics" to be **the descriptive project of specifying how the elements of our languages are related to the world**. Such an account might specify the relations that attain (e.g. references), or it might (also) specify states that are responsible for determining these relations (e.g. senses.) Most, and perhaps all, standard philosophical accounts that have been labeled "semantics" fall under this conception, not for any deep methodological reasons, but simply because then purport to specify one or another actual--as opposed to merely possible or recommended--"semantic feature"[3] of language--as opposed to specifying structural, syntactic features of language.

What I am asserting, then, is that any methodology (and view of meaning) falling under this general conception of

3. I use the phrase 'semantic features' as a placeholder for whatever sorts of states or relationships a theory of meaning ends up characterizing--e.g. sense, reference, truth conditions, etc.

semantics should be naturalized to scientific psychological inquiry. Such naturalization, as I understand it, has two consequences. First, empirical psychological facts about the processes and representations that speakers use in comprehending and producing utterances are an essential part of the evidential base for a theory of meaning. That is, any fact about language's role in our psychology could potentially be relevant to the truth or falsity of a semantic theory. E.g., a theory which successfully answers the question "what does 'dog' mean in English?" must be responsible to the psychological facts about how speakers comprehend utterances involving 'dog.' For instance, this data might help decide among competing accounts of the meaning of that term.[4] Such facts need not be obvious--in fact, I suspect that we have as yet little idea of what such evidence consists in. The facts I especially have in mind here are not facts about psychological states that common sense attributes such as beliefs and desires, but rather facts about our psychological language processing mechanisms that must be garnered through the experimental study of cognition and the brain.

The second implication of the naturalization of semantics to psychology is that the confirmation of a theory of meaning is part of the confirmation of empirical psychological theories in general, so that such a theory must plausibly

4. I shall provide several sample cases below.

cohere with other psychological theories. For instance, if a theory of meaning has implausible implications for a theory of language acquisition, and certainly if it implies or supports claims contrary to known facts about language acquisition, then this is good reason for rejecting or modifying the semantic theory. Thus, I am arguing that it is a mistake to conceive of either the data base or the confirmation of a theory of meaning as prior to or independent from empirical psychological theories of language processing.

I take it that both of these claims, if correct, will serve to modify existing philosophical approaches to semantics. The effect of the first claim is fairly concrete and obvious, i.e. it will widen the potential data base for semantics since most philosophical accounts assume a much narrower data base, e.g. only speaker's intuitions, or only facts about speaker's beliefs and desires. Acceptance of the claim about confirmation, though, would have less tangible effects. Roughly, the result would be that semantics could not be conducted in isolation from empirical theories of language processing, nor could semantic theories be used to legislate the explanatory goals of the latter. Put another way, if naturalization is appropriate, then it is possible that theories of language processing could show us that our languages do not have the semantic features that our intuitions and introspections suggest that they have.

Perhaps it might appear that there is an easy way out for those who wish to avoid these consequences. Someone might simply select their favorite methodology and conception of semantic features, and label the resulting investigation "semantics." Semantics, would, then, by definition not involve either the wider data base or co-confirmation with psychological theories. However, such a move trivializes the outcome of the investigation. The idea of the general conception of semantics that I formulated above is to prevent this sort of maneuver. To see this, let's suppose that the investigation in question was supposed to describe the references of terms. But, on the restricted view, the theorist with a complete theory could not claim to have a characterization of actual reference, but only reference-as-described-by-the-methodology. To claim that the methodology had yielded the real reference of terms, the theorist would need to maintain that there were no facts or issues other that those considered by the methodology which were relevant to the determination of the reference of terms. But this is to acknowledge that the methodology in question might potentially face both competing methodologies and additional falsifying data. Hence, any semanticist wishing to claim methodology-independent validity for the semantic features her investigation postulates must at least admit the possibility of naturalism.

A somewhat deeper objection to the possibility of naturalism is that the suggested naturalization of a theory of meaning to empirical psychology is a category mistake. This appears to be the upshot of the following passage from Dummett:

> A theory of meaning...is not intended as a psychological hypothesis. Its function is solely to present an analysis of the complex skill which constitutes mastery of a language, to display, in terms of what he may be said to know, just what it is that someone who possesses that mastery is able to do; it is not concerned to describe any inner psychological mechanisms which may account for his having those abilities. If a Martian could learn to speak a human language, or a robot be devised to behave in just the ways that are essential to a language-speaker, an implicit knowledge of the correct theory of meaning for the language could be attributed to the Martian or the robot with as much right as to a human speaker, even though their internal mechanisms were entirely different.[5]

Dummett might simply be making the seemingly trivial point that a theory of meaning will be more abstract than theories that characterize specific causal sequences of psychological states. However, there may be a stronger claim here as well. Specifically, we might ask in the present context, do these considerations provide any reason for thinking that a theory of meaning can be developed and confirmed without examining facts about psychological mechanisms?

The first point to note here is that while consideration of Martians or robots might immediately suggest radical

5. Dummett (1976), p. 70. I shall have more to say about Dummett's conception of a theory of meaning below.

psychological differences to us, this may be due to the fact that (unreflective) common sense suggests type-type identities for psychological states and physical make-up. Thus, we may immediately think that systems with radically different physical make-up than ours will have radically different psychological make-ups as well. But if we adopt a (more or less) standard functionalist construal of psychological events and properties,[6] where a given psychological event is conceived of as only token-identical, but not type-identical to some (complex) physical event, the examples become somewhat less intuitively forceful. Thus, the considerations in the passage provide no *a priori* reason for thinking that two systems could speak the same language and yet have radically different psychological mechanisms. For all we know at present, Dummett is wrong in asserting that beings with entirely different internal mechanisms could speak the same language.

However, there is, no doubt, room for massive variation in how a given psychological task could be accomplished. To see this, we need simply note that cognitive explanations of individual processes typically characterize them as taking a certain input then proceeding through a set of rule-governed transitions to an output state. In most cases, it is easy to find alternative sets of rules and transitions that will yield the same input-output relations. And such considerations might

6. See Fodor (1975), pp. 9 ff.

lead us to grant that it is likely that we could produce identical behaviors to those of any individual using radically different internal states. So, we probably could have two beings with isomorphic language behaviors but with completely different sets of internal states that are causally responsible for the behaviors. According to Dummett, it would follow that these beings satisfied the same theory of meaning.[7] Does this create problems for semantic naturalism, i.e. by showing either that facts about psychological mechanisms are irrelevant to semantic investigations or that the truth of semantic claims is independent from such facts?

I do not see how. For suppose that psychological processing actually differs in every possible way in us vs. other kinds of speakers (machines and aliens.) Suppose knowledge of English, S, is realized in us by one set of cognitive mechanism states S_0, in machines by a different set of states S_1, in Martians by a third set S_2 and so forth. How would this show that the study of S could not be conducted by studying one or more of the S_i's? The only possible basis for support of this claim would seem to be that the psychologist given a S_i is **unable** to abstract and idealize so as to reach a characterization of S. But this is simply false. Psychologists do idealize and abstract. Thus, psychologists might want to

7. I shall criticize Dummett's behaviorism below--for now I will grant this behavioristic conception of understanding and thus of semantics.

idealize away from features of the processes that seemed in some way irrelevant to what we were trying to account for. They might want to ignore features of processes that were irrelevant to the contents of the representations--i.e. they might want to describe the processes just in terms of the sequence of representational states involved, and they might also want to idealize away from apparent errors or breakdowns in the (normally) rule-governed processes. And psychologists might want to classify several different sequences as being of the same type, abstracting away from certain details of how the process was accomplished. Perhaps they would want to classify all processes that were input-output identical as being of a certain type. And such idealization and abstraction can continue indefinitely. But then, psychologists will arrive at Dummett's conception of knowledge of meaning sooner or later via abstraction and idealization, since a theory of meaning, on his view, will characterize some set of functions from impinging stimuli (e.g. others' utterances) to behaviors and such functions are abstractions and idealizations out of inner transitions. Thus the fact, if it is one, that beings with radically different psychological mechanisms--i.e. concrete psychological states--could achieve the same knowledge of meaning does not show that a theory of meaning is not to be based on a study of psychological states.

Nor does this example provide any basis for thinking that the truth about the nature of knowledge of meaning, S does not

depend on the nature of any one of the specific mechanisms, S_i. Apparently, it would do so only if there was some dividing point at which facts about less abstract states ceased to be relevant to facts about more abstract states. But we have been given no reason for thinking that such a point exists between S and the S_i's. A given state of semantic knowledge might exist independently of any given set of facts about processing, but only because other processing facts could realize the same knowledge state. It would not follow that the existence of the knowledge state was independent of the collective set of mechanism states.

There are cases, thought, where abstract states are governed by laws and principles that are largely independent of the laws and principles that govern more concrete states. The obvious example here is macro vs. micro physics. I.e. an explanation of the properties of the macro physical states is verified or rejected (for most practical purposes) completely independently of considerations about micro physical states and vice-versa.[8] But there is no reason to think that the case of physics is analogous to semantics and psychology. First, it appears that semantic knowledge is relevant to specific processing sequences--i.e. we explain cases of comprehension or inference in terms of the application of the speaker's knowledge of meaning. So, unlike macro and micro physics, the

8. See the wave example in Haugeland (1982).

two domains do not appear to be explanatorily independent. Second, ordinary language suggests that skills are still psychological states, which would certainly seem contrary to the idea that knowledge of meaning constitutes an independent levels of existence from psychological mechanisms. Indeed, (mental) skills and abilities, including understanding, are **paradigm** psychological states, albeit dispositional ones. Moreover, when we speak of, e.g., a person's intelligence, we often acknowledge that we are not characterizing a particular bit of a psychological mechanism, but something much more abstract. Yet, it is silly to insist that, e.g., "Jill is smart" is a non-psychological claim. Thus, there would appear to be no justification for the claim that truths about knowledge of meaning are independent of truths about psychological processes.

I conclude that Dummett's considerations give us no reason to think that semantics is somehow separate from the study of psychological processes, and hence no reason at all for thinking that the proposed naturalization of semantics to psychology involves some sort of conceptual mistake.

II. The Prima Facie Conception of Naturalized Semantics

I will now begin my defense of naturalism by sketching out a tentative model of how a theory of meaning might be developed through research into our psychological states. This is not to say that naturalism, nor even naturalization to

psychological inquiry must rest on the following methodological recipe. However, as I have suggested above, part of the burden for the naturalist is to show the *prima facie* plausibility of his position, and doing this requires having some fairly specific idea of how a scientific investigation of meaning could possibly be developed.

Like any scientific theory, a theory of meaning will begin by attempting to isolate the phenomena to be explained. It appears that our languages are systematically related to the world. Specifically, it appears that communication, including linguistic comprehension and production, and other uses of language such as inference and judgement depend at least partially on our possession of knowledge of the meaning. We want to know what this knowledge consists in and how it enables us to accomplish these tasks. Specifically, we want explanations of the nature of the cognitive states which embody knowledge of meaning.

Here, a remark or two about the nature of cognitive psychology itself is in order. The cognitive program seeks to explain our knowledge and abilities through the postulation of various representations and processes that operate on those representations. The representations and processes need be neither consciously accessible nor recognized by common sense psychology. In fact, many current cognitive theories characterize representations that are both sub-conscious and are not readily identifiable with common sense attitudes such

as beliefs or desires.[9] (On the other hand, it is not unreasonable to suppose that cognitive theories will eventually either include reference to common sense states or show them to supervene on states that the scientific theories refer to.)

I suggest that the best means of naturalizing semantics to cognitive psychology--let us call the resulting view **cognitivist semantics**--is through the following line of investigation. Theorists begin with a set of postulations about the meaning of natural language words and statements. These postulations might be drawn directly from our ordinary intuitions about meaning, or they might be more complex models, developed on the basis of these intuitions. The initial postulations will, no doubt, include a number of non-equivalent competing views about the semantic features[10] of linguistic (i.e. syntacticly specified) structures. The next step is to determine the role that these semantic features might play in various psychological tasks. Here, pre-theoretical beliefs suggest that semantic comprehension and production, inference and judgement will be the most

9. Thus, I am not necessarily advocating an intention-based semantics, if 'intention' is understood as referring to the common sense propositional attitudes such as belief and desire, but something broader--a "representation-based" semantics.

10. Note that competing theories may provide alternative accounts of what these states or relationships are as well as alternative assignments of specific semantic features to words and phrases.

significant abilities that draw upon or are in some way dependent on languages possessing the postulated semantic features. This in turn leads to the postulation of psychological states and processes which are sufficient to enable the realization of these abilities. Non-equivalent, competing views of semantic features should, in most cases, lead to alternative postulations of underlying states and processes.

Perhaps the first stage, or set of stages might be conducted along the lines that have proven successful for theories of syntax. That is, we might attempt to develop a competence theory for semantics. A competence theory, a methodology introduced by Chomsky,[11] is a characterization of the knowledge which will, under ideal conditions, enable the accomplishment of a certain goal or task (e.g. communication.) Standardly, a linguistic competence theory is constructed by developing rules which characterize structural descriptions that in turn reflect speaker's intuitions about the features (e.g. well-formedness, deviance) of various phrases.[12]

11. See Chomsky (1965), pp. 1 ff. See also Chomsky (1986), Chapters 1-2, where a terminological shift occurs, from "theory of competence", to "theory of I-language."

12. A competence theory may also include rules whose application yields transitions between postulated underlying structures. However, such rules should not be confused with rules that are actually used in processing. The former are perhaps best viewed as abstractions out of or over the latter processes.

The next stage is the difficult task of attempting to determine observable measures that will help decide which of the postulated representations and processes we actually possess. Standard measures for representational states and processes include reaction times, error rates and types, and task performance under certain unique circumstances. While there is no guarantee that data can be found which will decisively determine which states and processes are present, the situation is no different than that faced by any other theoretical pursuit in natural science which postulates unobservables. Suitable linkages of unique data to theories will provide a basis for deciding between competing views. At this stage, it may also be useful to draw upon results from independently established theories of non-linguistic abilities and processes, e.g. of theories of perceptual processes. As with any domain of science, compatibility of related theories yields support for the theory under investigation, while apparent conflict may lead to revision or even theory abandonment.

Finally, any revisions that occur at any stage of theorization may potentially lead to modifications of the originally postulated semantic features. This may include revising or rejecting some of the intuitions which form the original basis for the postulation of semantic features. In this way a theory of meaning can be regarded as a fully empirical undertaking, where the theory is potentially subject

to revision with each piece of recalcitrant data concerning psychological states and processes.

Although this methodology has not as yet been developed to the point where sample explanations can be presented, it will be useful to consider several highly speculative examples in order to illustrate the relationship between questions about content and questions about our cognitive processes. Consider, for instance, the issue of what logical form underlies definite descriptions. Russellians maintain that a phrase of the form:

> the Φ is *P*

has the logical form:

> there is something that is Φ and nothing else is Φ and it is *P*

By contrast, the (Fregean/)Strawsonian view holds that phrases such as the Φ have the same logical form as names--i.e. they are singular terms. On this view, it is standardly assumed that the uniqueness of reference of definite descriptions in presupposed when such phrases are used, but the uniqueness of reference is not specified by the logical form itself. What might each view predict about the mental representations underlying our understanding of definite descriptions? The obvious difference is that the Russellian view predicts that mental representations underlying "the Φ" will be complex while the Strawsonian view predicts that such representations will be simple. Of course, complexity is relative to function.

Specifically, we should expect a difference between these views when it comes to inferences based on the representations which underlie our understanding of definite descriptions. I.e. on the Russellian view, we should expect the representations underlying our understanding of "the Φ" to be complex when it comes to inferences. That is, we should expect the representations themselves to license the inference from "the Φ is *F*" to "something is *F*". By contrast, defenders of the Strawsonian view should hypothesize that the representations in question do not license any such inferences, i.e. that they such representations are computationally simple as regards inferences. This view will need to explain (apparent) inferences from "the Φ is *F*" to "something is *F*" somehow, e.g. as based on a storehouse of facts about conversational presuppositions. Hence, evidence supporting either the inferential complexity or simplicity of the representations that underlie our understanding of definite descriptions should help decide between these completing accounts.[13]

As a second example, consider, a disputed case of analyticity. E.g. the statement 'cats are animals' seems analytic to most speakers. But suppose, as Putnam has

13. It probably won't be this simple, of course. For one sort of developments, see Hornstein (1984) for a defense of the Strawsonian view based on facts about types of quantification drawn from a (largely) syntactic competence theory.

imagined, that we discover cats to be cleverly constructed robots, of alien manufacture (e.g. Mars), placed on the Earth to spy on us. In such a scenario, Putnam claims, we would admit that our experience has shown the falsity of 'cats are animals.'[14] In response, a defender of analyticity might claim that the statement is analytic, and that in such a case we would say that we have discovered that there are, in fact, no such things as cats (at least around here), or that we have discovered that some animals are robots.

Who is right? While cognitive semantics has hardly advanced far enough to provide an answer to this question, it is possible to sketch a speculative development which will demonstrate how cognitive findings could help settle this matter. Suppose, that it turns out that there is a cognitive semantic framework used in the comprehension of sentences, a framework that is not alterable in the way that ordinary stored facts (e.g. "cats like to play with yarn") are revisable. Thus, 'cats are animals' might ordinarily be comprehended in a way that makes it analytic, since, e.g., the 'cat' node is subordinate to the 'animal' node in the semantic framework. But this may not constitute the entire cognitive realization of 'cat's meaning--the term is also tied to cat recognition, in that it is the label used for such instances.

14. See Putnam (1962) for introduction of this case--though he does not use it to attack the intuition of analyticity there--he calls the statement "analytic."

Thus, a positive case of "cat-recognition" may excite the 'cat' node.

We can apply this crude cognitive model to see where the conflicting intuitions arise. Consideration of robot cats suggests that in such a case, it is not appropriate to have the 'cat' node--for what we recognize as 'cats'--subordinate to the 'animal' node; a different categorical structure must be used for comprehension of the term, e.g. one where the 'cat' node is subordinate to 'machine.'

We might say that this cognitive analysis has shown us that 'cat' in fact has two components to its meaning, "cat_1" for which it is analytic that cats are animals and "cat_2" which means roughly "the things we typically recognize as cats." Thus, 'cats are animals' has both an analytic and a synthetic component to it. Ordinarily, these coincide. However, in the imagined case they diverge, we have synthetic grounds for rejecting '$cats_2$ are animals', so '$cats_1$' must be abandoned in favor of, say, '$cats_3$', where '$cats_3$ are machines' is analytically true. Thus, cognitive theory might provide us with a basis for claiming, e.g. that 'cats are animals' is analytic and that cases such as the robot discovery would lead to changes of meaning. (Or perhaps we will find that the meaning is determined by the perceptual component alone--or, e.g., if it turns out that only at least one component must be used in comprehension--disjunctively as it were--then it seem

that the reaction to the original case that has it that the statement is not analytic is correct.)

Let us return to general considerations about the methodology for a cognitivist semantics. As I have presented it, this methodology appears as a rational progression of stages, each dependent on the previous one's near completion. Roughly, 1) construct competence theories--which may take a number of stages in itself, 2) postulate (abstract) processes and abilities that realize competence, 3) develop specific theories about the representations and computations that realize these processes and abilities 4) tie the postulated representations and computations to specific behavioral predictions and, coinciding with (2-4), 5) seek corroboration from independent psychological theories of the realizations of other abilities and processes. However, with science being the methodologically messy enterprise that it is, we might expect that investigation will, to some extent, be conducted at all stages simultaneously. Indeed, it appears that work is already underway at most stages,[15] although with the early stages

15. For instance, Jackendoff (1983) presents a partial semantic competence theory, Johnson-Laird (1983) presents a theory of the psychological processes underlying semantic comprehension and Jackendoff (1987) is, consciousness aside, an attempt at linking postulated linguistic faculties with other postulated faculties, notably visual and musical faculties.

These are only token, highly representative examples. Much of the work in cognitive psychology concerning concepts and categorization, and concerning language-processing could be construed as providing partial, tentative accounts for one or another of the stages I have outlined.

relatively undeveloped, theories at the later stages become extremely tentative. Whether or not a given stage will prove more difficult than others is an open question that I will not address here. I merely emphasize that all stages in this conception appear to require substantial inquiry--this model suggests how empirical semantics can be conducted, but it does not imply that it will be easy.

Is this, then, a plausible naturalization of semantics? One major issue that I will not examine here is whether or not a methodology which postulates representations and representational processes is indeed legitimate and can in fact be successful. I assume that the answers to both questions are affirmative for the same reason that I assume that various other scientific methodologies are legitimate, namely that there are, at present, modestly successful scientific theories that those methods have produced.

Beyond such worries, there is, I claim, nothing incoherent or conceptually impossible about any particular stage of the undertaking. While there is no guarantee of success at any particular stage, just as there is never any guarantee of success for any attempt to produce a scientific explanation of any particular phenomenon, there is also no reason to think that any particular stage cannot be successfully accomplished. Thus, this model of cognitivist semantics provides a *prima facie* case for the naturalization of a theory of meaning to cognitive psychology.

A. Meaning and Representation

I will now consider several objections to the view of naturalized semantics that I have just sketched. The first concerns the assumption of a representational psychology. Specifically, a cognitivist semantics will rely on theories that postulate various mental representations. Perhaps it might be objected that this presupposes that the content of such representations can be specified. But providing an account of the notion of representation that is at work here,[16] it might be claimed, is an *a priori* line of investigation, and moreover one that, if successful, will already provide an answer to the question which the cognitivist semantics was supposed to answer, namely what the content of linguistic expressions is. Thus, it might be argued, the proposed account relies on an *a priori* investigation and therefore does not provide a naturalization of semantics after all.

In response to this line of objection, I will distinguish between **meaning** and **representation**, and also between a theory of meaning and an analysis of representation. Doing so shows us that even if the latter is an exercise in *a priori* analysis, the former can still plausibly be regarded as an empirical theory. Let us then let 'meaning' stand for the

16. See, e.g., Cummins (1989) for presentation of the current alternatives, and defensive of his interpretational view. Also see Fodor (forthcoming) and Block (1986).

relation(s) that elements of natural language bear to items in the world, and 'representation' stand for the relation(s) that mental entities bear to items in the world. Notice that it appears to be an open question as to whether or not these are the same type of relation (or properties, etc.) Thus, having a conceptual analysis of "representation" does not automatically imply that the same analysis applies to "meaning", or even that there is a conceptual analysis of the latter notion.

And if there is an *a priori* analyses of "representation," it does not follow that a theory of meaning is an *a priori* undertaking. Given a broadly cognitivist model for meaning, a theory of meaning will need to tell us what mental representations and states are relevant to the meanings of natural language expressions, what the relationship between the mental states and expressions consists in, and what the representational content of the mental states is. An analysis of the notion of representation would seem to provide us with nothing beyond a partial answer to this third question. The bulk of a theory of meaning would be undetermined--we would need to investigate what mental states we have, which are relevant to a theory of meaning, and how the relevant ones determine meaning. And further, even the question of the content of particular representations would seem to require empirical investigation. E.g., if the analysis is in terms of certain causal relations, then we still need to look to see when and if such causal relations hold between specific mental

states and items in the world. So the possession of an analysis of "representation" would not affect the plausibility of the claim that these other tasks are part of an empirical program.

To help see this point, it is important to note that the cognitive picture is not in any way committed to the view that each word in a language is mapped 1-1 to a representation, where that representation's content is equivalent to the meaning of the word.[17] On such a model, the use of representations seems almost superfluous--adding a simple mapping from words to representations accomplishes very little. And the questions of what representations we have and how they relate to language are trivialized. What is more likely, though, is that a complex set of representations and cognitive processes is associated with each term, or linguistic structure. For instance, it may be that most common sense kind terms are given their meaning via an association with detection mechanisms, a "semantic network" that places the kind in a conceptual hierarchy, and a description or image of an exemplar of the kind. Thinking of the cognitive approach to semantics only in the first way, in terms of an almost trivial word to representation mapping, may suggest that the only interesting project for this model of semantic investigation is a specification of "representation." But this

17. The most well-known version of this view is found in Fodor (1975).

is only one possible model of how cognition connects language to the world, and it is an open matter as to what sort of model is correct. While the word-representation relationship may turn out to be more or less trivial, we cannot know this prior to empirical investigation, and can only establish it by non-trivial investigation.

So cognitivist semantics does not rely on an *a priori* analysis of representation that will actually do all of the work in an account of meaning. But I wish to make an even stronger claim here, namely that there is no reason to think that there is any *a priori* investigation proceeding here at all. Philosophers who are currently seeking a naturalistic analysis of the notion of representation may be viewed as engaging in a certain empirical task, i.e. the job of attempting a theoretical explanation of the instantiation of "representation" in non-psychological terms.

Cummins has argued that sciences, psychology in particular, not only involve transition theories that explain causal properties but also analyses that explain the instantiation of properties:

> Many scientific theories are not designed to explain changes but are rather designed to explain properties. The point of what I call a property theory is to explain the properties of a system not in the sense in which this means ""Why did S acquire P?" or "What caused S to acquire P?" but, rather, "What is it for S to instantiate P?" or, "In virtue of what does S have P?" Just as we can ask, "Why did the gas get hotter (or expand)?", we can ask, "In virtue of what does a gas have a temperature (volume)?" Understood as an answer to the latter questions, the kinetic theory of heat (and the

> molecular theory of gases that it presupposes) is not a transition theory but a property theory: it explains temperature in a gas by explaining how temperature is instantiated in a gas; it does not, by itself, explain changes in temperature.[18]

It seems that the task of providing a naturalistic analysis of the notion of representation might be reasonably viewed as just this sort of undertaking--a bit of very abstract biology or physics, or perhaps a bit of "theory of theories." Therefore, the model of cognitivist semantics, in so far as it relies on a notion of representation, need not, contrary to the objection, rely on any sort of *a priori* account.

And finally, note that even if we never get a naturalistic account of "representation," it is possible that we will have successful theories that postulate mental representations and specify their content. I.e. cognitive psychology might succeed without reductionistic foundations in the same way that mathematics has succeeded (apparently) without decisive foundations. The proposed semantic methodology merely requires that we have successful theories of representational states--no explicit appeal is ever made to an analysis of "representation." So it is not really apparent that cognitivist semantics requires any treatment of the notion of representation at all.

18. Cummins (1983), pp. 14-5.

B. Psychologism

Perhaps some readers will be worried that the view of cognitivist semantics that I have presented is committed to an implausible claim, namely that meanings are in the head. I will now address such concerns.

First, it should be noted that the methodology I have endorsed is not committed in advance to any particular conception of what meanings--or "semantic features"--are. We might divide traditional accounts into two types--those that endorse only relational semantic features, e.g. reference, and those that also endorse intrinsic semantic features, e.g. sense. Cognitivist semantics, as I have said, is not committed to either type of view in advance of theoretical investigation. However, to the extent that there do turn out to be intrinsic semantic features, it seems plausible for the cognitivist to think of them as either being abstract mental states, or as being determined or fixed by mental states. And to the extent that there turn out to be relational semantic features, it seems plausible for the cognitivist to maintain that they are largely determined by mental states, although, of course, identification of an intrinsic mental state with a relational state amounts to a category mistake.

In this innocuous sense, the cognitivist is committed to the view that meanings are in the head--i.e. that meanings are determined by mental states. As I shall now point out, this

does not commit the cognitivist to either of two problematic psychologisitc views, namely subjectivity and individualism.

Traditional psychologism claims that meanings are the sort of mental states that can be known subjectively, i.e. from the first-person point of view, no matter what the external environment is like. As far as I can tell, the cognitivist methodology that I have outlined has no commitments at all concerning subjective knowledge. Mental states, as conceived by cognitive psychology are as objective as any other sort of states. While (apparent) introspective knowledge and intuitions might prove useful in forming theories of psychological states, there is no reason for the cognitivist to be committed to the view that such knowledge and intuitions are either infallible or provide complete semantic knowledge. In fact, below I will argue that it appears to be **false** that we have complete explicit, (potentially) conscious knowledge of meaning. Thus, the cognitivist is not committed to any sort of subjective psychologism.

A second problematic commitment for some forms of psychologism is the thesis of individualism, i.e. that the meanings of terms are determined solely by what is in the head. This traditional psychologistic thesis has come under sever attack through Putnam's twin-earth cases and Burge's subsequent development of them. But this has no obvious bearing on the cognitivist semantic methodology. That is,

there is no particular reason for the cognitive semanticist to be committed to the view that the content of mental states is determined solely by what's in the head. While the slogan "meaning ain't in the head" might seem to directly oppose psychologistic semantics, this is only because the slogan is highly misleading. All that the twin-earth and similar cases show is that meanings are not **solely** in the head. Yet, language comprehension and production and inferential abilities surely are in the head, and meaning is relevant to these, so there must at least be a substantial portion of meaning determination in the head even if anti-individualism is correct--a substantial portion worthy of investigation. And what's not in the head might be left to sociology or related sciences. So a semantics naturalized (mostly) to cognitive psychology is perfectly consistent with anti-individualism. In fact, we might even want to infer from the intuitions against individualism, as Burge does, that psychology isn't concerned solely with what's in the head. So even if individualism is incorrect, semantics might be completely naturalized to cognitive psychology.

One final point on these matters concerns Fodor's recommendation of "methodological solipsism" as a research strategy for cognitive psychology.[19] He argues that the most promising methodology for cognitive psychology to pursue is

19. See Fodor (1980).

one that focuses only on intrinsic mental features rather than on relational features. Perhaps the same recommendation might be apt for cognitivist semantics. Thus, there is no reason to believe that the relations that form the external, mind-independent part of word-world relations exhibit any salience or order that make them good candidates for scientific study. In this sense, the only **interesting** candidates for a scientific semantics may be mental states. For instance, one could imagine that we will be able to develop theories which specify how our words inherit their meaning from mental states and processes, although we will find the task of specifying the specific relations that our mental states stand in to external objects too difficult to capture theoretically. That is, such relations may be too complex and diverse for us to be able to provide general theories about them.

In any case, it is important to see that the methodological solipsist strategy is distinct from both subjectivism and individualism. The recommendation that the psychologist or semanticist study only intrinsic psychological states says nothing about what knowledge of those states consists in. Specifically, there is no reason for the methodological solipsist to be committed to either the view that we have subjective knowledge of all our intrinsic mental states, or even the view that we have (correct) subjective knowledge of any of these states. Second, the methodological solipsist strategy need not be conjoined with individualism,

which is a metaphysical thesis. One could reject the latter, but still maintain that the only the study of states from (e.g.) the skin in offered hopes for a promising research program. Thus, methodological solipsism is a strategy which may prove highly successful for cognitivist semantics, though I remain officially neutral on this matter in the rest of the essay.

III. Knowledge of Meaning--The Need for Scientific Investigation

So far I have been defending the plausibility of the specific form of naturalism that I have proposed for semantics--naturalization to cognitive psychology. I now turn to the second main portion of my case for semantic naturalization, the claim that there is no alternative, non-scientific source of knowledge sufficient to yield a non-scientific methodology for a theory of meaning. My argument will have two parts. First, I will address the general view that non-scientific knowledge of meaning sufficient for a semantic theory is provided by ordinary knowledge of language. Then I examine a number of specific conceptions of meaning and semantic methodologies, each of which might appear to offer an alternative source of semantic knowledge and thus an alternative, non-naturalistic methodology. I shall argue that they in fact do not.

A. Kinds of Knowledge

It would appear that the main opposition the recommendation that semantics be naturalized to scientific inquiry is a position that maintains that we already have (relatively) complete knowledge of the meanings of our terms in virtue of being able to speak a language, so there is no need for empirical investigation since we already know all that we need to in order to develop a theory of meaning.

In response, I claim that the knowledge that we have in virtue of being able to speak a language need not be explicit, conscious knowledge that is either readily accessible or can be expressed through a bit of reflection. We are able to speak languages, and thus can be said to have knowledge of the semantic aspects of these languages. But this fact alone does not show that we have explicit knowledge of the meanings of the terms of the languages. It is possible that our knowledge of meaning is either tacit or implicit.

By saying we have **tacit knowledge** of *p*, I mean we may have explicit representations of *p* that are accessed by certain processes, e.g. language comprehension processes, but which we are not able to access in a way that allows us to generally report on them. Thus, tacit knowledge is essentially unconscious (or **non**-conscious) knowledge. By saying that we have **implicit knowledge** of *p*, I mean that while we do not have an explicit representation of *p*, some behaviors or ability can be successfully explained by postulating such a

representation. However, in most such cases it is likely that some other representations and processes actually account for the ability or behaviors. Thus, implicit knowledge ascription is essentially an instrumental means of preserving certain successful knowledge explanations. I shall illustrate each sort of non-explicit knowledge and suggest how each type might realize our semantic knowledge.[20]

The paradigm case of tacit knowledge is our knowledge of the syntax of language, as characterized by theories from the research program stemming founded by Chomsky. Certain evidence, e.g. speaker's intuitions about the well-formedness of sentences, suggests that we possess explicit rules which allow us to produce (and apply) representations of the underlying (syntactic) structures of sentences. However, we

20. A deep issue here, which I shall not attempt to address, is the question of when we can definitively assign implicit knowledge. E.g. any bodily movement could be explained as though we had representations of the complete laws of physics at our disposal. Perhaps the whole idea of implicit knowledge is spurious, and we should simply abandon such explanations, no matter how successful, in favor of the alternative explanations. Nevertheless, I present the idea here because it has seen some popularity in both philosophical and psychological circles.

Also note that I am not necessarily using the terms 'tacit' and 'implicit' as others use them. There is a frustrating lack of agreement about how to use these terms, just as there is a lack of agreement about what the alternatives to explicit knowledge are. Nor is there any agreement about whether, e.g. non-conscious knowledge is really "knowledge." However one uses these terms, the present claim is that some cases we ordinary label as "knowledge" (of *p*) could turn out to be cases where a representation that *p* is present but not consciously accessible or where alternative representations underlie these ascriptions.

clearly have no explicit, conscious knowledge of such rules. Thus, the following method has proven successful for the study of syntax: native speakers' intuitions concerning the (e.g. syntactic) well-formedness of sentences are used as data for formulating theories that are taken to characterize rules that speakers' tacitly know.

This yields somewhat of an "armchair" enterprise, in that the relevant data is readily accessible to native speakers. But this should not be confused with the idea that linguistics is an autonomous enterprise independent of empirical psychology. Chomsky's view is that this methodology is no more than a means of attempting to abstract away from inappropriate psychological data towards formulating a theory of one aspect of our psychology, viz. our knowledge of the syntax of language. Which is to say that the "armchair" methodology is not to be conceived of as yielding a domain of inquiry that is in principle distinct from psychology. So we should expect that a competence theory will be subject to potential modification when it comes to producing a performance theory of language use--data and theories of the latter may potentially cause revisions in the former theory. And, while there is an initial reliance on speaker's intuitions, we should on this view of the matter, regard such intuitions as potentially fallible, so that other data may ultimately cause us to revise or reject these data. Thus, the idea of a (more or less) armchair methodology in the study of competence does

not show that such an area of study is something other that a branch of empirical psychology.[21] And, if there is also a set of semantic rules which can be studied based largely on native speakers' intuitions, this does not provide any support for the claim that semantics is not a part of empirical psychology.

While there is no corresponding paradigm case for implicit knowledge, there are plausible examples. For instance, in knowing how to ride a bike, it is likely that we do not have any explicit representations of any of the actual laws of mechanics. But it is plausible to suppose that knowing how to ride a bike does involve the possession of appropriate rules (or heuristics) for coordinating motor behaviors with sensory input (and intentions). However, we can successfully explain our ability to balance and steer by postulating explicit knowledge of certain mechanical principles. Thus, we might say that our bike-riding know-how involves implicit knowledge of such principles.

If our semantic knowledge is implicit, then while there is no explicit representation of the meanings, various

21. Note that a long running debate in the foundations of linguistics concerns the psychological reality of the grammars that linguists study. Given the present distinctions, the central question may be formulated as, is a grammar tacitly known, or (merely) implicitly known? For two recent views, see George (1989) and Peacocke (1989). However, neither does justice to Chomsky's views of competence as idealization, or to his conceptualist (as opposed to platonist) views about the nature of language itself.

representations and processes may function in a way that produces the appropriate relations between our words and the world. For instance, it may be the case that the meaning of 'cup' is determined by a number of different representations and processes, none of which individually means "cup". It may be that the psychological determination of this meaning is a result of both (perceptual) prototypes used in recognizing cups, e.g. a set of features used in detecting cups, and a set of features (or "markers") in a lexical entry, e.g. +artifact, +object, etc. In such a case there need be no single, accessible representation that is the "meaning of 'cup,'" yet it is (apparently) as though we had such a representation.

Another possibility as far as implicit knowledge of meaning is concerned is that while individual representations with appropriate representational contents underlie our semantic competence, we do not access that representational content either consciously or unconsciously. E.g., suppose that knowing that 'dog' means dog is a matter of possessing a token in mentalese that refers to dogs, but also suppose that we do not access the representation's semantic features. Those who do not like the label 'implicit knowledge' here might prefer to say that while we **possess** or **instantiate** meanings--e.g. just as we instantiate the digestive process--we have no representation of the meanings. However, this is not to say that there is no difference between cases we now label as "knowing the meanings" and cases we label as "not knowing the

meanings." The suggestion is that in the former cases but not the latter, the speakers have correctly associated the relevant terms with appropriate mental representations, even though it may be inappropriate to describe them as having access to the content of these representations.

Thus, the fact that we know the meanings of the terms in the languages we speak does not show that such knowledge can form the basis of a semantic methodology that does not require recourse to empirical psychology. Such knowledge may be either non-conscious or it may not be knowledge at all, so much as a matter of possessing appropriate representations that are appropriately associated with terms in the language. In either case, we require scientific investigation of cognition in order to reveal and explain our non-explicit knowledge.

B. Specific Conceptions of Semantic Knowledge and Methodology

The general concern of this section is to show that there is not any non-scientific methodology that is capable of yielding a semantic theory. I have argued that ordinary knowledge of language does not, *per se*, provide the materials for a non-scientific theory of meaning. What I will now do is turn to an examination of a number of specific semantic methodologies and conceptions of meaning that might appear to provide such a basis. In each case, I will argue that the source of knowledge of meaning that the methodology relies on

does not provide a sufficient basis for a non-scientific semantics.

1. Truth and Reference

Davidson's suggestion, following Quine, that we substitute the notion of truth for that of meaning in a recursive semantical theory has been widely adopted in philosophical circles, thus inviting the conclusion that truth and reference are the sole or at least the key notions in a theory of meaning. Perhaps it might be suggested that we have explicit (or potentially explicit) knowledge of the references of terms or of the truth conditions of sentences, and such knowledge is sufficient for constructing a semantics, in a Davidsonian manner, without the need for empirical psychology.

Do all competent speakers have explicit, or potentially explicit knowledge of reference and truth conditions for the words and sentences of their languages? Perhaps it might be suggested that any competent speaker will assent to disquotational facts such as:

'Snow is white' is true iff snow is white

'Snow' refers to (or, e.g., "stands for") snow

But, as has been noted on several occasions,[22] knowledge of such facts is not appropriately characterized as knowledge of truth conditions or reference. Thus, suppose that, in the

22. See Higginbotham (1989) who in turn credits Harman with this insight.

ordinary sense of the phrase, you do not know the meaning of 'hautbois.' However, if I tell you that 'hautbois' is an English count noun, then you will known that:

(1) 'hautbois' refers to hautbois.

and:

(2) 'Something is a hautbois' is true iff something is a hautbois.

However, those who **really** know the meaning of this term will know further truths, e.g.:

(3) A hautbois is an oboe.

(4) 'Hautbois' refers to oboes.

Thus, it seems that knowledge of disquotational truths such as (1) and (2) is not knowledge of reference or truth conditions, or is at best a very partial knowledge that is insufficient for attributing semantic competence.

Therefore, if the claim that speakers have potentially explicit knowledge of reference and truth conditions is to be sustained, it must be on grounds other than the fact that speakers assent to disquotational truths. However, when we look beyond disquotational knowledge, the claim that we have (potentially) explicit knowledge of the reference and truth conditions of our languages becomes extremely dubious.

Consider reference. Often, there is much about the extension of terms that competent speakers do not know. For instance, most speakers do not know that 'brown' does not refer to a spectral color. Nor do most know that the primary

spectral color terms exhibit the following features: There is a range of the spectrum which each term will be judged to correctly classify, and this range trails off gradually into adjacent ones. Within this range there is a narrow band of the spectrum that will be judged as the "true" variety of this color (e.g. true red), although the width and location of this band on the spectrum varies from person to person. Or consider, is a virus an animal? Surely, most speakers can't answer this, yet, if they knew the references of the terms 'animal' and 'virus,' we might expect them to be able to determine if those extensions overlapped or not. Or think of false beliefs about reference. Many competent speakers think that a tomato is not a fruit, and some think that whales are fish. Again, it would seem that if semantic competence included some sort of explicit knowledge of the reference of these terms, they would know otherwise.

Further, consider that there are problematic questions concerning reference that speakers surely cannot answer. E.g., what is the reference of 'two' or 'justice?' Do abstract notions refer to abstract objects, or to mental states or to collections of concrete objects? Does 'phlogiston' refer to something which doesn't exist, or does it not refer at all? Does 'red' refer to a dispositional or intrinsic property of the world or neither? Does 'I' refer to consciousnesses, brains, or what? Again, if we had any sort of substantial (potentially) explicit knowledge of the reference of our

terms, it would seem that we should have answers for such questions, but we do not.

Finally, speakers need not be able to give much of a recipe for specifying the extensions of terms whose meaning they know. E.g. someone who knows what 'fruit' means might be able to do nothing but name a number of typical fruits, which certainly wouldn't show us that (e.g.) tomatoes or kiwis are in the extension. Yet, if we had explicit knowledge of reference, we might expect that such a specification would issue from this knowledge.

The idea that we have (potentially) explicit knowledge of truth conditions, where truth conditions are viewed as something other than knowledge of reference, is even more problematic. Thus, consider "logical form." Surely, we do not have explicit knowledge of anything like the propositional calculus or (the linguists') LF--knowledge of these forms is explicitly taught in the same way as any other (scientific) theory, and sometimes competent speakers cannot explicitly master the appropriate concepts. And, in any case, there are ongoing disputes about the logical forms of various sentences, but we should expect such disputes to be readily decidable if all competent speakers possessed (potentially) explicit knowledge of logical forms.

It might be suggested (but not by Quine or Davidson!) that what we do have is knowledge of semantical entailments, and that this could be used to determine the logical forms of

statements without recourse to psychology. However, there are well-known problems with this view. As Quine and others have pointed out, it's simply not apparent that we can distinguish conceptual from non-conceptual entailments. Thus, the first inference seems acceptable and the second unacceptable:

Bob or Sally will win.
Therefore, Bob and Sally will not both win.

John is greedy.
Therefore, if John is a Republican, then he is greedy.

There is no obvious "common sense" way to tell if these acceptability intuitions are due to the meanings of the terms, or to pragmatic aspects of language use. In general, it is not apparent that we have definitive intuitions about a wide enough class of semantic entailments in order to establish the logical form of sentences. What we have, at present, are conflicting, competing accounts of the logical forms of natural language expressions and there is no reason to believe that all such conflicts can be resolved without recourse to the data and theories of empirical psychology.

Is there some other construal of "truth conditions" which makes it plausible that we have potentially explicit knowledge of them? If truth-conditions are understood in some metaphysical sense, then we face the problem that outside of

philosophy, few competent speakers have knowledge of such matters. E.g., few competent speakers have explicit knowledge of the set-theoretic universe championed by some model theories.

Thus, it seems that speakers who know the meanings of terms and sentences do not necessarily have potentially explicit knowledge of the reference or truth conditions of the elements of their languages. Therefore, if we are going to buy into the Davidsonian view that a theory of meaning is to be developed using the notions of truth and reference, we have no reason not to maintain that this is an empirical psychological theory, e.g. one which will characterize speaker's tacit or implicit knowledge of reference and truth conditions.

2. Interpretation

The other cornerstone of the Davidsonian conception of semantics is the view that a theory of meaning should be a theory which provides interpretations of speaker's utterances. Perhaps it might be maintained that a theory of meaning could be developed that is based solely on ordinary interpretation practices, without any recourse to scientific psychology.

The independence of an interpretation-based semantics from a scientific semantics could be understood in any of the following ways: it could be held that the interpretation methodology will be **complete**--that it will explain everything,

or most everything about meaning that could possibly be explained. Or it could be
claimed that while the results of interpretation methodology will not explain all aspects of meaning, the explanations the methodology does produce will be insulated, as a result of the methodology, from overthrow by alternative scientific accounts. That is, it might be maintained that the results of the investigation are **autonomous** in relation to scientific semantics. Or, if not complete or autonomous, it might be maintained that the goals and results (to date) of interpretation methodology will at least be **required** by a scientific semantics, even if the latter reforms the results of the former to some extent.

However, as I shall now argue, none of these claims is plausible. First, consider the view that interpretation methodology will yield a complete semantic theory. The problem here is that our ordinary interpretation skills rest on tacit or implicit knowledge of meaning, knowledge which is apparently not revealed by ordinary interpretation. To see this, note that to interpret another's utterance, on the Davidsonian view, is to relate it to a phrase that the interpreter understands. Thus, a sample interpretation will be:

> A's utterance, "snow is white", on occasion O, is true iff snow is white.

In order to maintain that interpretation methodology will provide a complete semantics, the interpretation theorist must show that we have good reason to think that such native understanding will be explained by interpretation methodology.

But this is implausible. As we have seen in the preceding section, it seems that we do not have potentially explicit knowledge of many facts concerning the reference and truth conditions which we assign to statements in ordinary understanding. Nor is it clear how interpretation practice could uncover such facts. Suppose that some of us possess mental state *M* which contributes some semantic feature as part of understanding--let us suppose it fixes the reference of some term--although we do not have explicit knowledge of this fact. Now, what will happen as far as interpretation methodology is concerned? Either an interpreter will have *M* as part of her non-explicit understanding or she will not. Suppose she does. Then she will assign the reference that *M* fixes to terms of other speaker's utterances, when interpretation methodology so dictates (or recommends.) But doing so in no way uncovers, explains or explicates the facts about what reference *M* fixes. Nor does any sort of self-interpretation serve to uncover these facts. Making use of non-explicit knowledge does not make it explicit. On the other hand, suppose the interpreter does not possess *M*, or something equivalent as far as semantic theory is concerned. Then, as Davidson has repeatedly emphasized, she will not be able to

interpret others' terms as having the appropriate reference.[23] So it appears that interpretation methodology will not uncover non-explicitly known semantic facts. Given that there appear to be some such facts, it seems that interpretation methodology cannot provide a complete semantic theory.

At this point it might be tempting to argue that understanding really is nothing more than interpretation, so interpretation methodology must yield a complete semantics. Consider, for instance, Davidson's claim that "translation begins at home":

> The problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same? Speakers of the same language can go on the assumption that for them the same expressions are to be interpreted in the same way, but this does not indicate what justifies the assumption. All understanding of the speech of another involves radical interpretation.[24]

Here I think Davidson is primarily making the point that, if Quine's mythical radical translator thought-experiment is valid, then the results hold not just for actual radical translation, but for any interpretation of another's speech. In other words, the available evidence which we use to assign interpretations to our neighbor's utterances does not overcome the (apparent) gaps in determinacy that the case of radic l translation is supposed to reveal. However, the stated

23. E.g. see Essay 13 in Davidson (1984).

24. Davidson (1984), p. 125.

conclusion appears to run this point together with another. I.e. the conclusion might be read as the claim that the process of language comprehension is the same process as that of interpretation. It might be claimed that whenever we comprehend the meaning of any utterance, we are using the very same processes, or doing the very same thing that we are doing when we attempt to translate a foreign language, or when we attempt to make rational sense of the behaviors and speech acts of another. But Davidson's point in the quoted passage does not support the identification of meaning-assigning and intention-assigning processes. The fact that, as Davidson claims, we are not justified in assigning a given meaning says nothing about how such tentative meaning assignments were produced. For instance, we may think of our (largely subconscious) linguistic processors as making the assumption of homophony. If we grant that there is an epistemic equivalence between our case and the mythical radical translator--we are no more justified in our homophonic translation that the radical translator is in affirming one of a variety of evidentially equivalent translation manuals[25]--it does not follow that the processes that produce our homophonic translations are the same ones that the imaginary linguist would use in the process of developing an interpretation.

25. See Quine (1960), Chapter 2.

Moreover, there is no reason to think that by attempting the radical interpreter's task we could duplicate the meaning assignments that we typically produce tacitly or implicitly. For instance, it could well be that case that our ordinary knowledge includes knowledge (tacit or implicit) of semantic universals, and other universals of cognition for that matter, which enable us to acquire meanings from a relatively small set of choices, and which in turn enable the selection of one, or occasionally several, meaning assignments for a given utterance. The only way to develop explicit knowledge of such universals and their role in cognition is to empirically investigate our cognitive processes--that is, in essence, the point of this essay.

Notice that the position I am advocating leaves it open that we may really have no good grounds for believing that our neighbors are speaking the same language that we are. There is a tendency to assume (as we shall see with Dummett below) that there must not only be some ultimate evidence concerning which language, if any, someone is speaking, but that such evidence must be available to every speaker. However, it seems perfectly intelligible that the average speaker not have determinate evidence about what language is being spoken, in just the same way that he has no justification for induction and no explanation of the foundations of his mathematical and ethical beliefs. Some utterances (e.g. those in "English," for me) sound intelligible, some do not. Perhaps this means that

my language faculty "assumes" that it is receiving English input in much the same way that my visual system "assumes" that it is receiving input caused by a three-dimensional, spatial-temporal world of relatively discrete, medium-sized objects. The fact that I might, in my full cognitive capacity, resort to interpretation to justify the assumption shows nothing about how I ordinarily comprehend speech, just as facts about how philosophers might attempt to defend claims about the existence of an external world shows nothing about how our visual systems operates.

Thus, there is no apparent basis for the claim that interpretation methodology will yield a complete semantics. What of the claim that it will yield an autonomous portion of semantics? The problem with this claim is simply that in all other areas of investigation, we (speaking in the western, pro-scientific voice) as a rule decide conflicts between (well-established) scientific theories and ordinary, pre-scientific knowledge in favor of science, even if we remain disposed to the ordinary views in daily life. Why should we do otherwise in the case of semantics?

Suppose, for instance that we discover that some way of representing the world is innate, and typically underlies ordinary understanding. E.g. suppose we discover that we have an innate Euclidean representation of space. The defender of interpretation autonomy must claim that such a discovery would be **wholly irrelevant** to how we interpret other speakers. But

this is completely implausible. Surely this would have some impact on how we interpret speakers' spatial terms. E.g., we might maintain that unless we can uncover evidence that the speaker has acquired, or had the opportunity (e.g. training) to acquire a deviant set of spatial concepts, a Euclidean interpretation of his spatial terms is correct--even if this makes him appear rather irrational. But this is to say that interpretation methodology, which holds that rationality is to be maintained in interpretation above all else, would be (at least slightly) reformed in the case of such a discovery. Thus, it is readily conceivable that a scientific semantics could yield results which were in conflict with those of interpretation methodology. So it does not seem that the results of interpretation methodology are autonomous.

Could we at least maintain that something like interpretation theory, i.e. a theory of what interpretation methodology attempts to explain, must be a required part of any semantic theory? Again the answer appears negative. This is simply because, by anyone's estimates, cognition is a vastly complex place. And interpretations of utterances appear to rest on various and sundry facts about cognition--as Davidson and others have pointed out, anything you believe might be relevant to what interpretation a given utterance is assigned. The conclusion to draw from this is not that there is no hope of developing a scientific semantics, but rather that we should seek a semantic theory which idealizes away

from as many background facts as possible. I.e. instead of seeking a performance theory which explains behaviors that are the complex outcomes of diverse data bases and processes, we should seek a competence theory (or theories) of the knowledge that lies behind such behaviors.

This strategy has proven enormously successful for the syntax of language. Is there any basis for thinking that it will succeed for semantics as well? The following seems like a reasonable *prima facie* case. It appears that when we interpret speech, we make use of knowledge that is distinct from general considerations about the speaker's psychology. First, much of what we read we find meaningful without assigning any intentions to the author. When we read impersonal accounts, such as textbooks or journalistic reports, we often do not conceive of them as authored by anyone--but this does not affect their meaningfulness for us at all. In the extreme, it is commonplace nowadays to encounter sentences that have been produced by computer programs. We find them no less semantically comprehensible than conversational utterances. Yet we do not, in most cases, think of them as utterances of anyone. There is no interpretation of any sort in such cases--we do not assign propositional attitudes to anyone in deciding what is meant.

Second, when we do adjust our assignment of the meaning of another's utterance based on their apparent attitudes and intentions, we virtually never do so because the utterance

sounds like bare noise, the way an utterance in a language that we do not speak often sounds. Rather, we hear the former sort of utterances as meaningful, and then readjust accordingly, based on various other sorts of information. For instance, we may decide that the speaker is joking or engaging in metaphor or inaccurately expressing her beliefs. Or we may hear an utterance as ambiguous--we may hear it as having two or more possible meanings--and attempt to decide on one on the basis of the context. Such cases strongly suggest that there are at least two distinct types of processes and outcomes here. It seems a reasonable hypothesis that semantic comprehension is a largely non-conscious process that typically results in the assignment of a semantic representation to the utterance. On the other hand, there appears to be a mostly conscious process of attempting to assign rational intentions to the speaker. The latter appears to make use of the former's output. And this, in turn, suggests that we might do well to seek a theory of meaning that idealizes away from the latter knowledge and procedures and seeks to isolate the knowledge which appears to underlie the former procedures. Which is to say that the goal of providing interpretations of utterances with a theory of meaning may not be an appropriate one.

I therefore conclude that there is no basis for the view that an interpretation-based semantics will yield a theory of meaning that is independent from a scientific semantics.

3. Conventionalism

Recall that the general question is whether or not semantics is possible without recourse to scientific, and in particular, psychological theories. I have been contending that this is a reasonable view only if speakers have largely explicit knowledge of meaning, and, I claim, we do not. Perhaps, though, the explicitness of knowledge of meaning may appear to be already given by the idea that the selection of meanings is conventional. The reasoning might go roughly as follows:

> To follow a convention, you need to be aware, in a potentially explicit way, of what the convention is. For instance, in Lewis's analysis of a convention, the final condition is that conformity to the regularity and the mutual interest in this conformity "are matters of **common** (or **mutual**) **knowledge**: they are known to everyone" or at least this is "knowledge that would be available to everyone if one bothered to think hard enough."[26] What is conventional in (our) languages is the relationship between words and the world. E.g., 'red' could stand for any object or property you like. That it stands for redness is a matter of conventionally associating this term with the property, and to do this, in a way that we can agree on, we need to be (potentially) explicitly aware of the word-property association, which is to say that we need (potentially) explicit knowledge of the meaning.

The fault with this line of argument is that it overlooks the possibility of substantial mental states which do the work as far as the meanings of natural language terms are concerned

26. Lewis (1983), p. 165.

and are explicitly associated with elements of language, but are not themselves explicitly known. Assume that the term *t* is conventionally associated with property P. One possible way to achieve this association is the following: If a given speaker already has mental states that represent P, then all that is required is that she associate *t* with those mental states. Specifically, the speaker does not need any explicit knowledge of the nature of the mental states in question in order to achieve the association; all that is necessary is knowledge of when the appropriate states are occurring, i.e. knowledge that they are occurring.

Let us consider an example. Take any common sense natural kind concept, such as the concept "cat." None of us know how it is that we recognize cats, nor do we appear to know much else about how we represent cats. Nor do all speakers have much (completely) common knowledge about the property of "cathood." What we do know is when we have recognized a cat. This, I suggest, is the sort of knowledge appropriate to conventionally associating 'cat' with cats. To put it a slightly different way, we need to know when we are having "cat ideas" so we can associate them with the term 'cat' (if conventionalism is correct), but we don't need explicit knowledge of "cat ideas" themselves to do this. And if "cat ideas" are meanings, as I have been (more or less) arguing throughout, this is to say that we do not need explicit knowledge of meanings in order to achieve conventionalistic

associations. Thus, it seems that there is actually no basis in the conventionalistic picture for claiming that we must have explicit knowledge of meanings, and thus no basis for thinking that a theory of meaning can be constructed independent of empirical psychology.

4. Dummett on Meaning

Another influential account of the nature of meaning and semantics comes from Dummett, who, while not recommending a specific methodology for semantics, has had a lot to say about the nature of a theory of meaning itself. We have already considered one of Dummett's arguments against psychologistic semantics above. I will examine two other aspects of his views which bear on the idea of a non-scientific basis for a theory of meaning, namely, verificationism and his rejection of the idea of non-conscious knowledge of meaning.

First, consider Dummett's verificationistic view of truth and truth conditions. He recommends a theory of meaning on which a statement's meaning is given in terms of conditions of warranted assertability:

> For any [decidable] sentence, we may say that the speaker's knowledge of the condition for it to be true consists in his mastery of the procedure for deciding it, that is, his ability, under suitable prompting, to carry out the procedure and display, at the end of it, his recognition that the condition does or does not, obtain.[27]

27. Dummett (1976), p. 81.

And this, together with the following:

> The sense of a statement is determined by knowing in what circumstances it is true and in what false.[28]

yields the claim that the meaning of a statement is determined by knowledge of an effective decision procedure for the statement's truth. Does this support the claim that empirical psychology is not required for a theory of meaning? It might be taken as doing so if we read Dummett as suggesting that the relevant decision procedures are potentially explicit--that is, by sufficient self-examination, or "coaxing" of others, it might be suggested, we can bring the relevant decision procedures out into the open for examination, so that all relevant knowledge for a theory of meaning is already possessed by competent speakers, thus ruling out the need for recourse to a psychological theory.[29]

However, the conception of meaning as justification conditions just sketched seems wildly implausible, if we also suppose that the relevant decision procedure must be possessed by competent speakers and available upon a little coaxing or reflection. First, the sort of justification or assertability conditions that we know explicitly, or can give upon a little

28. Dummett (1978), p. 8.

29. I don't really think that this is Dummett's view, I read him as holding that while the outcome must be explicit, the process itself need not be explicitly accessible and frequently isn't. But even this view is dubious in light of the cases I suggest below--most notably, knowledge of the meaning of scientific terms.

reflection, are typically largely or completely irrelevant to a statement's meaning. Thus, an average person justifies many of their beliefs by appeal to authority. What do most people take as warranting assertions such as "2+2=4", "objects are made of atoms", "there are angels" or "the free market is the best possible economic system"? Typically, the answer is the authority of teachers, religious and political leaders. But clearly this is irrelevant to the statements' meanings.

Further, it often seems that we understand statements with little if any explicit knowledge of what would justify them. Thus, consider the case of scientific theories. Often, a theory is proposed for which its observable consequences are not immediately obvious. It takes years in some cases to deduce specific observational tests which play a significant or even decisive role in the confirmation or rejection of the theory. And indeed, it is not clear that there is ever any effective decision procedure for the confirmation of high-level scientific theories. However, on the more or less brute verificationist view we are considering, until such conditions are determined, scientists cannot be said to know the meaning of the sentences expressing the theory. But this is completely implausible. Moreover, from time to time, new empirical tests are deduced from longstanding theories, e.g. the theory of general relativity. Again, it seems completely implausible to claim that such deductions alter the knowledge of meaning of statements expressing such theories.

Or consider moral statements. It is extremely dubious to claim that anyone who knows the meaning of "abortion is permissible" has an explicit decision procedure for determining the truth of this claim. While virtually all English speakers appear to know the meaning of this statement, some are deeply convinced it is true, others are equally certain that it is false and many others remain uncertain. Nor is there any agreement on how the issue is to be settled.

Even in what constitutes the paradigm case for Dummett's verificationism, namely mathematics, knowledge of meaning as explicit knowledge of warranted assertability seems deeply counter-intuitive. Now, it may be (though this seems questionable) that when most mathematicians comprehend a mathematical expression, they have a good idea of what would constitute a proof or disproof of the statement. But it seems perfectly obvious that most other people can comprehend statements in logic or mathematics while having no idea at all how to go about proving or disproving them. For instance, one of the reasons that Gödel's proof of his first incompleteness theory is regarded as brilliant is that one can get a very good idea of the meaning and consequences of the theorem without ever conceiving of the proof's key notion of Gödel numbering. And, indeed, this is one way to teach this and related theorems--first teach the students the relevant formal language, then explain the meaning of the theorem and its significance, and finally teach them the technique of the

proof. In such cases, it seems clear that people are often in a state of understanding the relevant theorem without having the slightest idea how to prove it (or even if it is provable.)

For such reasons, there seems little point in attempting to identify meaning with explicit knowledge or potentially explicit knowledge of conditions of warranted assertability. Or, at least, it seems clear that the explicit and potentially explicit knowledge we have of decision procedures for sentences is not plausibly viewed as the sole material needed for constructing a theory of meaning (if it is needed at all.) On the other hand, it appears that we have non-explicit knowledge of verification conditions, at least for certain concepts. For instance, the verificationist might claim that our knowledge of the meanings of 'dog,' 'cup,' 'green' and the like is to be identified with our means of recognition for such types or attributes. And here, there is really no reason to suppose that there is no role for empirical psychology to play. On the contrary, introspection and reflection tell us very little about how we recognize dogs, cups, or greenness. So if the possession of such recognitional abilities is to be identified as the possession of knowledge of the meanings of these terms, then the study of this meaning would seem obviously linked to the ways we are seeking to study such abilities, i.e. psychological theories of recognition and perception. Which is to say that if verificationism is to be

plausible it is best viewed as a doctrine that is consistent with a naturalized semantics.

Let us turn to a second aspect of Dummett's views. Perhaps in virtue of some of the facts such as those just cited, he does not claim that we have explicit knowledge of meaning, but rather that such knowledge is **implicit**. Now, if implicit knowledge is understood as I have characterized it above, i.e. in terms of the possession (or "instantiation") of representations and processes which accomplish what explicit knowledge of meaning would accomplish,[30] where such representations and processes are not consciously accessible, there is no reason to think that such implicit knowledge would allow for anything but empirical (psychological) study.

However, an important aspect of Dummett's conception of implicit knowledge is that its attribution requires an associated set of observable behaviors:

> An individual cannot communicate what he cannot be observed to communicate: if one individual associated with a...symbol or formula some mental content, where the association did not lie in the use he made of the symbol or formula, then he could not convey that content by means of the symbol or formula, for his audience would be unaware of the association and would have no means of becoming aware of it.
>
> ...Implicit knowledge cannot, however, be meaningfully ascribed to someone unless it is possible to say in what the manifestation of that knowledge consists: there must be an observable difference between the behavior or capacities of

30. See Kirkham (1980), pp. 212-3 for this "as if" interpretation of Dummett's use of the phrase "implicit knowledge."

> someone who is said to have that knowledge and someone who is said to lack it.[31]

This might be read as supporting the claim that a theory of meaning can be developed without recourse to empirical psychology, because all relevant data is already available in terms of the observable behaviors that form the basis for the ascription of implicit knowledge of meaning.

But Dummett's claim that content can only be conveyed by use seems plainly false. If a given speaker non-observably associates a meaning with a term and her audience likewise associates the same meaning, then, it would seem, all is well for effective communication. No explicit "use," or observable signs of the meaning are required. Suppose, for instance, that speakers have innate, or typically acquire, a concept *C*, and suppose also that this becomes associated with a given term, *t*, by a few triggering experiences, some training, or whatever. And suppose further that no behavioral signs of the acquired association are ever present. Now, the individuals in question may be fortunate, in the sense that they all live in a community which has such associations, so that when they speak to one another and use *t* they interpret it as meaning *C* and are correct. So they communicate successfully, despite the fact that their use of *t* would fail to indicate the association of *C* to anyone who had not had the appropriate experiences or training. This, I suggest, is not only

31. Dummett (1978), pp. 216-7.

intelligible, it may correspond approximate to the state we are all in at least with regard to most of our "common sense" notions which were attained in early childhood. Thus, it is false that "an observer cannot communicate what he cannot be observed to communicate."

What seems to underlie this argument of Dummett's is the view that to have successful communication one must not only achieve the desired (i.e. shared) mental states with one's audience, but one must also be in a position to fully justify the attribution of those states. For instance, in discussing the view that knowledge of meaning might be unconscious, he writes:

> meaning becomes private and hence no longer in principle communicable. This is to say that faith is required if we are to believe that we communicate with one another. The hearer must presuppose that he is interpreting the speaker as the speaker intends: but the speaker's intention and the hearer's interpretation are, at best, constituted by inner states of each respectively, not accessible to themselves, let alone to the other.
> ...If communication is not to rest on faith, it is necessary to maintain that any misunderstanding can come to light.[32]

But why shouldn't communication rest partly on faith, or rather, on **luck**? Two speakers communicate when it is their good fortune to share tacit or implicit knowledge of rules of language including (e.g.) appropriate syntax-mental state associations. When the relatively degenerate evidence is evaluated, we might suppose, the speakers come to believe that

32. Dummett (1989), p. 202.

they are communicating, although, we might suppose further, their evidence is insufficient to rule out various alternative mental state hypotheses and they have no in-principle means of ruling out such hypotheses. What is counter-intuitive about this? Certainly, this is the position most speakers are in with regard to many claims, e.g. that inductions are acceptable, or that the causal principles they apply to the world are correct. The former might be established by some (relatively) obscure philosophical argument, and the latter by the theories of physics. But ordinary people, lacking substantial training in philosophy or physics, can provide no real justification for such beliefs. But, so what? Skepticism is always possible. What we want to know is not if they have a suitable justification, for they certainly do not, but rather if there is some justification for their beliefs, even if it is not accessible to them. And the cognitivist who holds that knowledge of meaning is not explicit may maintain this, in that there are some observable consequences of a scientific theory of meaning, or at least of the conjunction of such a theory with various other psychological theories. But such consequences need not be immediately obvious or tied to the isolated ascriptions--they may be linked in quite complicated ways to the entire cognitive theory and take years to determine. So they need not be accessible in any reasonable way at all for the ordinary communicator.

Note that actual verification of communication is typically very unprincipled. Outside of academics, we rarely, if ever, test one another's semantic competence in any serious way. Rather, as long as we are able to interpret one another's speech in ways that we find rational, we assume that we have communicated. Perhaps many such assumptions are false. But, then again, there is nothing bizarre about allowing that science may show our ordinary beliefs and property attributions about a given domain to be either largely false or systematically mistaken.

Of course, there are times when we fail initially to communicate and are able to achieve successful communication after some negotiations about meaning, but it is important to see that this need provide no support for the claim that knowledge of meaning must be in principle publicly manifest. One ordinary means of fixing a common meaning is by ostension--if I'm not sure that you mean baseball by 'baseball,' I might show you one and label it. What is going on here might be described as follows. I am attempting to activate an appropriate concept in you, the concept BASEBALL, and thus produce an appropriate association, if I'm lucky. But doing so does nothing to demonstrate that this concept, this knowledge of meaning, must itself be behaviorally manifestable. As noted above in discussion of conventionalism, all that is going on in such cases is that we can make good guesses about when a given concept has been activated--we do not have the concept

(or mental state) available for conscious, explicit examination, and we can't provide certain evidence that it has been activated in others.

Thus, it seems that Dummett's insistence that we must be able to fully justify our beliefs in the success of ordinary communication is unfounded, and therefore, there is no basis for the claim that there must be in principle manifestations of knowledge of meaning, if these are understood to be manifestations accessible to ordinary speakers, as opposed to manifestations accessible to (ideal) theoretical psychology.[33]

I will close this discussion of Dummett's views by echoing a point of Chomsky's concerning the issues just discussed. Dummett is guided by the motto that meaning is use. But, as Chomsky has pointed out, use is not to be equated with observable behavior. How one uses a symbol (or an object in general) is partly a matter of what intentions lie behind behaviors. To take a simple example, my uttering 'cat' only in circumstances appropriate for ostending cats (let us generously imagine that we can define such situations) does not show that I am using 'cat' to mean cat. I might be using it to mean pet and doing a poor job of it, or perhaps I believe that only cats are pets. Without some basis of determining my intentions and beliefs, observable behavior

33. Nor is it even evident that ideal science must rest entirely on observable data. After all, it is generally agreed that scientific theories are confirmed by factors other than observable data.

does not determine use. And, as several decades of criticisms of logical behaviorism have made clear, it is false that isolated attitudes must have necessary behavioral manifestations. (E.g. what are the necessary behavioral manifestations of believing that baseballs exist?) So the claim that meaning must be manifest in use need not lead to any sort of claim about meaning being manifest in observable behaviors.

Thus, to the extent that Dummett s views concerning verificationism and meaning as use are plausible, they offer no basis for the assertion that a theory of meaning is anything other than an enterprize of scientific psychology.

5. A Priori Knowledge

Perhaps the most obvious competitor to the view of naturalized semantics is the view that knowledge of meaning is *a priori*, and that such knowledge forms the basis for an *a priori* theory of meaning. However, few philosophers have actually explicitly endorsed either the view that ordinary knowledge of meaning is *a priori*, or the claim that a given semantic methodology is *a priori*, perhaps owing to the fact that the notion of the *a priori* has generally fallen into disrepute, or at least disuse in recent times. For instance, there is no well-received view of what *a priori* knowledge consists in.

One semantic theorist who does explicitly champion an *a priori* approach to semantics is Katz. On his account, meanings are abstract, platonic objects, and we know of them via a faculty of intuition which enables consciousness experience of platonic objects (in the form of intuitions.) Specifically, this faculty is viewed as giving us *a priori* knowledge of meaning. And, if we do have *a priori* knowledge of meaning, then it would seem that we have a basis for constructing a semantics independently of empirical psychological inquiry, i.e. by drawing on this *a priori* knowledge.[34]

Now, we should first note that an endorsement of a platonic ontology does not appear to have any direct bearing on the question of whether or not the study of meanings is part of empirical psychology. To see this, consider that if the general platonistic view is correct, then all properties are platonic objects. But it does not follow that the study of such objects is not a matter of empirical science, for after all, all scientists investigate properties of one sort or another. Thus, the cognitivist conception of semantics appears to face no problems from platonism that would not also apply to the rest of the sciences were this view to turn out to be correct. It is only the view that there is a faculty of intuition which gives us *a priori* knowledge of meanings, along

34. See Katz (1981), especially pp. 202 ff.

with Katz's proposed methodology, which directly bears on the issue.[35]

As noted, Katz' methodology rests on appeal to native intuitions about meaning--i.e. the semantic properties of words and sentences. His main theoretical construct is the semantic marker, which is a decomposition of a concept into a structured (tree-like) set of semantic primitives. Using intuitions about decomposition, semantic markers can be developed to represent the (apparent) underlying structure of meanings. This structure is then used to explain other semantic intuitions concerning synonymy, analyticity, analytic entailment, ambiguity, subordination, etc. For instance, the fact that:

Anyone who strolls walks

is (apparently) analytic can be explained via the fact that the semantic marker for "walks" is part of the semantic marker for "strolls."[36]

35. As far as I can see, one could adopt a naturalized approach, specifically, a competence theory approach to semantics, and either maintain a platonistic view of meanings and language, a conceptualist view such a Chomsky's, or some other ontological view. This is not to say that there would be no deciding between these ontological views in the long run. Indeed, it does seem to me that the conceptualist view--that languages are, roughly, abstractions and idealizations out of psychological states--is the most plausible given this approach. But there is certainly room for debate here.

36. See Katz (1987) for an overview and defense of this methodology.

Applying the cases I suggested above, we can ask whether there are reasons for thinking that this methodology could yield a complete theory of meaning independent of a scientific semantics, whether it will produce results whose validity is autonomous from any results of scientific semantics, and whether the sort of explanation this methodology seeks to provide will be required by any theory of meaning. The idea of *a priori* knowledge would seem to be relevant only to the claim of autonomy, but I shall briefly consider the other two as well.

It is fairly apparent that semantic marker theory will not produce a complete semantics, since it does not appear to say anything about reference, but this is something that it is plausible to think a complete semantics will explain.[37] For instance, suppose that "red" is a primitive element in "Markerese." This is to say that it does not decompose into any another sub-meanings. But it seems that one might ask what the extension of the term 'red' consists in. E.g. Is there a property, physical or otherwise that all red things share? Is the extension absolute, or are there fuzzy boundary cases? Marker theory provides no answers to these type of questions. Yet, there appear to be answers to these questions (tentatively, "no", and "fuzzy") and it seems that the pursuit

37. See Lewis (1983) p. 190.

of these answers might reasonably be regarded as part of the explanation of the meaning of the term 'red.'

Should the results of Katz' approach be regarded as valid no matter what any other sort of semantic enterprise, in particular a scientific semantics, produces? While Katz claims that the intuitions his approach is based on are *a priori* knowledge of meanings, he has not provided a means of establishing theoretical semantic claims *a priori*, in the way proof establishes mathematical or logical theorems. For instance, suppose that a psychological semantics is developed which takes the same intuitions that Katz uses as a basis for producing a theory of competence, and this in turn leads to a theory of performance. But suppose, as seems likely, that the competence theory will make some claims that conflict with marker theory, and that both the competence and performance theories will cause us to reject or revise some of out intuitions. In such a case, it seems, we would regard the scientific semantics as providing us with the real meanings of our terms, whereas marker theory merely shows us the meanings (and semantic structure) we pre-theoretically think that our terms have.

I take it that Katz would dispute the idea of throwing out or revising any of our native semantic intuitions, on the grounds that they are the result of a faculty of *a priori* intuition. But what is to provide the warrant for the judgments of a faculty of intuition? Thus, suppose we

determine that we have a faculty that spits out claims such as "to kill is to cause to die." How are we to tell that this is a faculty of *a priori* intuition? After all, it might also be a faculty that provides theoretical guesses about the meanings of our terms. Perhaps feelings of certainty accompany such claims, but feelings of certainty accompany many other clearly empirical claims that we make as well, e.g. "the sun rose yesterday." The only tenable response here would seem to be that a true faculty of *a priori* intuition, as compared to an "imitator," is actually in contact with meanings, e.g. platonic entities. But we have no idea of what this means, and we certainly have no idea if we have such a faculty or not. Thus, it seems that the postulation of a faculty of intuition is insufficient to explain or justify the *a priori* status of the statements which result from its operation. And this is to say that there is no basis for the claim that such intuitions are *a priori* true, and thus no basis for thinking that the intuitions which marker theory relies on could not be revised or rejected by a scientific semantics, which is to say that marker theory is not autonomous in relation to a scientific semantics.

Finally, will the sort of explanations that marker theory seeks to provide be required by any theory of meaning? The obvious answer is no, since it is conceivable that semantics could concern only reference and truth conditions. However, it seems plausible to think that something like marker theory

will form the first stage of a competence theory for a semantics naturalized to cognitive psychology. Katz, in fact, once conceived of marker theory in this way. He notes that "very little changes in the formal theory of semantics" when the theory is reinterpreted as being a explanation of *a priori* knowledge of platonic objects.[38] This also suggests that little would change for someone who wanted to reinterpret it as a competence theory, a move which the present considerations support.

6. Model Theory

There is a substantial amount of work devoted to model theoretic semantics, in the tradition of Carnap, with Montague Grammar being the most well-developed version of this approach. The methodology consists in interpreting a fragment of natural language into an intensional logic, which is in turn given an interpretation in terms of possible worlds. In effect, the approach characterizes the meanings of natural language expressions as consisting of functions from expressions and contexts to sets of possible worlds.

Does the model-theoretic approach yield a methodology that promises a semantics that is independent from empirical psychological investigation? Montague apparently thought so--the view that is usually attributed to him is that model-

38. Katz (1987), p. 173.

theoretic semantics is not in any sense a psychological competence theory, but is rather a branch of mathematics.[39] However, this outlook appears to run into problems that have been noted in our previous considerations. While the study of intensional logics themselves can reasonably be regarded as a completely non-psychological, mathematical investigation, when it comes to interpreting bits of natural language in terms of logics interpreted in terms of possible worlds, it seems that we need information about how our mental states relate our words to the world, and here psychology comes into the picture. Specifically, the questions of **which** intensional logics, if any, correctly interpret natural languages, and what the correct interpretation of these logics should be are questions which appear to require recourse to facts about our knowledge of meaning, including our tacit and implicit knowledge of meaning. For instance, we might ask if a first-order logic will suffice for natural languages, or if a higher-order logic is required. Or, do the predicates of natural language terms have exact or "fuzzy" extensions? Does our use of definite descriptions imply existential commitment, as Russell thought, or does some other logical form underlie these expressions?

To answer such questions--which model-theorists typically claim or presume to have answers to--we must move from the

39. See Thomason (1974), p. 2.

realm of the purely mathematical and consult our knowledge of meaning. In providing interpretations of fragments of natural languages, model-theorists appear to consult native intuitions about truth-conditions, semantic well-formedness and the like.[40] Perhaps it might be claimed that this is different from a full-blown psychological investigation of our semantic competence. However, such a position runs afoul of two issues that have been noted above, the incompleteness of explicit knowledge of reference and truth conditions, and the non *a priori* nature of our intuitions. First, we have seen that it appears that there are many questions about the truth conditions of our sentences that cannot be answered with our (potentially) explicit knowledge of meaning. It would seem that the only way to answer these questions is to move from complete reliance on intuitions to additional data and information that can be derived from a psychological study of the processes of comprehension, inference, etc. Second, as we have considered in the previous section, there is no apparent basis for the claim that our intuitions are immune to revision. Again, it would seem that the most reasonable construal of an investigation that relies on these intuitions is as an empirical theory that might be revised by further psychological investigation.

40. See Thomason (1974), pp. 51 ff.

It therefore seems that model-theoretic semantics is best viewed as one attempt at developing a semantic competence theory, a theory which might, in principle, be revised or rejected in due course as part of a naturalized semantics. The study of intensional logics themselves might best be viewed as relating to the task of semantics in much the way that mathematics relates to physics. There is much to be done in the development of various models in isolation from empirical inquiry, just as mathematics is typically developed in isolation from questions in the physical sciences. But, all the same, it is a confusion to think that semantics itself can be conducted in insolation from empirical psychology, just as it is a mistake to think that physics can be conducted without empirical investigation.

* * *

This concludes my examination of several of the more influential accounts of meaning and methodological approaches to the study of meaning. In each case, I have argued that the particular conception does not show us that there is any basis for a theory of meaning that does not require recourse to scientific psychology. Rather, it appears that our knowledge of meaning is largely tacit or implicit, i.e. it is knowledge that can only be fully revealed and explicated through the scientific study of cognition.

BIBLIOGRAPHY

Block, N. (1986) "Advertisement for a Semantics for Psychology." In P. French, T. Uehling and H. Wettstein (eds.), *Midwest Studies in Philosophy, Vol. X*, Minneapolis: University of Minnesota Press, pp. 615-678.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1980) *Rules and Representations*. New York: Columbia University Press.

Chomsky, N. (1986) *Knowledge of Language*. New York: Praeger.

Cummins, R. (1983) *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press/Bradford.

Cummins, R. (1989) *Meaning and Mental Representation*. Cambridge MA: MIT Press/Bradford.

Davidson, D. (1984) *Inquiries into Truth and Interpretation*. Oxford University Press.

Devitt, M. and K. Sterelny. (1987) *Language and Reality*. Cambridge, MA: MIT Press/Bradford.

Dummett, M. (1976) "What is a Theory of Meaning (II)?" In Evans and J. McDowell (eds.), *Truth and Meaning: Essays in Semantics*, Oxford: Clarendon Press, 67-137.

Dummett, M. (1978) *Truth and Other Enigmas*. Cambridge MA: Harvard University Press.

Dummett, M. (1989) "Language and Communication." In George, A. (ed.) (1989).

Fodor, J.A. (1975) *The Language of Thought*. New York: Crowell.

Fodor, J. A. (1980) "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology." *Behavior and Brain Sciences* 3, pp. 63-110. Reprinted in *Representations*, Cambridge, MA: MIT Press/Bradford pp. 225-263.

Fodor, J. A. (forthcoming) *A Theory of Content* Cambridge, MA: MIT Press/Bradford.

George, A. (1989) "How Not to Become Confused About Linguistics." In George, A. (ed.) (1989).

George, A. (ed.) (1989) *Reflections on Chomsky*. Oxford: Basil Blackwell.

Haugeland, J. "Weak Supervenience." *American Philosophical Quarterly*, 19, no. 1, pp. 93-103.

Higginbotham, J. (1989) "Knowledge of Reference." In George, A. (ed.) (1989).

Hornstein, N. (1984) *Logic as Grammar*. Cambridge MA: MIT Press/Bradford.

Jackendoff, R. (1983) *Semantics and Cognition*. Cambridge MA: MIT Press.

Jackendoff, R. (1987) *Consciousness and the Computational Mind*. Cambridge MA: MIT Press/Bradford.

Johnson-Laird, P. (1983) *Mental Models*. Cambridge MA: Harvard University Press.

Katz, J. J. (1981) *Language and Other Abstract Objects*. Totowa, NJ: Rowman and Littlefield.

Katz, J. J. (1987) "Common Sense in Semantics." In E. LePore (ed.), *New Directions in Semantics*, London: Academic Press, pp. 157-233.

Kirkham, R. (1989) "What Dummett Says About Truth and Linguistic Competence." *Mind*, 390, pp. 207-224.

Lewis, D. (1983) *Philosophical Papers, Vol. 1*, Oxford University Press.

Marr, D. (1982) *Vision*. San Francisco: W. H. Freeman.

Peacocke, C. (1989) "When is a Grammar Psychologically Real?" In George, A. (ed.) (1989).

Putnam, H. (1962) "It Ain't Necessarily So." *Journal of Philosophy*, LIX, no. 22, pp. 658-671.

Putnam, H. (1970) "Is Semantics Possible?" In H. Keifer and M. Munitz (eds.), *Language, Belief and Metaphysics*, State University of New York Press. Reprinted in Putnam, *Mind, Language and Reality: Philosophical Papers Vol. 2*, Cambridge University Press (1975), pp. 139-152.

Quine, W. V. O. (1960) *Word and Object*. Cambridge, MA: MIT Press.

Thomason, R. (1974) "Introduction." In R. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague*, New Haven: Yale University Press.

PERCEPTUAL NATIVISM, JUSTIFICATION AND NEO-KANTIANISM

At the heart of Kant's account of knowledge and metaphysics is the claim that the constitution of the mind is what is responsible for the truth of certain substantial general principles:

> If intuition must conform to the constitution of the objects, I do not see how we could know anything of the latter *a priori*; but if the object (as object of the senses) must conform to the constitution of the faculty of intuition, I have no difficulty in conceiving such a possibility...we can know *a priori* of things only what we ourselves put into them.[1]

While it is notoriously difficult to determine what exactly the arguments of the *Critique* are, the following general line is at least suggested.[2] First, establish that certain principles are *a priori*, specifically synthetic *a priori*. Next, show (by way of transcendental deduction) that this implies that the mind's constitution makes such principles true. Finally, show that the mind's making such principles true implies an (transcendental) idealist metaphysics, where the world we perceive and think about is understood as being mind-dependent.

1. Kant (1787), pp. 22-3.

2. See also Kant (1787) pp. 174-5. Nothing in the main body of the essay hangs on this being the correct reading of Kant's deduction.

In this essay I am going to develop a similar position by a much more modest route. I will begin with the empirically supported claim that the mind contributes many abstract concepts to our perceptions, and argue that this together with an acceptable account of justification supports both an ideal verificationist theory of truth and the view that the world is mind-dependent. I shall then conclude by suggesting that this may lead us to suppose that there is a substantial synthetic *a priori* element to our knowledge, although the exact specification of this element may be a quite difficult and elusive matter.

My allusion to Kant should not be misunderstood. I will not be attempting any sort of *a priori* investigation of knowledge or truth--my methodology is consistent with a conception of naturalized epistemology. Thus, while I shall use the term "neo-Kantian" for my position, my methodology will not resemble Kant's *a priori* investigations at all. Yet, as I have just noted, the resulting view bears an obvious resemblance to Kant's position--I believe it is closer to his view than any other current outlook in the analytic tradition. But whether or not the result is in any way Kantian, the investigation and its implications should be of interest to anyone concerned with the nature of justification, truth and metaphysics.

I. Overview

I have said that I will be arguing for a form of metaphysical anti-realism based on perceptual nativism I suspect that most readers will find this a very odd line of investigation to pursue. Why, it might be asked, should innateness, which is a concern of empirical psychology, have any bearing on the philosophical issue of metaphysics?[3] In order to put the reader in the proper frame of mind, I will present an introductory overview of the argument I will develop.

A standard view of our epistemic and metaphysical situation is that we are cognitive beings who live in a mind-independent world and attempt to learn about this world. Specifically, we tend to conceive of ourselves as getting samples of certain aspects of the world in the form of perceptual data, and attempting to produce theories that correspond to the (unobserved and unobservable) world on the basis of these samples. But suppose that this perceptual data depends substantially on innate concepts, and that there is no guarantee that the content of these concepts corresponds to the mind-independent world. This is to say that the data may

3. A possible worry that I shall not address at length is that while nativist doctrines are empirical hypotheses, metaphysical doctrines are *a priori*. In keeping with naturalism, I reject this latter claim, but not on conceptual or *a priori* grounds. Instead, I can only hope to demonstrate the empirical status of metaphysics by showing how empirical results can have a bearing on metaphysical issues.

not be accurate at all--it may not constitute a sample of aspects of the mind-independent world. Then it is no longer clear that what we are doing when we theorize, and, more generally, when we seek justification for our beliefs, is seeking to learn about the mind-independent world. For if enough of the innate perceptual concepts are inaccurate, then since justification is dependent on perceptions, increasing justification will take us no nearer to an accurate depiction of the mind-independent world, and may even lead us away from it. Nor is there any means of determining if our concepts correspond to this mind-independent world, since our methods of investigation rely on the very concepts in question.

Thus, however appealing the original picture, we find that substantial perceptual innateness requires us to be completely skeptical about ever having knowledge of the mind-independent world, and prevents us from being able to view the search for justification as the search for correspondence to the mind-independent world. Rather, our justificatory practices are more accurately viewed as attempts at producing an epistemically idealized development of our perceptions, which, given the innate contribution of the mind to perceptions, can be regarded as a mind-dependent world. This is the world we are seeking knowledge of, and have reasonable hope of gaining knowledge of.

It may also be useful to contrast this form of anti-realism with traditional opposition to realism. The old brand

of skepticism about the existence of external objects was based on the ideational view of the mind: having knowledge of something, e.g. perceiving an external object, was explained in terms of the mind's entertaining (conscious) ideas. The primary knowledge relation was understood as a relation between the self and its conscious states. Knowledge of external objects was understood as a secondary, mediate relation, between the mind and the causes of the (perceptual) ideas. But this view left no grounds for the justification of the inference from phenomenal to external claims. How could assertions about the external world be justified on the basis of the primary knowledge relation? It seemed that we were epistemically trapped within a veil of ideas, unable to determine the nature or even the existence of an external world.[4]

We no longer are subject to such worries, not because we have solved them from within the ideational view, i.e. by providing justification for claims about the external world on the basis of our knowledge of the internal world, but rather because we have abandoned the introspective, ideational view of the mind. We no longer explain perceptual knowledge--e.g. the fixation of perceptual belief--in terms of a presentation of phenomenal ideas to the self. And in giving up this view of

4. See Hume (1739-40), p. 212.

the mind, we happily leave behind the problem of the veil of ideas.

However, as I shall argue, the new picture of perception presents an equally difficult problem for mind-independent ("metaphysical") realism. We find that we are epistemically trapped, not in a phenomenal world, but in a world of innate concepts that are contributed to our perceptions by sub-personal (non-conscious) processes. To support realism, we must show either that these concepts correspond to the mind-independent world or that, whether they correspond or not, we have some hope of moving from perceptions which rely on these concepts to accurate depictions of the mind-independent world. But if we cannot do so, then we must admit that the world we live in, that matters to us, is a mind-dependent world whose nature is partially dependent on these innate concepts.

II. Perception and Nativism

My initial task is to show that current psychological research supports the hypothesis that a large number of abstract perceptual concepts are innate. First, a general word about cognitive psychology is in order. Cognitivism, as I understand it, involves the application of the computer metaphor to human psychology. The mind/brain is conceived as an information processor--a set of representational states and processes that are transformations of representations. Such states and processes need not be conscious. In fact, many of

the representations and processes currently postulated are "sub-personal"--i.e. they are neither conscious nor potentially conscious and they are typically not states that would be attributed reflectively or introspectively (i.e. as part of common sense psychology.)

The innateness of most abstract perceptual concepts, given the representational approach, is supported by three considerations. First, there is a substantial amount of evidence indicating that many abstract perceptual concepts are present within the first few months of life. I shall summarize this evidence shortly. The other support for the innateness of perceptual concepts comes from two versions of the poverty of the stimulus argument.[5] The first concerns the poverty of the input: Cognitive theories explain perceptions as resulting from transitions through informational stages from an initial input to a final perception.[6] When we examine the information

5. Chomsky has used this type of argument as a basis for postulating a substantial innate knowledge of (the formal structure of) language. See Chomsky (1965), pp. 47 ff. and Chomsky (1980), pp. 34 ff. As he notes, this general form of argument dates back to Plato's *Meno*.

6. It is not clear whether we should think of the final stage as a perceptual belief, or as something else, e.g. an input to the belief-system. For one version of the latter view, see Fodor (1984). Actually, I suspect that the answer is a complicated combination of these alternatives--something along the lines that common sense attitude states supervene on but do not always type-reduce to the states that perceptual theory characterizes, so that the outcome of perceptual processes can sometimes be regarded as a belief and sometimes not. In any case, all that matters for the present investigation is that the conceptual content is added to those representations as a result of innate concepts or processes, and that this content somehow turns up in perceptual beliefs

that is regarded as the initial input for perception, we find that it is greatly impoverished relative to the final perception. Thus, explanations must postulate that content is added at various transitional stages.[7] And since--as we shall see in consideration of sample theories--there is no basis for claiming that the input in any way determines or specifies what content is to be added,[8] the only explanation for the addition of the content is that it, along with the processes which add it, are innate. Specifically, examination of the information on the retina, the eardrum, etc. as well as information about the states of the body and sense organs, suggests that the innate, added content must include most of the abstract notions found in perceptions, viz. the notion of an object, and most "primary" spatio-temporal concepts including ideas of surfaces, rigidity, forms, depth, motion and the like.

as a result of these states. With this understood, I shall use the term 'perception' for states at this final stage.

7. There is a long tradition--dating back at least to Berkeley in philosophy and Helmholtz in psychology--of postulating inferences in the production of perceptions. What I suggest is that the important aspect of multi-stage theories is that information is added in the process. I suspect that the question of whether or not this is a genuine case of inference depends on the complicated issue of whether or not the stages in such transitions can be regarded as beliefs or not--see the previous note.

8. Unlike, e.g., most computer programs, whose initial segments typically dictate how later portions or further input is to be processed.

The other form of the poverty of the stimulus argument for perceptual innateness concerns what might be called poverty of training: There appears to be extreme uniformity in how most (physiologically normal) humans perceive the world--we not only all appear to perceive a world of medium-sized spatio-temporal objects, but also agree on most observable qualities of those objects. However, we do this without a substantial amount of training, and very non-uniform training, in perceptual concepts. This strongly suggests that few (abstract) perceptual concepts are acquired through learning, which is to say that the rest are innate.

I will now elaborate on each of these points. First, consider the direct evidence for perceptual innateness. Studies by Spelke and her colleagues indicate that infants aged 3-5 months perceive a world of physical objects. Since infants this young do not have sufficient motor skills to enable testing of perceptual competencies through performance, researchers must instead observe infants' reactions to various stimulus arrays. A methodology frequently utilized involves the measure of looking time after habituation. Infants are shown a display until they cease to look at it. Presentations are repeated until habituation--i.e. until looking time on each trial declines substantially. Then a new display is presented. If the looking time is the same as on the last few trials, then it is assumed that the infant perceives the new display as being the same as the previous display. And if

infant spends a significantly longer time looking at the new display, then it is assumed that the new display is perceived as different.[9]

Using this methodology, 4-month-old infants were habituated to a display of an object, e.g. a rod, whose top and bottom was visible, but whose center was occluded by another object, e.g. a box. The infants were then separately presented with two other stimuli, one display of the non-occluded object, and one display of two objects whose fragments corresponded to the visible portions of the object in the original display. It was found that if the visible portions of the occluded objects were moved together, in any direction, infants looked longer at the fragmented object in the subsequent display, but not at the whole object, indicating that they had originally perceived a complete but only partially visible object. However, when the displays were completely stationary, preferential looking occurred for both subsequent displays. This suggests that infants at this age represent a world of three-dimensional, mobile objects, and use motions as cues to determine objects' boundaries and

9. See Spelke (1985), Kellman and Spelke (1983).

unity, but, unlike adults do not use configurational properties of objects to determine boundaries and unity.[10]

Another set of studies provides evidence that four-month-old infants infer the continued existence of objects when they are hidden from view. The apparatus involved a block with a rotating screen in front of it. The screen, initially lying flat, was rotated upward 90° to hide the block. In one sequence, the screen was rotated until it reached the place where the block had been, stopped (as though it had hit the hidden block) and then was rotated back again revealing the block once more. In the other sequence, the screen was rotated 180° through the place where the block had been, (thus making it appear as though the block has vanished) and then back again re-revealing the block. Measures of looking times revealed that infants, like adults, showed a (statistically) significant greater interest in the second, "impossible" sequence, thus supporting the hypothesis that infants infer the continued existence of objects that become hidden from view.[11]

In general, these and similar studies reveal that 3 to 5-month-old infants perceive a world of cohesive, bounded,

10. See Kellman and Spelke (1983). Note that the fact that infants do not appear to use certain cues or represent certain properties at a given age does not show that such methods or representations are not innate. They may be innate but not programmed to emerge until a later age.

11. Baillargeon, Spelke and Wasserman (1985).

potentially movable three-dimensional objects that have spatio-temporal continuity.[12] And these results are corroborated by abilities that are observed when motor abilities begin to emerge. When reaching behavior emerges at 4-5 months of age, infants reach out for more distant objects, showing an apparent representation of depth, and reach ahead of moving objects (in order to catch them), showing a perception of the position and motion of objects.[13] And when infants begin to crawl, they are (fortunately!) able to successfully avoid a visual cliff, as a famous set of studies showed.[14]

This evidence strongly suggests that perceptual concepts sufficient for representing a three-dimensional world of temporally enduring, movable objects are innate. While only the detection of concepts at birth can conclusively establish innateness, three to five months leaves little time for acquisition. Certainly, no training is received in these concepts nor has any obvious trial and error learning of these concepts been observed in the first few months of life.

The second part of the case for perceptual nativism is the argument that, given the apparent poverty of the informational input, cognitive approaches to perception

12. See Spelke (1987), (1990) for summary and discussion of this research.

13. von Hofsten (1986).

14. Gibson and Walk (1960).

require the postulation of substantial innate contributions of content to perception. To illustrate, I will briefly examine Marr's computational explanation of vision.[15] His theoretical framework, which incorporates the results of a number of researchers, provides an explanation of how the visual system produces representations of the shape and spatial arrangements of perceived objects from the retinal image. The theory postulates various representational stages, each of which is computed from the information available at the previous stage. There are four primary stages: the first is the image, the information available immediately from the retina, which consists of a two dimensional array of intensity values for each point in the array--a "gray array." The next stage is the primal sketch, which is a representation of geometrical information about the image, i.e. where various sorts of patterns can be found on the image, including zero crossings, blobs, edge segments, virtual lines and boundaries. The following stage is the 2½ dimensional sketch that represents visible surfaces from a viewer-centered perspective. Represented features include the local surface orientation, relative distance, depth and surface orientation discontinuities of objects. The final stage is the 3 dimensional model representation that consists of

15. Marr (1982). Page numbers in the text in this section refer to this work.

hierarchically arranged models of the shapes and spatial arrangements of objects.

The two points I wish to make about this theory are that content is added in the transitions between stages and that this content is innate. Marr does not explicitly endorse either claim--nor does he discuss either one--but as I shall now show, the theory readily supports both assertions.

Consider first the point that content is added in processing. This is most easy to see from the fact that not all information in a given stage is present in the initial stages. Marr typically characterizes this content enrichment in terms of the assumptions the theorist must make about the physical world:

> the [structure of surfaces] is strictly underdetermined from the information in images alone, and the secret [for the theorist] of formulating the process accurately lies in discovering precisely what additional information can safely be assumed about the world that provides powerful enough constrains for the process to run.(p. 266)

These assumptions include the rigidity of surfaces, spatial coincidence, and that there is a uniform light source. But, the later stages have content that represents the world--as Marr puts it:

> ...the true heart of visual perception is the inference from the structure of an image about the structure of the real world outside. The theory of vision is exactly the theory of how to do this, and its central concern is with the physical constraints and assumptions that make this inference possible. (p. 68)

So the theorists' assumptions become equivalent to content being added in processing--in effect the visual system gets characterized as making those assumptions.[16] However, we should probably not think of such assumptions as being explicitly represented in visual processes--the content is typically added in virtue of certain input as being treated as an indication of certain external states.

The second point I wish to make here is that the added content is not present because it is determined or dictated by experience/prior input. Rather, such content is added as a result of innate features of the perceptual system. Put another way, we don't choose to perceive things the way we do, we're designed to see them that way. The visual system does not produce visual descriptions of objects from the gray array input because it, or some other system, has determined that there really are objects. It's just built to produce these sorts of descriptions. While Marr's theory does not require that the computations in question are innate, it is difficult to see how we could acquire processes and computations of such unbelievable complexity, particularly given that we have absolutely no explicit training in such matters.

16. Compare Chomsky, *op. cit.*, who equates the theorist's assumptions about a universal grammar with the innate knowledge of language that the child possesses.
As noted above, I do not mean to apply that the perceptual system's "assumptions" are to be equated with ordinary propositional attitudes, particularly not in terms of functional role.

To illustrate these two points, consider the first stage of stereopsis. Stereopsis is the process of comparing the information from images from the two eyes to determine the disparity between the two representations of an object, in order to determine, among other things, the depth of the objects in the visual field. The initial task is to measure disparity by selecting a particular location on one image, comparing it with the same location on the other image, and measure the discrepancy. The problem, however, is to determine what is to count as the same location on each image--i.e. which portions of a given image pairs can reasonably be assumed to represent the same object? Marr cites two apparent facts ("physical constraints") about the visual world: "1) a given point on a surface has a unique position in space at any one time and 2) matter is cohesive, it is separated into objects, and the surfaces of objects are generally smooth in the sense that the surface variations...are small compared with the overall distance from the viewer."(p. 113) This leads to three specific rules that it is assumed the stereopsis module follows in determining the areas to compare for disparity: i) black dots can only match black dots, since it is assumed that there can be matches just in case the images have arisen from the same physical situation.
ii) Almost always, a black dot from one image can match no more than one black dot from the other image, by physical constraint (1) and iii) the disparity of the matches varies

smoothly almost everywhere over the image, by physical constraint (2).

It is assumed that if a correspondence is established between the two images in accord with these rules, then that correspondence is "physically correct" (114-5), i.e. that it yields a (partially) veridical representation of the environment. In effect, this is to say that this process assumes that there are external (at a spatial distance) stimuli of shapes and surfaces giving rise to the information in the two images. It is clear that the input, the images themselves (and information about eye movements) hardly dictate the specific matching assumptions. And it is difficult to imagine acquiring such rules and assumptions on a trial and error basis--if this were required, then surely many, perhaps the majority of us would never learn to see.[17] Thus, it appears that if the computational theory of vision is correct, then substantial innate content is added in the course of the derivation of perceptions from retinal inputs.

The third consideration that supports the innateness of many perceptual conceptions, and thus, as we have just seen, of the addition of innate content to perception, given the cognitive approach, is simply that there is virtually no evidence that any of a basic core abstract perceptual concepts

17. Stereopsis appears to emerge in approximately the fourth month of life. Held (1985) suggests that this is the result of cortical maturation.

are learned. Specifically, there is a basic set of perceptual notions that characterize the features of objects existing in space and time, e.g. surface, boundary, shape concepts, texture concepts, solidity, etc.--concepts representing what were traditionally known as the "primary qualities," and perhaps the "secondary qualities" as well.[18] These notions appear to be uniformly present in humans. That is, we do not encounter groups of people that, e.g., fail to conceive of spatially bounded objects, or fail to conceive of the texture of surfaces, or fail to conceive of surfaces. Moreover, there is no evidence to indicate that any of these notions are acquired through trial and error training. Nor is there any uniform instruction throughout the world's diverse social environments that could account for this uniform presence. But this is to say that there is virtually no positive case for the view that some or all of these concepts are acquired.[19]

I do not mean to enter into the much discussed issue from the philosophy of science of the theory-ladenness of observation concepts. Critics of positivist views of science have argued that what someone observes is relative to what

18. I leave it as an open question as to what exactly this set includes.

19. Note that even non-uniform perceptual competence which mirrors non-uniform environments and training is not decisive evidence for acquisition--as Fodor (1981) points out, the relevant concepts could be innate but dormant, waiting to be triggered in appropriate environments.

theoretical concepts that individual has.[20] For instance, scientists observe electro-magnetic fields while Eskimos observe dozens of varieties of snow, but most of us observe neither. While such critics may have shown that there is no theory-neutral notion of observation that will provide an epistemological foundation for the testing of theories, they have not provided any evidence against the view that theory-laden observation concepts are ways of re-categorizing or re-conceiving states that are produced using some basic set of (mostly innate) perceptual concepts.[21]

I conclude that the claim that a substantial amount of innate content is added in perception is quite plausible. This is not, however, to say that it must be true. The theories just presented are just in their infancy, and there is certainly no definitive data concerning the absence of acquisition for perceptual processes (although the lack of training in observation should be fairly evident from ordinary social facts.) Yet, it is certainly a tenable enough account to warrant an investigation of the consequences of this view for philosophical doctrines--as I proceed to do.

20. See Hanson (1961), Kuhn (1962/70).

21. See Fodor (1983) for some development and defense of this point.

III. Realism and Anti-Realism

I will now argue that the existence of a substantial innate contribution to perception can be shown to support a neo-kantian/anti-realist view of knowledge and the world against a metaphysical realist view. In this section I will present these two opposing views and make some introductory remarks about them.

First a caveat. I am not claiming that the position that I will present captures all of the important issues in the various realist/anti-realist disputes. However, I do claim to have found a substantial and important point of contention, one that I believe most of those who have labeled themselves realists would not want to accept. And, moreover, I claim that this issue should be a main consideration when framing a metaphysical and epistemological point of view.

My formulation of the issues owes much to Putnam's recent work,[22] although as I shall discuss below, I believe that I am defending a position that differs substantially from his. Throughout I will also attempt to provide some indication of how the position I favor contrasts with those of other notable anti-realists.

Both positions that I will describe maintain that there is world that exists that is mind-independent and evidence-independent. That is, there is an existence that has the

22. See Putnam (1978), (1980) and (1983).

attributes (or whatever) it has independently of our representations or knowledge of it. The dispute arises over the role this world plays in our knowledge. The position that I shall call **metaphysical realism** maintains that what we are seeking when we are seeking knowledge (whether we actually ever get it or not, by perception or science or any other means) is a correct representation of this completely mind-independent existence. Assuming that a "correct representation" involves, among other things, truth, then metaphysical realism will amount to a correspondence theory of truth:[23]

> *Correspondence theory of truth*: A belief or proposition (etc.) is true if the state of the mind-independent world it represents actually obtains, whether or not anyone could ever have knowledge of its obtaining.

Note that this might be understood as the conjunction of two doctrines--what Putnam has called the non-epistemic theory of truth, namely that a statement's truth is independent of any knowledge that we might have, and second that the world we represent is the mind-independent world. However, I shall consider these two doctrines together until it is time to reject them, since together they form a tenable position about truth and reality. I.e. maintaining the non-epistemic view but

23. When I speak here and throughout of "theories of truth" I mean only partial theories of truth--thus, the theories listed here should be understood as supplementing whatever else will be required of a theory of truth, e.g. the satisfaction of certain formal constraints.

granting mind-dependence is highly problematic, since it would seem that correspondence to a mind-dependent world must in some sense be epistemic. And, as I shall show below, the conjunction of a verificationist view of truth and the claim of mind-independent correspondence is inconsistent given perceptual nativism, for then there is no guarantee that true--i.e. ideally verified--statements will correspond.

Now, the metaphysical realist need not claim that all predicates represent a mind-independent reality. For instance, it could be maintained that while predicates that reflect ordinary perceptual concepts have mind-dependent truth-conditions, predicates of (true) science have mind-independent truth-conditions. Nonetheless, I assume that on the metaphysical realist view, at the very least some key, core set of predicates is thought to represent a mind-independent reality.

What I shall call the **anti-realist**, or **neo-kantian** view also grants the existence of a mind-independent world, but denies that we can have any knowledge of it, beyond acknowledging its existence. Instead, it is claimed that the objects and properties that we perceive and theorize about make up a mind-**dependent** world, one whose ultimate nature is in part determined by (some) of the ways in which we represent the world.

The neo-kantian view can be understood as a combination of two general types of claims, first that we make a

substantial contribution to our knowledge of the world, and second that truth or rightness is relative to our knowledge, conceptual schemes or theories. The second thus amounts to an ideal verificationist, or Peircean theory of truth:

> *Ideal verificationist theory of truth*: A belief or proposition (etc.) is true if would be accepted under ideal evidential conditions.[24]

24. A verificationist theory of truth need not, in and of itself, lead to neo-kantianism. One could maintain verificationism while denying that we make any substantial contribution to our representations of the world. For instance, I believe this is the position of Dummett.

Also note that it is possible to hold an non-ideal verificationist theory of truth--both traditional positivists and, again, Dummett appear to present such a view. While I think that this account is shown to be wildly implausible when we consider our actual practices of justification, I will not argue this point here, so those that accept a non-ideal verificationist view of truth may read my argument as supporting a slightly different form of neo-kantianism from the one I describe below.

Finally, it is also possible to develop an ideal verificationist view in several different ways. One development equates truth with those beliefs that we humans would be left with after actually contemplating all evidence that we are able to get our hands on. On the other hand, truth might be identified with those beliefs that an ideal reasoner would achieve after considering all evidence that is potentially available to beings in our world with our sense faculties. This latter view treats truth as a sort of normative ideal, something that we might never quite reach because of inherent irrationality, failure to pay attention or seek data diligently, lack of storage capacity, or failure tc develop or contemplate the most explanatory theories. Such a view, while still verificationist, allows for something like what the realist wants to maintain--i.e. that there is no guarantee that inquiry must ultimately lead to the truth. I suspect that the former view is closer to the way we think of truth in everyday contexts, while the latter is more fitting for our conception of scientific truth. In any case, the discussion will be neutral between these (or any other) further developments of the ideal verification theory.

The general moral of these theses is that metaphysics must be replaced by epistemology. While we can conceive of the world's existing independently of our particular representations, we also, in acknowledging that we make a contribution to our representations, limit our ability to make claims about the contribution-independent nature of the world. If the contribution is substantial enough, the questions "what exists mind-independently and what is it like?" cannot be answered. Instead, the neo-kantian urges the relativization of metaphysical claims to our representations, so that metaphysical inquiries become inquiries about our best-justified theories. That is, the question "what really exists?" can only be answered by presenting the best (e.g. scientific) theory we have to date. Thus, the neo-kantian holds both (1) a very pragmatic, "internal" realism, which is mind-dependent in that it includes an explicit relativization to the contribution we make to our knowledge and also (2) acknowledges the intelligibility of contribution-independent existence but (3) maintains a complete and total skepticism about the possibility of knowledge about this existence, i.e. complete skepticism about ever knowing that we are right or wrong in any claims we might make about the contribution-independent world.

The nativist view of perception does not in and of itself conflict with the supposed metaphysical reality of represented objects and properties. However, it does naturally lend itself

to the neo-kantian view, since it provides an obvious basis for one of the two crucial neo-kantian claims, namely that we make a substantial contribution to our representations of the world, a contribution that is not determined by the states of the world that cause those representations.[25] The crucial question, however, is whether or not this innate, added content can (potentially) be shown to correspond to a mind-independent world. If so, then the existence of this contribution to our knowledge is consistent with metaphysical realism. If not, then we should instead adopt the neo-kantian view. What I will argue in the next section is that when we consider any of the standard accounts of justification, we find that a substantial innate contribution cannot be shown to correspond to a mind-independent reality. Put another way, what I will be arguing is that the conjunction of the existence of this innate contribution and any standard theory of justification is not compatible with metaphysical realism--the correspondence theory of truth--but is compatible with an ideal verificationist theory of truth. Thus, I will argue that perceptual innateness supports the neo-kantian view.

25. The neo-kantian positional might eventually be substantially bolstered by the additional empirical finding that our theory-forming and belief-fixating processes also involve a substantial innate conceptual contribution. However, I do not think that there is much evidence at present about what such processes are like, let alone about whether they depend on substantial innate conceptual contributions or not.

A. Putnam's Internal Realism

Before turning to this line of argument, it will be useful to briefly consider Putnam's view of these matters. He sees the realist/anti-realist dispute as turning solely on the correspondence vs. verificationist theories of truth. Thus, he would deny that it is reasonable to conceive of a mind-independent reality. And he also gives no indication of thinking that there is any sort of contribution to our knowledge on the part of the mind. On the other hand, he has provided no notable criticisms of either of these additional doctrines.[26]

If Putnam's arguments for an ideal-verificationist theory of truth are successful, then the anti-realist position I have presented above could rest on the plausibility of perceptual nativism together with his conclusion. I.e. given perceptual innateness, it is reasonable to conceive of a world that is

26. See, for instance, Putnam (1989), pp. 221-2 where he accuses Quine of suggesting the idea of a noumenal reality, and rejects it because this sort of Kantianism and metaphysical realism are "made for each other." However, it is one thing to suppose that a mind-independent reality exists, as the basis for our perceptions for instance, and quite another to suppose that it is this mind-independent world that we represent or know. If Putnam has, as he claims, shown metaphysical realism to be incoherent, it does not follow that the idea of a mind-independent world is incoherent in and of itself.

independent of these concepts--knowledge independent. But given truth as ideal verification, we also see that this cannot be the world we represent and claim to sometimes possess knowledge of.[27] However, as I shall now discuss, Putnam's main argument appears to rest on a questionable assumption, one that I shall attempt to avoid in my subsequent defense of an ideal-verificationist theory of truth.

The argument is designed to show that the metaphysical realist idea that truth involves correspondence to a theory-independent reality is, at bottom, incoherent. We are asked to consider an ideally verified theory of the world, one which satisfies all operational constraints on evidence and verification conditions. The metaphysical realist, argues Putnam, must claim that in such a situation, there is still a question of whether or not the theory corresponds to reality. And this is precisely what distinguishes the realist from the anti-realist, who identifies truth with ideal confirmation. However, Putnam argues, there is no way for the realist to make this view of correspondence intelligible:

> I assume that THE WORLD has (or can be broken into) infinitely many pieces...Pick a model M of the same cardinality as THE WORLD. Map the individuals of M one-to-one into pieces of THE WORLD, and use the mapping to define relations of M directly in THE WORLD. The result is a satisfaction relation SAT--a 'correspondence' between the terms of [the language] and set of pieces of THE WORLD--such that the theory T_1 comes out *true*--true of THE WORLD--provided we just interpret 'true' as TRUE(SAT). So what becomes of the

27. I shall enlarge on this claim when I examine each of the standard accounts of justification.

> claim that even the *ideal* theory T1 might *really* be false?[28]

The idea is that such an interpretation meets all operational and theoretical demands on the notion of reference. So, claims Putnam, there is no sense to the view that something more could be required for a theory to be true. Hence, we are left with a(n ideal) verificationist theory of truth.

However, it looks as though the metaphysical realist who is not also a **semantic** verificationist will be unmoved by this line of argument. Why, we might ask, should the distinction between the meaning of 'true' and 'ideally verified' be accessible to us, even ultimately? As Fodor points out, in a slightly different context, we are more than willing to admit that other creatures with concepts, e.g. all other species with concepts, cannot have access to the one true theory[29] (if there is such a thing), which is to say that they cannot have complete knowledge of what their concepts concern, so why should we be exempt from such worries? That is, it does seem intelligible to maintain that there is a difference between "ideal-by-our-lights" and "true," even if we shall never be able to gain knowledge of what this difference consists in.

If we grant an ideal verificationistic semantic principle, to the effect that:

28. Putnam (1978), p. 126.

29. Fodor (1983), pp. 125-6.

> if two terms differ in meaning then their applications must be distinguishable under ideal evidential conditions.

then the argument is shown to be sound, by applying this principle to 'true' and 'ideally verified.' But it is not obvious why the metaphysical realist, or anyone for that matter, should be required to accept this principle. In fact, this would seem to be precisely what the externalist movement in semantics--which is based in part on Putnam's own views[30]--has rejected. If, for instance, you grant that "meaning ain't in the head," it's difficult to see why ideal evidential conditions must reveal differences in meaning.

Thus, Putnam's argument appears to fail.[31] One could attempt to support it by defending a verificationist theory of meaning, but it is extremely difficult to find convincing, let alone decisive reasons in favor of this doctrine. What I shall now argue is that various considerations concerning justification, when joined with the hypothesis of perceptual nativism that we have examined in the previous section, lead us to the ideal verificationist theory of truth--a line of

30. E.g. Putnam (1975).

31. Putnam offers what may be an independent argument (or arguments) against a non-epistemic theory of truth that is based on the indeterminacy of reference--see Putnam (1980), chapter 2. See also chapter 1 and "Models and Reality" in Putnam (1983). However, this line seems to assume a verificationist semantics too. If not, or if semantic verificationism is ultimately tenable, then the case for an ideal verificationist theory of truth is over-determined.

argument that avoids appeal to a verificationist view of meaning.

IV. Justification

In this section I will argue that the doctrine of a substantial innate contribution to perception, when conjoined with any of the standard accounts of justification, implies that we should reject the correspondence theory of truth in favor of the ideal verificationist theory of truth. This will complete my argument for neo-kantianism.

A theory of justification is an account of how it is we go about justifying our beliefs--i.e. deciding what is true. Such an account is traditionally thought of as not just a descriptive view, but as a normative set of rules which legislate justification.[32] In what follows, I will consider in turn foundationalist, coherentist and reliabilist/externalist accounts of justification, arguing that in each case, if the account of justification is to be maintained in the face of perceptual nativism, we must abandon a correspondence theory of truth in favor of an ideal-verificationist theory.

32. Two things that will not matter to the argument below are 1) whether or not a theory of justification is normative or simply descriptive and 2) whether or not a theory of justification will be a substantial criterion for knowledge, e.g. if knowledge is justified true belief. I take it that even if one's account of knowledge does not mention justification at all, we still need a theory of justification, either to describe how we seek truth or to tell how to do so.

While the arguments for each case will vary a bit, it is worth noting the general strategy. I will claim that (increasing) justification must yield or approach truth.[33] Further, an account of justification must rely on some large subset of perceptions to link our beliefs to the world. Now, once we acknowledge a substantial innate conceptual contribution to perceptions, and if we take truth to be a correspondence to a mind-independent world, then we face the possibility that the innate concepts do not correspond at all. Given this, an examination of the role of perceptions in the theory of justification reveals that justification may not yield or even approach truth. Assuming that some account of justification must be correct, the culprit must be the theory of truth. Since on an ideal verificationist theory of truth, we do not face the same possibility of error for our innate perceptual concepts, a theory of justification together with

33. This is not to insist that our beliefs must be justified--i.e. that we must be able to reach the truth. Justification could turn out to be something that we can't actually get for our beliefs. To put the issue a slightly different way, I am not insisting that a theory of justification must defeat the skeptic, although this is something that those offering theories of justification usually are seeking. Thus, it is a mistake to think that in response to my position someone could maintain metaphysical realism while simply allowing that skepticism can never be ruled out. The available move in this regard is to reject the possibility of an account of justification, but this is to embrace complete skepticism about our ordinary justificatory practices, a highly implausible view. I shall discuss this option below, after having considered each of the candidate views of justification.

(substantial) perceptual nativism supports a verificationist rather than a correspondence theory of truth.

A. Foundationalism

A foundationalist account of justification seeks to identify a set of core beliefs--the basic beliefs--which can be regarded as self-justifying.[34] Other beliefs are justified if they are related appropriately to the basic beliefs, e.g. by deduction, induction, etc. Relative to our purposes, two important questions face the foundationalist. One is what degree of self-justification the basic beliefs receive. We can simply note two apparent extremes. On the one hand, it might be claimed that the basic beliefs must carry with them a guarantee of their truth. On the other hand, such beliefs might simply be viewed as carrying some substantial initial likelihood of truth--at least more so than most other beliefs. The other important question concerns what type of beliefs, psychologically speaking, are to be in the set of basic beliefs. Assuming that we must at some point have justification for beliefs about the world, and since we learn about the world through perception, there are two prominent classes that would appear to be candidates for basic beliefs about the world, namely ordinary perceptual beliefs and

34. For a recent version of a foundationalist view of knowledge, and thus justification, since knowledge is viewed as justified true belief, see Chisholm (1980).

phenomenal beliefs. I will begin by evaluating the position of the foundationalist who holds that the basic beliefs include some (large) subset of ordinary perceptual beliefs. Later, I shall consider the possibility of a foundationalism based on phenomenal beliefs.

Now, the obvious difficulty that arises when we conjoin the foundationalism just characterized with the correspondence theory of truth and the fact that there is a substantial innate contribution to our perceptions is that it seems possible that the innate concepts added in perception fail to correspond to the actual (mind-independent) state of the world. Suppose that *C* is some concept that is added in perception, in that some informational transitions in the course of perception are such that the presence of certain information at earlier stages leads to representations which represent the world as being *Cish*. Why should we think that there actually is a property or attribute in the (mind-independent) world that corresponds to *C*? That is, is seems perfectly possible--logically and physically--that any of the innate concepts applied in perception could turn out not to correspond to actual properties of the world. For instance, we could imagine building (assuming the cognitive approach to psychology is basically correct) an information processing device that added incorrect concepts in the course of producing representations from impoverished initial input. And we sometimes consciously mis-apply concepts to the world--e.g.

we believe in unicorns or phlogiston. So it would seem that it is possible that we have been built with inaccurate concepts too.

And what if they fail to correspond? It seems, then, that the foundationalist is in trouble, for the correspondence theory of truth suggests that then perceptual beliefs involving these concepts are false, as they do not correspond to any external properties, any more than do beliefs about the existence of unicorns or phlogiston. Thus, it seems that the possibility of non-correspondence undermines claims about guaranteed truth for perceptual beliefs. And it also appears to undermine a weaker foundationalist view, which holds only that such beliefs are very likely to be true. For, while there may be arguments for the likely truth of perceptual beliefs, such arguments would not appear to satisfy the foundationalist's concept of self-justification. For instance, it might be suggested that the best hypothesis based on our best confirmed theories is that most such concepts correspond, or that evolutionary theory shows us that it is likely that most of our perceptual concepts correspond to external properties, since otherwise nature would not have selected for perceptual mechanisms that make use of such concepts. But such explanations, if successful,[35] would be of no help to the foundationalist. The perceptual foundation is supposed to

35. I will return to these suggestions in relation to a coherence theory of justification in the next section.

provide justification, not receive it from the statements which it supposed to (potentially) justify. And, it should be apparent that such theories or explanations would not themselves be self-justified beliefs, but would instead depend on a host of other beliefs, including countless perceptual beliefs. So rather than serving as the foundation of all knowledge, perceptual beliefs with substantial innate content would seem to themselves be in need of justification, which is to say that the beliefs which the foundations claimed were self-justifying, on whatever grounds, are not self-justifying. So the suggested version of foundationalism, when conjoined with a correspondence theory of truth, will fail in the face of nativism.

On the other hand, if truth is ideal justification, then perceptual nativism need not constitute a threat to the claim that some subset of perceptual beliefs are self-justifying. For these beliefs will either be fully justified, and thus true, or, more plausibly, very likely to be true, i.e. very likely to remain acceptable as more evidence comes in, with exceptions in cases of the detection of error and the development of scientific theories that undermine the beliefs.

The foundationalist who wants to cling to metaphysical realism might seek to identify something other than ordinary perceptual beliefs as those that form the basic self-justifying set. The only obvious candidate is the set of phenomenal beliefs. But this is not much of an option.

Phenomenalist foundationalism is the traditional cornerstone of idealist or skeptical views, which reject belief in a mind-independent world. Thus, there is no apparent means of justifying beliefs about the non-phenomenal world on the basis of phenomenal beliefs. Moreover, perceptual nativism would seem to undermine any alleged connection between phenomenal appearances and external attributes. E.g., the fact that it appears that there is an object in front of me, and no other appearance conflicts with this does not support the claim that there really is a mind-independent object there unless the (apparently) innate notion of an object actually corresponds to mind-independent objects. Since it appears that phenomenal beliefs could not themselves justify a belief in this correspondence, a phenomenalistic foundationalism will not support metaphysical realism.

Nor is a phenomenalistic foundationalism particularly plausible in and of itself. A theory of justification, it would seem, must at least provide some hope of showing how to justify many of those beliefs that we think are justified, pre-theoretically. For instance, it is not clear how we could function on a daily basis if we did not actually accept countless beliefs about the external environment. A foundational theory of empirical knowledge based solely on phenomenal states would not appear to justify sufficient beliefs to enable this. I.e. attempts to reduce either ordinary perceptual beliefs or scientific beliefs to

phenomenal beliefs have ended in failure. Nor is it particularly plausible to think that ordinary and scientific non-phenomenal beliefs can be induced or otherwise epistemically based on phenomenal beliefs. Thus, we rarely cite phenomenal beliefs in ordinary contexts of justification, or in scientific methodology. Therefore, it would seem that a tenable foundationalism must identify at least some non-phenomenal perceptual beliefs as basic, and if there is a substantial innate conceptual contribution to perceptual beliefs, then as we have seen, these views are incompatible with the correspondence theory of truth, but are highly compatible with the ideal verification theory.

B. Coherentism

The basic idea of the coherentist view is to make justification system-wide instead of basing it on a privileged set of statements. Thus, the coherentist can allow that everything is potentially up for revision, while at the same time maintaining that each element in the system supports and is supported by the other elements, in that they exhibit a mutual coherence. While there have been several extensive developments of the general form of a coherentist picture as a replacement for foundationalism, e.g. by Lehrer and by BonJour[36], it has been somewhat difficult to develop a precise

36. Lehrer (1974), BonJour (1985).

specification of what coherence is. E.g. it is certainly at least consistency, but probably a lot more as well. But for our purposes, the intuitive idea of coherence should suffice.

Initially, the combination of a coherence theory of justification, a correspondence theory of truth and the doctrine of perceptual nativism appears quite compatible. Indeed, the coherence view may seem just the solution for the troubles which plague the foundationalist, since the coherentist does not grant a self-justifying status to perceptual beliefs, so the fact that our innate perceptual concepts may fail to correspond to external attributes need not be bothersome. The metaphysical realist may maintain that truth is external correspondence, and that perceptions are justified if they cohere with the rest of our beliefs, particularly our scientific beliefs. So everything appears in order, *prima facie*.

The difficultly is that the coherentist must ultimately accept something similar to the foundationalist view of perception. The reason is that there are various different conceivable sets of statements that exhibit a high degree of coherence. Think of sets of statements describing various possible worlds, for instance, especially if you think that there are possible worlds where our science isn't true. What is to keep someone from adopting any such set he pleases, and receiving the stamp of approval, in terms of justification, from the coherentist? Take for instance, the flat-earth

theory. It is consistent with some evidence, and there appear to be cosmological hypotheses that cohere with it. Suppose that someone were to adopt this view, or any other "crazy" yet coherent set of statements, and simply reject outright all evidence to the contrary. We do not want to say that such an individual is justified in his beliefs. He is, we assume, refusing to look at obvious evidence that would show the falsity of his presumptions. Clearly, the coherentist must make some move to rule out allowing that such cases exhibit genuine justification. Indeed, almost any statement appears to cohere with at least some others. So if we allow the blatant rejection of negative evidence, we must admit that virtually any claim is justifiable and this is simply too counter-intuitive for an account of justification.

The solution is to adopt a non-foundationalist dependence on observation. BonJour, who sees this problem clearly, provides the following resolution:

> as a straightforward consequence of the idea that epistemic justification must be truth-conductive, a coherence theory of empirical justification must require that in order for the beliefs of a cognitive system to be even candidates for empirical justification, that system must contain laws attributing a high degree of reliability to a reasonable variety of cognitively spontaneous [e.g. perceptual] beliefs.
>
> This requirement, which I will refer to as the *Observation Requirement,* is obviously quite vague, and I can see no way to make it very much more precise...The underlying idea is that any claim in the system which is not *a priori* should in principle be capable of being observationally checked, either

> directly or indirectly, and thereby either confirmed or refuted.[37]

As Bonjour makes clear, the requirement is not that observations must be accepted as true or correct, but only that they must be seriously considered in relation to one's prior theory. So this is not an endorsement of foundationalism, but merely a means of insuring that justification is "truth-conductive." And our current system of beliefs, including science, does attribute a fairly high degree of reliability to our perceptual beliefs, so again, everything would appear to be fine for the realist-coherentist.

But now consider what happens if it is the case that most of our innate perceptual concepts do not correspond to external objects and properties. The metaphysical realist supposes that there is a true theory--i.e. a set of beliefs that correspond to reality. Let us suppose that it is accessible to us. In the unfortunate situation we now consider, it looks like we must reject the true theory based either on lack of general coherence, or on BonJour's observation requirement. Thus, the true theory would appear to be largely inconsistent (and thus not coherent) with substantially inaccurate innate concepts--at least we can imagine possessing perceptual concepts that were inconsistent with the true theory in this way. In such a case, we would be

37. BonJour (1985), p. 141.

led to reject the true theory, even if we encounter it, since we can maintain it only by rejecting most, or all of our perceptual beliefs. Here, rejection of our perceptual beliefs is the right thing to do, from the external point of view, but from within our web on belief we could never distinguish such a case from a case such as the flat-earther who deliberately ignores contrary evidence.

Might increasing coherence lead us to the one true theory regardless of which set of perceptual concepts we begin with? It is difficult to see how this could be guaranteed. Specifically, it is difficult to see what could lead us to abandon all of our innate perceptual concepts (as far as theory is concerned.) After all, theories must explain observations--this is the point of the observation requirement. Even if we stumbled onto the ideal theory and even if it, internally, was ideally coherent, it is not clear how mere possession of the theory would lead us to accept it. Justificatory coherence must be system-wide, and given that perceptions make up a substantial proportion of our belief-set, and given that innate concepts contribute substantially to perceptions, it is difficult to see how the ideally internally coherent theory--which conflicted with inaccurate perceptual concepts--would lead to a greater total coherence than would the perceptions in conjunction with a highly coherent theory which included them.

This is the real problem with the set of views under consideration, i.e. it is quite conceivable that we could build alternative theories based on these inaccurate, innate concepts, theories which in and of themselves were quite coherent, and coherent in relation to our perceptual beliefs. Yet, the realist must maintain, we would actually be moving away from the truth by adopting such theories, or at least their justification would bring us no closer to the truth. But this is unacceptable, for it follows that on the coherentist's view, justification is not always truth-conductive, it does not always move us towards the truth even in the ideal. But then the account of justification cannot be correct, since even if ideal justification does not **yield** the truth, still, it would seem that ideal justification must at least approach the truth--i.e. the difference between completely unjustified and ideally justified beliefs must at least be that the latter have a better chance of being true than do the former. As BonJour writes:

> What then is the differentia which distinguishes epistemic justification, the species of justification appropriate to knowledge, from other species of justification?...The basic role of justification is that of a *means* to truth, a more directly attainable mediating link between our subjective starting point and our objective goal.
>
> ...it seems to follow as an unavoidable corollary that one can finally know that a given set of standards for epistemic justification is correct or reasonable only by knowing that the standards in

> question are genuinely conductive to the cognitive goal of truth.[38]

We have, however, just shown that this requirement is not satisfied for a coherence theory of justification. Basically, the problem is that the contribution of innate perceptual concepts to our webs of belief is substantial enough to allow that the pursuit of coherence will lead us away from the truth, if truth is correspondence with a mind independent reality.

On the other hand, if we reject the correspondence theory of truth in favor of the ideal verificationist view, then we find a solution. For on the verificationist view, there will be no set of "false" innate perceptual concepts. As long as an innate concept set is capable of sustaining a coherent development of some theory, then the ideal theory will at least partially vindicate those concepts, since it will partially depend on them through the observation requirement. So our innate, unjustified perceptual concepts, assuming there are a substantial number of them, cannot all be wrong![39]

38. BonJour (1985), pp. 7,9.

39. In accepting a verificationist theory of truth, there is room for some of our innate perceptual concepts to be mistaken. The observation requirement doesn't require that we accept all observations, and so maintaining that some innate concepts are mistaken, e.g. because they appear inconsistent with our best-justified scientific theories, does not threaten this constraint. While a substantial part of the totality of our innate perceptual concepts must cohere with our theories, any individual perceptual concept need not cohere.

1. The Case for Correspondence

The metaphysical realist could resist this move by showing that most of the innate concepts in perception do indeed correspond to a mind-independent world. If this can be shown, then, it might be argued, coherence does lead to truth after all, for us at least, and therefore there is no problem with the conjunction of a coherence view of justification and a correspondence theory of truth.[40]

First, consider the argument from natural selection. The coherentist/realist might argue that we have the innate perceptual concepts that we do because they have been selected for their survival value. But this, it might be argued, means that they must be roughly correct. That is, if we had a large number of incorrect innate perceptual concepts, then we would never have survived. But we have done quite well in terms of general survival, and in terms of moving around the environment and exploiting aspects of the environment. So, it might seem, natural selection guarantees that most of our innate perceptual concepts are correct.

The difficulty with this line of argument is that it drastically overstates what selection can guarantee. The most

40. Someone taking this position would probably have to concede that for other creatures with different innate concepts than ours, and incorrect ones, coherence wouldn't constitute justification. The trouble with this is that it seems that justification should be the same, wherever there are beliefs. However, I shall not worry about such issues since the arguments for our concepts' corresponding fail.

we can say from evolutionary considerations is that our concepts are good enough to have enabled us to survive this long. However, we have absolutely no idea how good they are, e.g. we don't that any of them pick out actual properties of (mind-independent) objects, and we certainly can't say by considering natural selection, which are correct and which are not.[41] What must also be shown is that being good for survival is the same as corresponding to mind-independent features. Of course, if such concepts did correspond, then we could explain their survival value, but this is not what we are looking for. What we need to show is how the fact that these concepts (jointly) have survival value supports correspondence.

It seems that there could be ways of representing the world that had survival value but did not actually correspond to mind-independent features of the world. In fact, we may already have discovered an instance of this situation in the case of color. Consider Hardin's recent presentation of the scientific evidence against the claim that color is a mind-independent property. He summarizes his case as follows:

> there is nothing in the world as described by the physicist which corresponds to the division of colors into hues. If we suppose hues to be physical properties that are neither on the physicist's list nor derivable from anything on the list, our knowledge of object color becomes totally mysterious. If, on the other hand, we identify colors with *bona fide* physical properties such as spectral reflectance or emitance profiles, we shall indeed have object characteristics that are

41. There are also methodological problems with adaptationist explanation. See Gould and Lewontin (1978).

> typically essential ingredients of explanations of why we have the color experiences we do. Distinct reflectance profiles then become distinct colors regardless of whether they are distinguishable by any human observers, and indefinitely many objects will be taken by us to be qualified by the same hue family despite marked dissimilarities in their reflectance properties. Colors will thus be properties of objects, but red, green and yellow will not. This does not seem to be a satisfactory solution to the problem of the ontological status of colors.
>
> An appeal to the color experiences of normal observers under standard conditions will assign colors to objects only approximately and relatively to particular interest and purposes. It is not just that colors turn out to be, as J.J.C. Smart supposes, disjunctive, gerrymandered physical properties when assigned according to *the* normal observer/standard condition procedure; it is, rather, that there is *no* such single, purpose-free procedure. In consequence, we are not entitled to say that physical properties have determinate colors *simpliciter*. Given a particular observer in a particular adaptational state and particular standard conditions, a color can be assigned to an object as precisely as the observer's perceptual condition warrants, but we cannot expect the assignment to remain the same when the set of conditions or the observer's adaptation state is changed. Assignments of colors to physical dispositions would thus not be just homocentric, or even ideocentric, but ideocentric *and* situational.[42]

While the case is far from settled, at the very least Hardin's considerations show that it is conceivable that our color concepts could fail to correspond to any sort of physical properties. Let us suppose that his view is correct--what can we say about the survival value of colors? Specifically, could color concepts have survival value in spite of failing to

42. In Hardin (1988), pp. 80-81. Also see Boghossian and Velleman (1989), who argue that "the best interpretation of colour experience ends up convicting it of widespread and systematic error" (p. 81.)

correspond? It seems that the answer is "yes", if for no other reason than because color concepts have great utility in our daily lives. And this utility is not, apparently, decreased by the discovery that color concepts do not correspond to physical properties.

While it is difficult to say exactly what this utility amounts to, we can consider a simple example that will provide a rough initial indication that will be sufficient for our purposes. Suppose that we discover that berries of a certain shade of red are poisonous, and from that point on avoid eating these berries. Our concept of redness need not correspond to any actual property of the berries. But what is important, in this case, is that this non-actual property is approximately co-extensive with (what we may suppose for the sake of example is) an actual property, i.e. being poisonous. We manage to do what it takes to survive--avoiding certain berries in this case--by following certain rules--e.g. rules for attributing colors, and, in this case, avoiding berries of a certain color. In this case, the predicates that the rules are formulated in do not correspond to actual, mind-independent features of the rules. But, nonetheless these rules do allow us to act in a way that achieves positive results. Roughly, what we do is associate negative affects--e.g. getting sick from eating the berries--with our non-corresponding predicate--"such-and-such a shade of red" and thereby manage to produce an appropriate generality in our

behavior--i.e. avoiding the "noumenal situations" that appear to us as red-berryhood, all without having any actual knowledge or representations of the **real** attributes of these situations.

Thus, the case of color shows that it is false that a concept will have survival value only if it actually corresponds to features of the mind-independent world. Yet, the argument from natural selection may still seem seductively plausible. I suspect that what has happened is that those tempted by it have mistaken explanation for justification in this case. We might explain our successful survival (to date) by claiming that as innate perceptual concepts go, ours are pretty good--i.e. they allow us to act so as to avoid a lot of harm and achieve a lot of survival-enabling benefits. However, what is required to support the correspondence hypothesis is something different, namely evidence which will justify the claim that our concepts (individually or jointly) are not just "good", but metaphysically correct--that they correspond to actual properties. And to do this, it would seem that we must show that there are no other concepts that would not provide information that would allow us to achieve equal or better survival. I.e. a successful argument for correspondence from natural selection would need to show something incredibly stronger, namely, that we have ideal survival information, but the facts of our success and survival don't show this.

To see this last point more clearly, consider an apparent case of inaccurate representation in a lower species. Most readers are probably familiar with the depictions of how certain insects' vision is supposed to work--we are shown a field of vision that has dozens of separate images for each eye, apparently implying that these insects do not form a single representation of external objects, but rather form many separate representations (let us suppose that this describes actual insect vision.) When we encounter such representations, we judge them wrong--the insects have failed to represent what we regard as the correct representation, that of a single, uniform world of objects. At the same time, though, we can imagine how such a multi-image visual field could be extremely useful, enabling the insects to avoid danger, determine directions and identify food sources. As long as actions are coordinated with appropriate totals of features or distributions of features of the sub-images, then the failure to have a representation of a single world of objects does not undermine survival. The point of this example is that to support the argument from selection to correspondence, the metaphysical realist must apparently show that there is little or no chance of some superior beings standing in the same sort of relation to our perceptual capacities that we (apparently) stand in to insect capacities. That is, we must be able to rule out the possibility of there being beings much more complicated than us who view our

"primitive" perceptual capacities as inaccurate, yet good enough to enable a substantial survival success. But there would seem to be no forthcoming evidence for such a claim, particularly not from evolutionary theory. And this is to say that the argument from selection to correspondence of innate perceptual concepts fails.

These same considerations undermine several similar attempts at supporting the claim that most of our innate perceptual concepts correspond to mind-independent features of the world. One of these arguments is that our perceptual concepts succeed relative to certain standards, e.g. we don't bump into things too much. We manage to get around in the world pretty well. This, it might be argued, shows that our perceptual concepts must generally correspond to external, mind-independent features of the world, otherwise we would encounter problems, such as waking into walls. However, these criteria for success are **internal** to the (apparently) innate set of perceptual concepts. I.e. bumping into things presupposes objects and spatial location, that are, by hypothesis, part of the innate, unjustified perceptual concept set. Someone advancing such an argument might imagine alternative possibilities that either falsely indicate the presence of objects when some are present (by our standards) or which give no indication of the external situation at all. It is clear that such alternatives are inferior. But what must also be shown is that there are no alternative possibilities

which, by standards internal to those concepts, could allow us to do as well or better in terms of getting about and generally coping with our environment than we do relative to our standards. Since there is no apparent way of sketching out such a possible concept-space, let alone evaluating our position in it, it seems that there is no hope for this sort of defense of innate concept correspondence.

As a final possibility for a defense of innate perceptual concept correspondence, consider the claim that the best hypothesis, given everything that we know--all the evidence that we have--is that most of our perceptual concepts (e.g. color aside) do in fact correspond to mind-independent features of the world. But what would justify such an explanation? To put it a slightly different way, what would there be to stop creatures with the sort of insect vision that we considered above from advancing the same argument--what would justify us and not them in accepting the correspondence hypothesis? It cannot be that we can achieve a better science than such creatures, for then we would need to demonstrate that no creatures with different concepts could achieve a science better than ours, and this we surely can't do. Moreover, it is perfectly possible for the anti-realist to accept the hypothesis that most of our innate concepts are correct. Given the ideal verificationist theory of truth, this is to say that under ideal justificatory conditions, we would still accept beliefs involving those concepts--i.e. we would

never develop theories which are inconsistent with most of our ordinary perceptual beliefs. In a manner of speaking, it could even be granted that such concepts "correspond" to a mind-dependent world. But it is not apparent that anyone seeking an explanation of the status of innate perceptual concepts vis-a-vis the world would require anything more than this. Which is to say that there is no support here for the metaphysical realist.

But without support for the claim that most of our innate perceptual concepts correspond to features of mind-independent reality, the presence of such concepts in perceptions implies that if justification is coherence, then there is no guarantee that increasing coherence will approach truth unless truth is ideal justification.

C. Reliabilism

The final view of justification that I will consider is a reliabilist, or, more generally an externalist account of justification. What is notable about externalist accounts is that they propose that a belief is justified just in case certain conditions in the external environment are satisfied, whether or not the believer knows of their satisfaction. This approach may seem highly favorable to metaphysical realism, since the externalist account seems to resemble the metaphysical realists' account of representation and truth. However, as I shall now argue, an externalist account,

considered in light of perceptual nativism, is as incompatible with metaphysical realism as foundationalism and coherentism.

I will examine Goldman's reliabilist view of justification, since this is the most well-developed externalist account. Goldman stresses that justification should be something that leads to true beliefs. And not just true beliefs (think of a theory of justification that sanctioned countless true beliefs by sanctioning ten times as many false beliefs), but a high percentage of true beliefs. Specifically, what Goldman suggests is that a theory of justification, or set of justifying rules, would evaluate belief processes and sanction only those that produced an acceptable ratio of true beliefs, some unspecified ratio greater than 50%.[43] A belief will thus be justified on this view if it is produced by a process that generally produces true beliefs--a reliable process.

A reliabilist view of justification would seem to harmonize well with metaphysical realism, even in the face of perceptual nativism. Thus, on a metaphysical realist view, beliefs will be justified just in case they are produced by processes which reliably produce beliefs that accurately represent the mind-independent world. If most of our innate perceptual concepts do not correspond, then few is any of our beliefs are justified, whether we know it or not--in fact we

43. See Goldman (1986), p. 106.

probably would not know it in such circumstances. But, it would appear, there is no inconsistency here.

However, there is an unacceptable result in the imagined situation where most innate concepts fail to correspond. The reliabilist account implies that few, if any of our perceptual beliefs would be justified. But this is surely quite implausible. Suppose someone in the imagined situation had come to an inconsistent pair of beliefs by two difference perceptual means. On the reliabilist view we are considering, that person would be no less justified in believing the contradiction than in believing only one or the other of the beliefs. But it seems reasonable to think that we are always less justified in believing an explicitly contradictory set of beliefs than we are in believing a non-contradictory set.

Or consider someone who came to some perceptual belief *p* in the imagined situation, and then discovered by some ordinary means that *p* was in error--e.g. an illusion. Again, the reliabilist would maintain that the individual would be no more justified in believing not *p* than in believing *p*. But this seems unacceptable. For all we know, we are indeed in this situation, however, it seems, we are justified in rejecting those beliefs that we find, by ordinary evidential means, to be mistaken. Ultimately, the suggested reliabilist account implies that unless our perceptual processes are mind-independently reliable, any appearance-reality distinction we draw will be unjustified. But surely we are at least partially

justified in distinguishing appearance and reality, even if what we think is reality turns out to be incorrect.

Finally, note that people in the imagined case could develop rudimentary, and even sophisticated theories that were based largely on perceptual observations. But processes which based their theory-production on non-corresponding observation concepts would be ruled unreliable by the proposed account. This is to say that, for two theories and some set of observations, even though theory *A* provided a full and complete explanation of the data, accurate predictions, etc., while theory *B* failed on all these scores, individuals in this situation would be no more justified in believing theory *A* than in believing theory *B*. But this is surely unacceptable--individuals in this circumstance are more justified in believing *A* than *B*.

This illustrates a more general point, which is problematic for reliabilist views as far as science is concerned, namely that not all false theories are equally unjustified. Some false theories are better than others, and it is by rejecting the less justified theories that we produce a better science. Thus, it is likely that most of the scientific theories we now hold are false, whether our innate concepts are "veridical" or not. The search for better theories is a progress through more and more highly justified false theories, in pursuit of the truth.

The suggested combination of reliabilism and metaphysical realism is therefore unacceptable. Goldman appears to recognize the problem through consideration of Cartesian-demon examples.[44] His solution is to tie justification to the evaluation of belief-forming processes in a certain range of possible worlds:

> We have a large set of common beliefs about the actual world: general beliefs about the sorts of objects, events, and changes that occur in it. We have beliefs about the kinds of things that, realistically, do and can happen. Our beliefs on this score generate what I shall call the set of *normal worlds*. These are worlds consistent with our general beliefs about the actual world. Our conception of justification is constructed against the backdrop of such a set of normal worlds. My proposal is that, according to our ordinary conception of justifiedness, a rule system [for justification] is [acceptable] in any world W just in case it has a sufficiently high truth ration *in normal worlds*.[45]

Clearly, normal worlds are those, that, among other things, have properties that correspond to most of our innate perceptual concepts. The suggestion is that we decide on reliability for belief-forming processes in worlds where the concepts do correspond, so that there will be some beliefs that are justified even in worlds where the innate perceptual concepts do not correspond to mind-independent properties at all.

44. See Goldman (1986), p. 113.

45. Goldman (1986), p. 107.

However, the revised account no longer connects justification with truth. To see this, simply consider that if we are not in a normal world, then none of the beliefs yielded by acceptable processes will be true. Therefore, sets of increasingly justified beliefs will approach the truth just in case the world is normal. But this is to say that normal world reliabilism in conjunction with metaphysical realism faces precisely the problem that we have seen for coherence theories and metaphysical realism: the rules for justification that are sanctioned on the reliabilist view will not generally promote true belief, even under ideal application, if most innate perceptual concepts are mistaken. Thus, justification will be "verific" only if most of our innate perceptual concepts are correct. But this is unacceptable. Justification should be lead us towards the truth whether or not we start with true beliefs.

Goldman notes that this is a problem, but he doesn't take it too seriously. He merely suggests that by "epistemic bootstrapping" we can escape false initial beliefs.

> We start with a set of available processes with varying degrees of reliability. We use the more reliable processes to identify good methods. We then use the more reliable processes, together with some of the good methods, to identify the various processes and their respective degrees of reliability. The superior specimens are identified, and their use is said to be justification-conferring. The inferior specimens are so

> identified, and their use is said to be non-justification-conferring.[46]

How, though, are we to tell initially which processes are reliable and which are not reliable? If reliability is mind-independent correspondence, then there is no apparent way to initially identify "good" methods and get the bootstrapping started. On the other hand, if reliability is normal world correspondence, then the problem has not been solved. For identification of good methods and the subsequent selection of reliable processes will in fact move towards true beliefs only if the world really is normal. So again, the account implies that we can approach truth through justification just in case we are lucky enough to start with substantially true beliefs. But this is unacceptable, since justification should be a means toward truth regardless of the starting point.

Unlike the cases of foundationalism and coherentism, it is not apparent that an ideal verificationist theory of truth will solve the difficulty. This is because the substitution of ideal verificationism for truth in the reliabilist formula appears to yield a circular account:

> the belief that *p* is justified iff it is the result of processes that yield a suitable ratio (i.e. > 50%) of ideally justified beliefs.

This apparent circularity leads Goldman to endorse a correspondence theory of truth.[47] However, as I shall now

46. Goldman (1986), p. 120.

47. See Goldman (1986), pp. 116-7 and chapter 7.

argue, an ideal verificationist theory of truth is strictly speaking compatible with reliabilism, although the former shows the latter to be an somewhat limited account of justification.

The proposed account is clearly circular as a criteria of justification--we cannot apply it to a given belief to tell if it is justified or not. Let us consider, though, how it would be improved with a correspondence theory of truth. With such a theory, the reliabilist dictum might be read as telling us to accept a belief just in case it is produced by a means that yields a suitably high ratio of beliefs that correspond to the mind-independent world. But this does nothing towards telling us how to determine if a given belief correspond or not. And in fact, it might be argued, that is precisely the task facing the justification theorist, viz., provide a set of rules that will enable us as much as possible to believe the true and reject the false. Noting this, and taking a clue from the passage quoted above on bootstrapping, we might suppose that the reliabilist presupposes a range of ordinary methods and belief-forming processes which will allow us to distinguish apparent truth from apparent falsehoods. Ideally justified--i.e. true--beliefs will thus be those that those that are produce by processes or methods that have withstood the test of evaluation by all other reliable methods, original or derived. If we term such methods "acceptable", we get the following criteria:

> the belief that *p* is justified iff it is the result of processes that do not yield an unacceptable ratio (i.e. > 50%) of beliefs that are undermined by acceptable methods.

which is no longer circular. In effect, this is to naturalize justification to our original, common sense methods. Here, the reliabilist rule serves not as a criteria applicable in isolation from all other methods of justification, but rather as a constraint on what sort of methods and processes of justification are acceptable.[48]

This solves one of our difficulties, in that it enables reliabilist justification to approach the truth, even given perceptual nativism, since whatever innate processes and methods we begin with, for such processes and methods will either be truth-conducive, on the ideal verificationist theory, or they will enable bootstrapping to, again by definition, truth conducive processes and methods. However, we have not as yet met the other challenge that I raised above, namely that it seems that some false beliefs may still be at least partially justified. This is a problem, I suggest, for externalism generally, since by linking justification too closely with truth, the reliabilist fails to allow for false yet (partially) justified belief. The remedy, it would appear,

48. This is how Goldman (1980) sees an externalist view of justification.
Also note that the account is still externalist, or as Goldman (1986) puts it objectivist, in that whether or not a given belief is justified is independent of any belief anyone ever actually holds.

is to make justification depend on the available evidence. This seems appropriate if we consider scientific theories--often, a theory can be justified relative to a certain amount of evidence, even though it is shown false, and thus not justified, when more evidence comes in. The revised criteria would be:

> the belief that *p* is justified iff it is the result of processes that do not yield an unacceptable ratio (i.e. > 50%) of beliefs that are undermined by presently acceptable methods.

While this constitutes a move away from externalism to internalism, it nonetheless appears to be the most plausible form of a reliabilist account.

D. Conclusion

We thus have the following argument for an ideal verificationist theory of truth, and more generally, for neo-kantianism: psychological theories of perception suggest that it is likely that most of our very abstract perceptual concepts are innate. We add them to our perceptual inputs because we are genetically programmed to do so, not because we have learned to do so from experience. But such nativism renders metaphysical realism--a correspondence theory of truth--untenable in conjunction with either a foundationalist, coherentist or reliabilist theory of justification. In the case of foundationalism, we see that nativism allows an unacceptable potential for falsity in the perceptual

foundation. In the case of coherentism, the possibility of such falsity allows that justification will not always approach truth, even in the long run. And in the case of reliabilism, there will be justification only if the perceptual concepts correspond to begin with. But none of these outcomes is acceptable by the standards of what an account of justification should give us. Since foundationalism, coherentism and reliabilism exhaust the available alternatives for an account of justification, this line of reasoning constitutes an argument from elimination against metaphysical realism (i.e. it isn't compatible with any of the alternatives.) Moreover, replacing a realist theory of truth with an ideal verificationist theory overcomes these difficulties, thus supporting the ideal verificationist view.

We can also consider the possibility of accepting the ideal verificationist theory of truth, but maintaining that truth is correspondence to the mind-independent world. This is not acceptable, since, as we have seen, each account of justification implies that, given a substantial innate contribution to perception and the possibility of a failure of mind-independent correspondence for these concepts, ideal justification will not guarantee correspondence. Thus, if the concepts in the perceptions in the foundationalist's basic beliefs do not correspond, then they will still fail to correspond no matter what, i.e. when all evidence is in. And we have also seen that there is no reason to think that an

ideally coherent set of beliefs which began from non-corresponding innate perceptual concepts would converge on beliefs which corresponded to the mind-independent world. Finally, we have seen that ideally reliable beliefs will correspond to the mind-independent world just in case the initial processes which form the basis for reliabilist bootstrapping themselves correspond. So none of these views support the claim that ideally justified beliefs will correspond to the mind-independent world. And any of them thus supports neo-kantianism.

The metaphysical realist might argue that this only shows that there is no acceptable account of justification. If this is merely to suggest that there could be some other account of justification which is consistent with metaphysical realism and perceptual innateness, then the burden of proof is clearly on the metaphysical realist to produce a compatible view. And not just any compatible view, but one that is at least as independently plausible as competing accounts of justification.

However, the argument might be, not that there is some alternative account of justification, but rather that no such account is possible. Specifically, the metaphysical realist might object that I have maintained that an account of justification will show us how increasing justification yields or approaches truth. But, it might be argued, this is too strong a demand--it is reasonable to allow that skepticism

might be true, that no account of justification in this sense is possible.

Note, though, that I have not insisted that any of our beliefs must be justified, but only that we have an account of what it takes to justify them. Given that justification will point towards truth, skepticism says that none of the beliefs we now hold are, as a matter of fact, justified. So I am not assuming that the metaphysical realist must provide an answer to (the standard, global form of) skepticism.

The position implied by this line of reply thus amounts to the suggestion that, while maintaining belief in a mind-independent reality, we give up hope of ever knowing that we have any true or nearly true beliefs about it--for this is the result of giving up on ever having an account of justification. I.e. if we can never know if or when any of our epistemic practices are truth-conductive, then we cannot claim to have any basis for holding that we have or ever can have any true beliefs. This is, in effect, to grant part of the neo-kantian position that I have characterized above--i.e. complete skepticism about the possibility of knowledge of the mind-independent world. However, the suggested position also leaves us without any understanding or support for our justificatory practices, e.g. detection of perceptual errors, or scientific methodology. That is, giving up on an account of justification means that we will not have any explanation of why we should reject the beliefs we decide are false while

continuing to believe those that we do not find to be mistaken. If we have no possible grounding, either explanatory or normative, for our ordinary justificatory practices, then why should we engage in them at all? Here, it seems to me that the most plausible move is to adopt the other half of the neo-kantian position too, namely that truth is idealized justification, where justification is understood according to one or another of the standard accounts. Such an account grounds at least some of these practices, or offers new ones, and thus provides an acceptable account of the notion of truth that plays a role in our lives.

Thus, the rejection of the possibility of an account of justification is no real option for the metaphysical realist. Confusion on this point may come over mistaking a fallibilist response for the rejection of an account of justification. If the view is that we have certain well-confirmed theories, but we shall perhaps never know if they are true or not, then an account of justification is still required, i.e. an account which describes or prescribes what we can or should count as a well-confirmed theory. This is probably not going to be a traditional, foundationalist view, that offers certainty for some beliefs, but it will be an account of justification all the same (probably a coherentist account.) On the other hand, the rejection of the possibility of an account of justification, as we have just seen, involves complete skepticism about our justificatory practices, leaving us in

the implausible and unpalatable position of having no grounds for endorsing some theories while rejecting others.

Since there are no other options left for the metaphysical realist, I conclude that we should abandon the correspondence theory of truth and metaphysical realism in favor of the ideal verificationist theory of truth and neo-kantianism.

V. Anti-Realisms

There are several aspects of the present view that separate it from other forms of neo-kantianism or anti-realism, as well as from metaphysical realist views, and I will examine these in this and the final section.[49] A crucial issue concerns the distinction between the mind-dependent and the mind-independent. If the neo-kantian is unable to convincingly draw this distinction, then the view becomes idealism or perhaps transcendental realism. The former type of account claims that all existence is mind-dependent, while the latter view holds that metaphysics is altogether impossible--thus there is no saying that the world is mind-dependent or mind-independent, it is simply what we theorize about--and all we can do is examine (e.g. the ontology of) our beliefs and theories.

49. I shall not discuss anti-realist views such as Dummett's which are verificationist as far as truth is concerned but which reject the idea of there being a contribution to knowledge on the part of the mind.

The present view draws this distinction in terms of the fact that we are cognitive beings that have input to our perceptual systems, the fact that we add content to our perceptions, and the fact that these perceptions form the basis for our knowledge, in one of the ways discussed above. We can understand the possibility of having different innate concepts--or processing the same input differently. And we must acknowledge that the same input, together with different innate concepts, could yield a different ideally confirmed, i.e. true, theory of the world. And we can also conceive of our innate concepts corresponding or failing to correspond to the mind-independent world, although we could never know of this, since our inquiry is rooted in the (conceptual) world which is constituted by the perceptual categories that we possess. Thus, we must acknowledge limits to our ability to make metaphysical claims--all our metaphysical claims must be relativized to the innate basis from which we build our theories. And, while we can conceive of a mind-independent metaphysics, we also see that we cannot conduct an investigation of its nature.

Kant himself held that the way to answer mind-dependent metaphysical or ontological questions (about *a priori* concepts), such as, "do objects exist?" was through *a priori* analysis. I would suggest instead that such questions can only be answered by evaluating our best confirmed theories and beliefs (to date.) Thus, this brand of neo-kantianism turns

out to be quite close to a very moderate, pragmatic realism, which answers all metaphysical questions by pointing to our best justified scientific (and common sense) theories. However, the neo-kantian adds to this moderate realism a qualification that our knowledge is mind-dependent.

A second important dimension of neo-kantianism concerns the question of how we are to think of alternative conceptual contributions. **Pluralists**, such as Goodman or Kuhn, urge that there are many alternatives--many different mind-dependent "worlds" which are actually accessible in human history. Goodman suggests that we are free to move from "world" to "world"[50] while Kuhn holds that different eras of science constitute different "worlds."[51] Such pluralists also endorse the claim that there is no single, best, "right" world. On the other hand, there are the "**absolutists**" such as Sellars or (maybe) Putnam who hold that we are progressing through various "worlds" or conceptual systems toward a single, ideal theory.[52]

While I cannot discuss these views in detail here, I suggest that we reject both alternatives on the grounds that it appears that nativistic perceptual theories tell us that

50. See Goodman (1978).

51. See Kuhn (1962/70) and especially Kuhn (1989).

52. See Sellars (1968) and Putnam (1980), especially p. 216.

our observational concepts are not very plastic.[53] Moreover, it also seems plausible to claim that we have a set of relatively stable, perhaps largely innate, justification methods. It may be the relative invariance in these partial sets of observation concepts and justification methods that allows us to see our history as a more or less unbroken line of investigation into a single world, despite radical differences in theories and high-level observation concepts from different eras. Thus, nativist neo-kantianism fails to support the view that we ever move from "world" to "world." Instead, I recommend a "monistic" version of neo-kantianism in which there is only one world (barring such things as encounters with beings with radically different innate concepts than ours, if indeed we could communicate with such beings at all.) So, again, the present account comes very close to a moderate, extremely pragmatic realism.

VI. The Synthetic A Priori

A distinct feature of the present view is the implication that there is (non-trivial) synthetic *a priori* knowledge. The following is a recipe for determining what is synthetic *a priori* in perception, and perhaps in all of knowledge as well. Take the innate concepts that the perceptual system contributes to perceptions and subtract the elements not

53. See Fodor (1984).

vindicated by our ideally confirmed theories. The remaining concepts, when formulated in terms of truths concerning their instantiation, e.g. there are material objects, constitute statements that are true in the face of no, any or all evidence, yet that are not made true through their logical form or meanings alone, but rather, by their dependence on our perceptual systems.

Our considerations suggest that there must be at least some such truths, assuming that there are substantial innate contributions to perception, since either the perceptual foundation, the observation condition of a coherentist or the common sense perceptual methods of the reliabilist insures that not all of this substantial set can be rejected. On way to think of our result is the following. Our perceptions serve to fix the context of inquiry--we must use them to insure that we are inquiring about the actual world, rather than some fictional world. Now, Kripke-style causal reference examples suggest that reference fixers may sometimes succeed even when the attributed properties do not actually apply. Thus, it seems possible to ostend a certain substance that appears to be hard and blue as 'gold', and do it successfully, even though gold turns out to be yellow and malleable. However, our present considerations also suggest that such error can only go so far. Ostensions typically proceed against a relatively fixed background of perceived material objects. Thus, if there not only turns out to be no hard, blue substance present, but

no material substances at all, or if our spatio-temporal framework proves unworkable, such ostensions appear to fail. For then, anything at all could be the actual referent of the ostension. Likewise, it seems that we can't give up all of the perceptual qualities and concepts that our perceptual systems supply us with, for then, as far as justification and ultimately truth is concerned, the world could be anything at all.

However, it seems there is little reason at present to expect that we will ever have a clear formulation of exactly what the synthetic *a priori* in perception comes to. For it is not obvious that we will ever have a definitive and complete specification of what is innate and unjustified in perception--this is a monstrously complicated task, to say the least. Nor can we ever expect to actually have anything approaching the ideally justified belief-set. Even those more optimistic that I about accessing these items must surely agree that there is no hope for such a specification at present.[54]

Thus, we have a position that, though very Kantian in the final result, is directly opposite in methodology. We do not start with a neat, clearly defined set of synthetic *a priori* truths and proceed, *a priori*, to conclusions concerning the mind's contribution to knowledge and an advocation of mind-

54. It is also possible that what remains as the synthetic *a priori* is relative to our choices concerning which facts we want to explain. Such considerations might be a way of making the present view more pluralist.

dependent realism and transcendental idealism. Instead, we must begin with the messy psychological facts and inquire as to what this shows us about truth and reality.

Perhaps it will be objected that it is outrageous that such an important matter as the decision between metaphysical realism and neo-kantianism should rest on contingent psychological facts. But this is exactly what we should allow for if we are going to genuinely naturalize epistemology[55] and philosophy. One might expect, *prima facie*, that naturalistic enquiries would support popular realist positions. But I can see no reason why things should ultimately turn out this way.[56] Naturalism might well upset some of our most firmly held philosophical views, as the present inquiry demonstrates.

55. I have not argued that a theory of justification can be naturalized, although I think this is quite plausible.

56. In fact, Matheson (1989) argues that a naturalist epistemology is more readily compatible with an ideal verificationist theory of truth than with a realist, non-epistemic theory.

BIBLIOGRAPHY

Baillargeon, R., E. Spelke and S. Wasserman (1985) "Object Permanence in Five-Month-Old Infants." *Cognition*, 20, pp. 191-208.

Boghossian, P. and J. D. Velleman, (1989) "Color as a Secondary Quality." *Mind*, 389, pp. 81-103.

BonJour, L. (1985) *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Chisholm, R. (1980) "A Version of Foundationalism." In P. French, T. Uehling and H. Wettstein (eds.), *Midwest Studies in Philosophy, Vol. V*, Minneapolis: University of Minnesota Press, pp. 543-564.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1980) *Rules and Representations*. New York: Columbia University Press.

Fodor, J. A. (1981) "The Present Status of the Innateness Controversy." In *Representations*, Cambridge, MA: MIT Press/Bradford, Chapter 10.

Fodor, J. A. (1983) *Modularity of Mind*. Cambridge, MA: MIT Press/Bradford.

Fodor, J. A. (1984) "Observation Reconsidered." *Philosophy of Science, 51*, pp. 23-43.

Gibson, E. J. and R. D. Walk (1960) "The Visual Cliff." *Scientific American*, 202, pp. 64-71.

Goldman, A. (1980) "The Internalist Conception of Justification." In P. French, T. Uehling and H. Wettstein (eds.), *Midwest Studies in Philosophy, Vol. V*, Minneapolis: University of Minnesota Press, pp. 27-51.

Goldman, A. (1986) *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goodman, N. (1978) *Ways of Worldmaking*. Indianapolis: Hackett.

Gould, S. J. and R. C. Lewontin, (1978) "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." In *Proc. R. Soc., London*, 205, pp. 581-598. Reprinted in *Conceptual Issues in Evolutionary Biology*, Sober (ed.), MIT Press, pp. 252-270.

Hanson, N. (1961) *Patterns of Discover*. Cambridge University Press.

Hardin, C. L. (1988) *Color for Philosophers*. Indianapolis: Hackett.

Held, R. (1985) "Binocular Vision: Behavioral and Neuronal Development " In J. Mehler and R. Fox (eds.), *Neonate Cognition*, Hillsdale, NJ: L. Erlbaum Associates.

Hofsten, C. von (1986) "Early Spatial Perception Taken in Reference to Manual Action." *Acta Psychologica*, 63, pp. 323-335.

Hume, D. (1739-40) *A Treatise of Human Nature*. L. A. Selby-Bigge (ed.), (2nd. ed.) Oxford University Press, 1976.

Kant, I. (1787) *Critique of Pure Reason*. N. K. Smith (trans.), New York: Macmillan/St. Martin, 1929.

Kellman, P. and E. Spelke (1983) "Perception of Partly Occluded Objects in Infancy." *Cognitive Psychology*, 15, 483-524.

Kuhn, T. (1962/1970) *The Structure of Scientific Revolutions* (2nd. ed.) Chicago: University of Chicago Press.

Kuhn, T. (1989) "Possible Worlds in History of Science." In S. Allen (ed.), *Possible Worlds in Humanities, Arts and Sciences, Proceedings of Nobel Symposium 65*, New York: Walter de Gruyter.

Lehrer, K. (1974) *Knowledge*. Oxford University Press.

Marr, D. (1982) *Vision*. San Francisco: W. H. Freeman.

Matheson, C. (1989) "Is the Naturalist Really Naturally a Realist?" *Mind*, 390, pp. 247-258.

Putnam, H. (1975) "The Meaning of 'Meaning'." In K. Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII, Minneapolis: University of Minnesota Press. Reprinted in Putnam, *Mind, Language and Reality, Philosophical Papers Vol. 2*, Cambridge University Press, pp. 215-271.

Putnam, H. (1978) "Realism and Reason." In *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.

Putnam, H. (1981) *Reason Truth and History*. Cambridge University Press.

Putnam, H. (1983) *Realism and Reason, Philosophical Papers Vol. 3*. Cambridge University Press.

Putnam, H. (1989) "Model Theory and the 'Factuality' of Semantics." In A. George (ed.), *Reflections on Chomsky*, Oxford: Basil Blackwell, pp. 213-232.

Sellars, W. (1968) *Science and Metaphysics*. London: Routledge & Kegan Paul.

Spelke, E. (1985) "Perceptual Looking Methods as Tools for the Study of Cognition in Infancy." In G. Gottlieg and N. Krasnegor (eds.), *Measurement of Audition and Vision in the First Year of Post-Natal Life*, Hillsdale N.J.:L. Erlbaum Associates, pp. 323-64.

Spelke, E. (1987) "The Origins of Physical Knowledge." In L. Weiskrantz (ed.), *Thought Without Language*, Oxford University Press.

Spelke, E. (1990) "Origins of Visual Knowledge" In D. Osherson, S. Kosslyn and J. Hollerbach (eds.), *Visual Cognition and Action: An Invitation to Cognitive Science, Vol. 2*, Cambridge, MA: MIT Press.