Toward a Linguistic Conception of Thought

Benjamin J. Stenberg

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2006

Program Authorized to Offer Degree:
Department of Philosophy

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Benjamin J. Stenberg

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Cass Weller

Reading Committee:

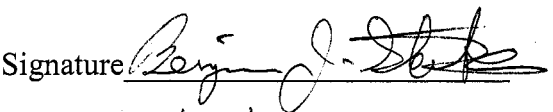
Lynn Hankinson Nelson

Cass Weller

Andrea Woody

Date: 10/19/2006

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date 12/6/06

University of Washington

Abstract

Toward a Linguistic Conception of Thought

Benjamin J. Stenberg

Chair of the Supervisory Committee:
Associate Professor Cass Weller
Department of Philosophy

The traditional, and still most common, view of the relationship between language and thought is that language is merely a tool for expressing thoughts. I argue that this view is mistaken and that language is the very thing that makes thought possible in the first place. I mean a number of things by this. I mean that no one can have any thoughts at all without first being a practiced member of a linguistic community. I mean literally that the thoughts themselves are linguistic entities: thoughts are, as it is sometimes put, just 'sentences in the head.' And, furthermore, I mean that thoughts are actually constituted by symbols of the natural language(s) one speaks. I call this view the Linguistic Theory of Thought (LTT), and it is a version of the Representational Theory of Mind. The main goal of my project is to lay the foundation for the LTT: to show that it is at the very least a coherent possibility, and find space for it among its representationalist peers. Having found a general place for the LTT to sit, I proceed to argue that this position is stable. First I develop a theory of meaning, adapting Wilfrid Sellars' view that specifying the meaning of a linguistic type involves classifying that type functionally. By treating meaning as functional classification I can specify the meanings of words without

appealing to any supposedly antecedent thoughts. Next I argue that thinking is a matter of social practice. In this part of the project I rely on the philosophical framework provided by Sellars (his 'Psychological Nominalism'), and then modify the Sellarsian framework to support my view that our thoughts themselves are constituted by the symbols of the natural language(s) we speak. Ultimately, my project is both a modernizing of Sellars' view, developing and broadening his arguments in order to critique many of the representationalists that wrote (and continue to write) after him, and the development of a modern representationalist theory of thought that takes seriously the insights of Sellars, Quine, and those who have followed in their footsteps.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

A graduate career is invariably a collaborative effort. I owe a debt of gratitude to a great number of people, both to those directly involved in guiding this dissertation, and to those who have influenced and been a part of my life in other ways. I cannot possibly thank everyone, but the following remarkable people deserve to be recognized.

First, there are those most directly responsible for enabling this project to be written: the members of my dissertation committee. Professor Cass Weller, who mentored me through every stage of my graduate career, and whose unfailing support, constant feedback, and tireless willingness to guide this project is evident in every single thing that it does well. Professor Andrea Woody, who was always available to advise me, with her characteristic objectivity, calm, and poise, as I navigated the ups and downs of a graduate career and took my first steps toward becoming a professional philosopher. And Lynn Hankinson Nelson, a brilliant and generous professional.

Furthermore, I would like to thank: Professor Ellen Klein, who first taught me what philosophy could be; Professors Jean Roberts and Ann Baker, two wonderful people from whom I have learned so much about what it means to be a professional philosopher; and the office staff of the University of Washington's Department of Philosophy—most especially Barbara Mack, a friend and a consummate professional.

I would also like to thank my three dearest friends: Katie Cole, first among equals, who has stood by me constantly since we were kids—I have leaned on her, and relied upon her steadfast love and support, at every stage of my life; Kevin Rennert, a man that anyone would be lucky to call a friend, whose unfailing encouragement through thick and thin has helped me so many times and in so many ways; and Karen Mazner, a colleague and confidant, whose friendship has helped me through these last few years more than she knows. I would not be where I am today without these three extraordinary people.

Finally, and most importantly, I wish to thank my parents, Bill and Kristy, who have believed in me all my life. No one could ask to be raised by two more loving and supportive people. They have, more than anyone else, made me who I am. Thank you both.

DEDICATION

To my wife, Amanda.
All that I do, all that I am,
is for you.

1. The Representational Theory of Mind

Here is a familiar sequence of events: I have a thought, decide it's worth sharing, and use words to communicate that thought to my fellow human beings. No doubt something like this happens with each of us countless times every day. This is as familiar an experience as one could hope for. So familiar, in fact, that it makes a particular philosophical thesis seem like common sense. The thesis in question is the idea that it is perfectly possible for me to have thoughts even if I lack a language with which to express them. That is, that language is no more than a *tool*, a convenient way to make my otherwise private thoughts part of the public domain. I think this is mistaken. One would be hard pressed, though, to deny that this particular way of conceiving the relation between thought and language has been largely accepted, by both philosophers and non-philosophers alike.

Nevertheless, I intend to argue that this widely accepted view is false, and I plan to show that an alternative view—*viz.*, that language, far from being merely a tool for expressing thoughts, actually makes thought possible in the first place—is both plausible and has a number of advantages to recommend it. When I say that language makes thought possible, I intend for this dependence to be understood as a very strong one: *no being could have any thought whatsoever without first having lots of experience participating in a social practice of language use.* I also believe that the thoughts themselves are, at some level of description, linguistic in nature.[1] This latter claim is

[1] As a physicalist I am committed to the idea that there is another level of description at which mental events are physical events. This has no bearing on the view that I will argue for here, however, since I do not think that any given physical event (say, a neurophysiological event) is *intrinsically* a thought. What makes something a thought, on my view, is, roughly, its functional and representational features, not its physical characteristics.

perhaps the definitive mark of what we might call a 'linguistic conception of thought.' I think it is the former claim—that thinking depends upon participation in a social practice of language use—that makes my linguistic conception of thought differ from most others. Let us call the view that I will endorse here the Linguistic Theory of Thought (LTT) to distinguish it from linguistic conceptions of thought generally.[2] The main goal, then, of this project will be to make room for, that is, to lay the foundation for, the LTT. I want to show that such a view is coherent and sensible; and I want to argue the virtues of such a view in the face of various other theories of mind.

The first step in laying the foundation for the LTT is to locate it within the general landscape of the philosophy of mind. To that end it will be necessary for me to create a number of general, and very broad, distinctions between different theories, or groups of theories, in the field. Any attempt to fit some of the very precise theories in philosophy of mind into rather generic categories is bound to make a mess of their more subtle points. My goal, however, is not to describe any of these views definitively; rather, I hope only to sketch them in enough of a rough outline to be able to say where the LTT fits amongst them. I begin this task in the later parts of this chapter, and continue it in more detail in Chapter Two. Much of this task (the detailed part that will occupy us in Chapter Two) will focus on trying to organize the vast field of views that might all be considered versions of *the representational theory of mind* (RTM). According to such theories to have a thought is to have a certain representation occur in one's mind or brain

[2] Throughout this project I often use the term 'thought' to refer what we might call 'thinkings', i.e., that general category of mental states which are in some sense 'acts of thinking' (a broad category that includes such things as believing, wanting, wondering, intending, hoping, questioning, etc.). The term 'thought' is, however, ambiguous between acts of thinking and the propositional contents of such acts. Context ought to make it clear when I mean by 'thought' the former, and when the latter.

(the nature of the representation will vary from theory to theory). That representation will have a certain content; and the content of the representation will be the content of the thought. That is, according to representationalists, to have the thought that-*p* is to have a representation that means *p* occur in one's mind or brain. The main task, then, for most representational theories of mind is to say how representations come to have the content that they do. This is also the main task for the LTT. Much of the rest of this chapter, on the other hand, will be dedicated to a brief consideration of some theories that are *not* versions of the representational theory of mind, as well as an argument for a very generic type of representationalism.

Once I have managed to locate the general space that I feel the LTT occupies, I will then need to begin to show how such a view could be considered plausible. This task will occupy us throughout Chapters Three, Four, and Five. In Chapter Three we will look at Sellars' 'functional classification' theory of meaning—a theory that treats meaning as a matter of classifying linguistic tokens in terms of the role they play in the linguistic economy of which they are a part. Words (and, by implication, according to the LTT, the representations which constitute our thoughts) on this view are meaningful because they are part of a functional system, a system that underlies and structures the social practice of language use. I will argue both that this theory of meaning is coherent as well as necessary for the conception of thought developed in the LTT. Chapter Four is concerned to explicate a number of details of Sellars' philosophies of mind and language, as well as to argue that the theories of most modern representationalists miss an essential feature of thoughts. In Chapter Five I will bring the pieces of all the preceding chapters together to spell out the fundamental details of the LTT. I see the LTT as a way of

applying the insights of Sellars (and others) to the modern representationalist landscape in the philosophy of mind. Having cleared out the space for the LTT and presented the fundamental components of the view in detail, I will turn to brief consideration of a couple of larger objections that people might have (Chapter Six).

Before we get to all that, however: I said earlier that Chapter Two would be devoted to a detailed examination of the various representational theories of mind, and the LTT's place among them. In the remainder of this chapter I will try to make the case for representationalism generally, in part by looking at some non-representationalist theories of mind, and then by presenting a general positive argument for RTM.

1.1 The Case For Representationalism

As I said above, very broadly speaking, the representational theory of mind says that to have a thought is to have a representation occur in one's mind, or brain, or whatever (silicon chips, Martian physiology, etc. still being open possibilities). Depending on who's representational theory of mind you're talking about, the nature of those representations will differ wildly.[3] Yet all representationalists would, I think, agree with the basic idea that for *S* to think that-*p* is a matter of some representation that means *p* occurring in *S*. And this idea, representationalism, has come (taking the collection of its various specific forms) to be the dominate view in philosophy of mind today. As Patricia Kitcher has put it, "since the fall of behaviorism, virtually everyone sees human behavior as mediated by the internal processing of information" (1984, p.215). William Lycan even claims that representationalism is not only true, but fairly *obviously* true. He

[3] This is almost an understatement since the 'representations' we're talking about could be anything from ideas in Cartesian mind-stuff to patterns of neural activity in human brains. Representationalism, broadly speaking, is thus quite insensitive to much of the debate in philosophy of mind from Descartes onward.

writes: "Representationalism is not supposed to be a hunch, a bold speculation or a wild throw of the dice, but rather a truth that is pretty plain once a few not very controversial assumptions have been accepted" (1993, p.405).[4] I will talk about the assumptions that Lycan thinks lead to representationalism a bit later. In fact, Lycan's view is probably more controversial than the generic representationalist idea that I've stated above.

So, what is the case for representationalism? How might one argue generally for this view? To begin with, one thing that all representationalist theories of mind have in common is a realism with regard to the so-called propositional attitudes (belief, desire, etc.). That is, intentionality is taken in some sense to be a real property of cognitive systems. There are non-representationalist theories that are realist with respect to the propositional attitudes, of course (Hilary Putnam's 'direct realism' in *The Threefold Chord* seems to be of this kind), but this commitment to taking intentionality seriously does serve nevertheless to mark a very sharp distinction between representationalist theories and those theories of mind (e.g., eliminativism, behaviorism, instrumentalism) that do *not* take intentionality to be a real property of cognitive systems. So, part of the argument for representationalism generally will be making a case for realism with regard to the propositional attitudes.

As I just mentioned, though, this does not separate representationalism from non-representationalist theories that are also realist with respect to the propositional attitudes.

---

[4] Given the breadth of views covered by representationalism (see note 2, above) it's easy to see where Lycan and Kitcher are coming from. If representationalism can cover everything from Cartesian dualism to type-identity theory, it may seem that there aren't really any views that wouldn't (or couldn't) count as representational. Behaviorism, of course, is a non-representational view, but it is generally defunct. Thus, the best current candidate for a non-representational theory of mind (though its proponents might not like it being called that, since their view does away with 'minds' altogether) is probably eliminative materialism. More on this below.

Fortunately, such views are, it seems, rather rare. And this seems to be the case not simply because such approaches have, by some quirk of fate, never come into vogue, but rather because representationalism has simply seemed so *obvious*. What could be more clear than the fact that human beings have thoughts *about* our world, and that those thoughts are things *different from* the things that they are thoughts about? Ideas in my mind are not the same as objects in the world, even though my ideas are *of* those objects. Given this seemingly obvious fact, the question is simply: How is it that my thoughts, my ideas, come to be about the world? And the answer, which has, I think, seemed to most people equally obvious, is: My thoughts are *about* the world because they *represent* the world. How else could a thought in my mind come to be about objects or states of affairs in the world other than by representing those objects or states of affairs in some way?

Of course, seldom have people (even philosophers) put things this way (i.e., been explicit about calling our thoughts 'representations'). Nevertheless, there is little doubt that most philosophers have believed our thoughts to work in this manner. The lack of an explicit endorsement of representationalism is likely just a reflection of the fact that such an endorsement would have seemed utterly unnecessary. Again, how else could the ideas in our minds be about objects in the world other than by somehow representing them?[5]

There is one distinction that is important and relevant here, a distinction between two 'kinds' of representationalism. Though many, if not most, philosophers have been representationalists (in the broadest sense) about our awareness of the external world, as we might put it, many of these same philosophers have avoided representationalism when it comes to our awareness of our own mental states. Descartes is a perfect example.

[5] This, of course, is a very big 'somehow.' Much of the history of philosophy of mind is taken over by debates that are, in essence, arguments about the nature of our mental representations themselves.

There is no doubt that he treated all of our awareness of the external world as mediated by representations, but he just as obviously denies that our awareness of our own thoughts is representational. (This point clearly shapes his skeptical worries in the *Meditations*; and one might reasonably suppose that someone like Putnam, in decrying representationalism all together, is concerned to avoid this very skeptical trap, wherein our knowledge of our own thoughts is immediate and direct, but our knowledge of the world is only indirect, mediated by representations.) The kind of representationalism that is popular today, and the kind that properly characterizes the LTT is a thoroughgoing representationalism that treats *all* awareness, both of the world and of our representations themselves, as a representational matter.[6] That is, my awareness of the chair in front of me, and my awareness of my thought that today is Thursday, are *both* tokenings of representations in me. To the extent that 'direct realist' views are an attempt to avoid the Cartesian problem of unequal access to facts about the world and facts about one's own thoughts, the form of representationalism that I endorse here should be perfectly unproblematic.

I plan, then, to leave so-called 'direct' realist views alone, focusing instead on the debate over realism about the propositional attitudes. The sensibilities of the eliminativists and the instrumentalists seem to be the main hold-outs against representationalism these days. Eliminativists (e.g., Paul and Patricia Churchland) deny that minds contain, or consist of, representations for the simple reason that they deny that there are any such things as minds. If there are no minds, no intentional systems, no beliefs, no desires, then there are *ipso facto* no representations in minds, no intentional

[6] See Sellars' "More on Givenness and Explanatory Coherence." At the beginning of that essay Sellars makes this same distinction.

representations, and no propositional attitudes adopted toward various representations. The eliminativist will not countenance talk of minds, so he will not countenance talk of representational minds. Instrumentalists (e.g., Dennett—though, as we'll see, Dennett is much harder to pin down to a clear and simple view than are the Churchlands) are less severe, in that they do allow talk of minds—but that's all it is: talk. For an instrumentalist, we can talk about systems *as though* they have intentionality, *as though* they have beliefs and desires; but no system is *really* intentional, no system *really* has beliefs and desires. Or, rather, there's simply no sense to the question of whether a system is really intentional or not: it's all a matter of the stance we take toward that system for the purposes of explaining and predicting its behavior. Since intentionality is merely a matter of interpretation, though, there is no role to be played by representations. If a system had representations, they would seem to be independent of our interpretive stance toward that system. Since, for an instrumentalist, there is nothing (as regards intentionality and the mental) that is independent of our interpretive stance toward a system, there is no stance-independent role for representations to play (i.e., the idea that real beliefs, etc. are constituted by actual tokens of distinct representations is rejected).

1.11 Eliminative Materialism

Eliminative Materialism is a view that arose primarily because of the advances made in neuroscience in the latter half of the Twentieth Century. For those philosophers of mind looking for a physicalist theory without the dramatic failings of behaviorism, eliminativism promised to provide everything that we could want in a theory of mind, and more. The central insight, according to its proponents, of an eliminative materialist

view is that philosophy of mind remained tethered to what eliminativists saw as simplistic and outdated 'folk' psychological concepts. Concepts like 'belief', 'desire', 'intention'—to the eliminativist these are primitive ways of referring to complex activities of the human brain. But if that is so, says the eliminativist, then why shouldn't we simply do away with those outdated concepts all together? Why not replace folk psychology with neuroscience (and the scientific psychology that it would engender)? Of course, it was always recognized that the science of the brain was still too much in its infancy to immediately replace ordinary ways of talking about the mind. Nevertheless, the eliminative materialists saw strong reasons to think that eventually neuroscience would be in a position to take over, if not in everyday contexts (though that would presumably follow in time), at least in philosophical and scientific contexts.

As I said above, eliminative materialism is one of the primary contenders against representationalism today. And perhaps its most famous proponents are Paul and Patricia Churchland. What I will do here is take a look at a typical line of argument from the eliminative materialist's perspective against folk psychology. I happen to think that folk psychology gets things more right than not, but the point of the discussion here is simply to show that the eliminativists, for all their insights, don't have a strong enough case to convince us to embrace the radical change they endorse. I will use as my target a typical argument for eliminative materialism as offered by Paul Churchland.

Churchland's endorsement of, and argument for, eliminativism is based for the most part on the supposed failures of so-called folk psychology. According to Paul Churchland, our common, everyday belief-desire psychology forms a theory of human minds that has proved largely inadequate in explaining and predicting human behavior,

and that has been stagnate for over two thousand years. These are, of course, contentious claims, as is (for some people) the claim that folk psychology (belief-desire psychology; common-sense psychology; hereafter simply 'FP') is an empirical *theory*. I am not going to rehearse Churchland's arguments for the theory-hood of FP; suffice it to say that he finds parallels between FP and other empirical theories (both scientific and common-sense). The parallels are so strong, in Churchland's view, that he concludes that not "only is folk psychology a theory, it is so obviously a theory that it must be held a major mystery why it has taken until the last half of the twentieth century for philosophers to realize it" (1981, p.186). I'm not particularly troubled by the idea that FP forms a common-sense empirical theory about human minds, but I disagree with Churchland that FP is so radically false that we must reject its categories all together. That is, I do not think that Churchland has made the case that there are really no such things as beliefs and desires, or intentional systems generally. Let us look briefly, then, at his arguments for anti-realism with regard to the propositional attitudes.

As I said above, Churchland's primary reasons for rejecting FP are its supposed explanatory failures and its general stagnation. He also believes that FP fails to cohere with the rest of what we know about the world that we live in. Let me take these points in order. The first of these claims is perhaps the most contentious. How can Churchland claim that FP is generally a failure when it comes to explaining and predicting human behavior? If FP has anything going for it, its ability to help us explain the behaviors of our fellow human beings would seem to be a bright shining star; we've been using FP successfully for at least as long as recorded history. The ancient Greeks no less than us today seemed to enjoy great success when explaining someone's behavior by reference to

his beliefs, and predicting someone's behavior by reference to her desires. FP seems to work so well, and has worked well for so long, that—unless we're doing psychology or philosophy—we're generally completely unaware of the fact that we're using FP. As Fodor has said: "Commonsense psychology works so well it disappears" (1987, p.3). Consider the following story, also from Fodor:

> 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. Or if I *don't* turn up, it's less likely that the theory has failed than that something went wrong with the airline. It's not possible to say, in quantitative terms, just how successfully commonsense psychology allows us to coordinate our behaviors. But I have the impression that we manage pretty well with one another; often rather better than we cope with less complex machines. (1987, p.3).[7]

This story, and countless real episodes like it that happen every day, suggest that FP is really quite successful at predicting and explaining behavior. But Churchland sees things otherwise. He does not deny that FP enjoys some success; but he says that "we must reckon not only with FP's successes, but with its explanatory *failures*, and with their extent and seriousness" (1981, p.187, emphasis added).

Here Churchland provides what he takes to be a damning list of FP's failures, of "central and important mental phenomena" that remain unexplained (and are perhaps, though Churchland does not argue for such a claim, unexplainable) by FP:

> the nature and dynamics of mental illness, the faculty of creative imagination, or the ground of intelligence differences between individuals

[7] Notice, as an aside, that Fodor freely refers to FP as a 'theory.' Like me, Fodor would seem not to object to this part of Churchland's position.

> ... [;] the nature and psychological functions of sleep ... the common ability to catch an outfield fly ball on the run, or hit a moving car with a snowball ... [;] the internal construction of a 3-D visual image from subtle differences in the 2-D array of stimulations in our respective retinas ... the rich variety of perceptual illusions, visual and otherwise ... [or] the miracle of memory, with its lightning capacity for relevant retrieval. On these and many other mental phenomena, FP sheds negligible light. (1981, p.187).

Yes, FP works passably well in our day-to-day dealings with one another; but according to Churchland it has also failed to shed any light on a great number of important mental phenomena. As a theory, we must evaluate it not only on its successes, but on these failures as well. Personally, I believe that this criticism is somewhat misguided, but let's look at Churchland's other complaints about FP before responding.

The supposed explanatory failures of FP are only part of the picture that Churchland is trying to paint. By themselves these failures might not even be all that significant (since no other psychological theory can explain all the phenomena Churchland mentions either). Churchland, however, combines these failures with the fact that FP has been around for millennia, and *still* has not been able to explain the phenomena in question. In fact, according to Churchland, FP has not noticeably matured *at all* in the entirety of its history. And that is strike two, as it were, as far as Churchland is concerned. The fact mentioned above that the folk psychology that we use today is more-or-less the same folk psychology that the ancient Greeks got along with, and that "we are negligibly better at explaining human behavior in its terms than was Sophocles" (p.188) is considered, by Churchland, to be evidence of the extreme stagnation of FP. This is, he writes, "a very long period of stagnation and infertility for any theory to display, especially when faced with such an enormous backlog of anomalies and

mysteries in its own explanatory domain" (p.188). Of course, if FP were a perfect theory, it would have no need to grow and evolve, but "FP is profoundly imperfect" according to Churchland, so its "failure to develop its resources and extend its range of success is therefore darkly curious, and one must query the integrity of its basic categories" (p.188). Given the explanatory failures that FP supposedly has, its additional failure to produce any new insights into human psychology, or to enhance its own explanatory or predictive power, over the last two or three millennia (at least) calls its basic concepts into question.

Nevertheless, if FP were the only theory of mind that fit well with "other theories about adjacent subject matters" or held the promise of being successfully reduced to more basic theories, it might still merit "patience and solicitude" (p.188). That is, current explanatory success is not the only reason to hold on to a theory; future promise can be a balancing virtue for a currently stagnated theory. But here, too, Churchland finds FP lacking, for far from lending itself to theoretical integration with the rest of science, FP's "intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus" (p.188) of scientific explanation. According to Churchland:

> If we approach *Homo sapiens* from the perspective of natural history and the physical sciences, we can tell a coherent story of his constitution, development, and behavioral capacities which encompasses particle physics, atomic and molecular theory, organic chemistry, evolutionary theory, biology, physiology, and materialistic neuroscience. That story, though still radically incomplete, is already extremely powerful.... In short, the greatest theoretical synthesis in the history of the human race is currently in our hands.... (P.188).

And folk psychology is just not part of this synthesis. Churchland does not completely rule out the prospect of a successful reduction of FP to this large synthesized body of scientific knowledge in the future, "but FP's explanatory impotence and long stagnation

inspire little faith that its categories will find themselves neatly reflected in the framework of neuroscience" (p.188). Without the prospect of future success, FP's failures and stagnation are, as far as Churchland is concerned, fatal.

Hence, eliminative materialism: the view that folk psychology is a radically false theory with no prospect of having its categories reflected in some future scientific psychology. Rather than adapting our theories of mind to the concepts of FP (intentionality, belief, desire, etc.), we should eliminate those concepts from our theorizing about minds, replacing them (presumably) with concepts from current, or future, neuroscience. And, again, if there are no such things as beliefs that-*p* (nor even other things which beliefs that-*p* reduce to), then there cannot be any such thing as standing in the *believing* relation to a *representation* that means *p*. Representationalism is incompatible with eliminativism for the simple reason that representationalists are invariably realists about the propositional attitudes—and eliminativists are not.[8]

How, then, should the representationalist respond to someone like Churchland? One way would be to reject the claim that FP is an empirical theory; if it's not an empirical theory then it is not subject to the kinds of defeating criticisms that Churchland proceeds to level at it. That would be one way to go, but as I said above, I have no problem with the claim that FP is an empirical theory. Rather, I would suggest that, while the criticisms that Churchland aims at FP are valid criticisms, they are not as

[8] Notice, of course, that there is no similar incompatibility between representationalists and *physicalists*. In fact, most representationalists these days probably *are* physicalists: they believe that the concepts of FP *do* reduce to concepts of the physical sciences (neuroscience, biology, physics, etc.). The familiar mental states of FP (beliefs, etc.) are considered real states of the cognitive system in question; and the assumption is that their description at the representational level will simply be reflected at the level of the physical sciences. Thus, obviously, realists about the propositional attitudes needn't believe in Cartesian spook-stuff.

defeating (nor nearly as damning) as Churchland takes them to be. Let's reconsider those criticisms.

I think that it's best to respond to Churchland's criticisms in reverse of the order that he presents them (and of the order in which I presented them, above). We begin, then, with the third criticism, that FP does not fit into the modern synthesis of scientific knowledge about the world, and that it shows no promise for future integration via a successful reduction of the concepts and categories of FP to those of the physical sciences. With the first part of this criticism I have no real disagreement: considered merely as concepts of FP (i.e., considered apart from any physicalist belief that the concepts of FP *can* and *will* be reduced to concepts of the physical sciences), things like beliefs, desires, and intentional mental states do not seem to fit into the 'great synthesis' that Churchland sees in modern science.[9] But recall that Churchland's only real basis for the second half of his third criticism, that there is little chance that FP will ever be successfully reduced, are his first two criticisms of FP. That is, he sees little reason for optimism regarding a future reduction of FP to the physical sciences precisely because he thinks that FP is so utterly ineffectual, and has been stagnant for so long. Hence, if we can show that these criticisms are not well founded, we will have disposed of Churchland's reasons for thinking that FP will never be successfully reduced. That, then, is precisely what I aim to do now.

The main point of my reply to Churchland's first two criticisms of FP is quite simple: I think that he has unfairly characterized the state and history of that theory. The

---

[9] One might also question whether there really is this 'great synthesis' that Churchland sees in modern science, but I'm happy to suppose that there is, and so will not push him on that point. (For an example of someone who *does* push Churchland on this point, see Patricia Kitcher (1984, p.222).)

modern scientific method is a relatively new invention, and its application to human psychology is an even newer endeavor—perhaps less than a century old. Thus FP as a science has not been stagnant for very long, if at all (and it certainly has not been stagnant for millennia, as Churchland claims). Nor is FP's list of things that it has failed to account for as damning as Churchland claims—FP's failure to explain, e.g., the faculty of creative imagination is no more problematic for psychologists than quantum physics' inability to reconcile general relativity with quantum mechanics was for physicists. What Churchland has done is unfairly apply modern criteria of theory evaluation to a theory that was not recognized nor treated as such for most of its history. Let me elaborate a bit.

I have no doubt that the FP that we use today is substantially the same FP that people have used for thousands of years; on that point I agree with Churchland. Churchland's mistake comes in supposing that because FP is an empirical theory it must have been treated as such for the entirety of that long history; it has not. To use Fodor's way of putting it, again: "Commonsense psychology works so well it disappears" (1987, p.3). That is to say, because FP has seemed so natural and so obviously true, it has never really been treated as a hypothesis that was open for question—it has not been treated as a hypothesis at all. FP simply has not seemed like a *theory* to most people for most of history. Partly for this reason FP has not, until recent history, come under any sort of scientific scrutiny; nor has it, again until recently, been an area for serious-minded research programs. The so-called stagnation that Churchland points to in his second criticism of FP is, therefore, not well founded. One could hardly expect a theory which was not recognized as a theory to have experienced any substantial scientific growth—since FP was not recognized as a scientific theory it was not subject to the kinds of

critical examinations that lead to important theory development.[10] Again, the modern scientific method is also a fairly recent invention, and so the types of theory critique and development that have contributed to the growth of other, more recent, empirical theories (in the physical sciences, especially) were simply not available for most of FP's existence. Even more importantly, though, since FP has not been viewed as a theory by anyone until even more recently (as Churchland himself admits—though he takes that fact to be a mystery and an embarrassment to philosophy) the time frame in which FP has been subject to the type of scrutiny bestowed upon other empirical theories is short indeed. (Recall: according to Churchland, it was not until the "last half of the twentieth century" (p.186) that anyone 'realized' that FP was, indeed, a theory; FP's 'stagnation,' then (if it has been stagnant), has really only lasted a few decades—and I'm not sure that even Churchland could seriously argue that there has been absolutely *no* growth in FP in the last half century. I think that psychologists and sociologists really have made some progress in understanding the human mind—progress which has in turn filtered down into FP; though admittedly scientific psychologists today lack a solid philosophical foundation on which to work.) So much, I say then, for Churchland's second criticism; what about the first, the long list of FP's supposed failures?

My response to Churchland's first criticism builds upon my response to his second criticism. Given that the period of FP's treatment as an empirical theory, and as an area for scientific research programs, is, at best, limited to the last fifty-or-so years, the fact that there remain a large number of phenomena that it cannot (yet) explain is hardly

[10] This is not to say that FP can't have made progress; I think, in fact, that it probably has. But any progress it made would be likely to occur slowly, in the background, as everyone went about using the concepts of FP without questioning their integrity. So, again, it is not surprising, to my mind, that FP wouldn't have experienced any substantial *scientific* growth.

surprising. Churchland would have us believe that FP's failure to fully explain mental illness, imagination, intelligence differences between individuals, hand-eye coordination, perceptual illusions, sleep, memory, etc. is reason to reject the theory completely (to eliminate it). Yet, what theory, in its scientific infancy, was free of a comparable list of unexplained phenomena assumed to be within its explanatory domain? Physicists have been working on various problems in their field for hundreds of years, and still there remain physical features of the universe that they cannot explain. No one takes this as a reason to abandon physics (to eliminate it) outright. Why should psychology be any different? If Churchland is right and FP is a theory, it is a theory that we have only very recently begun to explore in earnest. I agree with Churchland that there are many elements of human psychology that FP cannot currently explain, but unlike Churchland I am optimistic that a deeper understanding of these various elements will be integrated into FP in the future. We need only give the psychologists, philosophers of mind, and researchers in adjacent fields (neuroscience, linguistics, and the cognitive sciences generally) time to do their jobs.

So, to summarize: As an empirical research program, FP is still very young; there are many phenomena that it cannot yet explain, but that is to be expected. Since it is very young, there has not been any great period of stagnation as Churchland would have us believe. And since there is no stagnation to point to, the prospect for a successful integration of FP with the rest of our empirical theories about human beings and the

world around us is still well intact. All of Churchland's objections are based on an unfair mischaracterization of the history and nature of FP as an empirical theory.[11]

I want to be clear at this point, however, that I don't think 'folk' psychology holds the strongest promise for the future. Insofar as FP is just our 'ordinary way of talking' about the mind, it isn't going to be much better than 'folk' physics (say). My point, rather, is that I think we have good reason to believe that scientific psychology will ultimately reflect and incorporate the basic concepts of FP. The sheer volume of successes that FP has enjoyed over the millennia (think of Fodor's story again) are the strongest reasons to suppose that we are warranted in being realists about the propositional attitudes. Even some philosophers who, as we'll see in just a moment when we look at instrumentalism, are anti-realists about the propositional attitudes nevertheless recognize the practical indispensability of the concepts of FP. In the absence of any good reasons to reject those concepts, then, I think that we are justified in remaining realists. And eliminativism simply fails to offer a compelling case to abandon the propositional attitudes.

1.12 Instrumentalism

If eliminative materialism leaves us unconvinced, then, what of instrumentalism? The instrumentalist believes that there are really no such things as beliefs and desires, not because he finds FP to be bankrupt, as the eliminativist does, but rather because he takes ascriptions of beliefs and desires (etc.) to be simply a matter of taking up what Dennett has called the 'intentional stance.' That is, to ascribe, e.g., a belief that-*p* to some system

[11] William Lycan, in a footnote in his "A Deductive Argument for the Representational Theory of Thinking" (1993, note 14, p.413), offers a reason for rejecting eliminativism that is somewhat similar to the way I have reasoned against eliminativism here.

is not, on the instrumentalist's view, to describe a real property of that system.[12] Rather, it is to take up a particular stance toward that system for the purposes of explanation and prediction. We ascribe the belief that-*p* to the system, on this view, because doing so makes it easier (perhaps) to explain the behavior of the system. As a matter of fact, though, this is only one among a number of different 'stances' that we could take toward the system in question. We could, for example, describe the system from the physical stance instead, predicting and explaining its behavior by appeal to the concepts of the physical sciences. In many cases (e.g., when dealing with thermostats) the physical stance will be just as informative, and just as easy to use, as the intentional stance. As the system becomes more complex, though, we might find ourselves pushed toward using the intentional stance. For an instrumentalist, this is the primary thing in FP's favor: so far, the only generally successful and user-friendly way to predict and explain *human* behavior is with the concepts of FP.

FP may be generally successful and user-friendly, but it is ultimately (at least in theory) unnecessary on an instrumentalist view. To see why, we need to elaborate on just how instrumentalism is supposed to work. The following picture is drawn from Dennett's work, as he is both the originator of, and most significant philosopher to endorse, instrumentalism—though, again, Dennett himself is rather cagey about his view, so much so that the instrumentalism described here would probably not be endorsed by Dennett (at least, not without a lot of caveats); and to my knowledge, Dennett never calls

[12] Dennett, though seemingly an instrumentalist, is somewhat cagey on this point. He claims that he is *both* a 'realist' and an 'interpretationist,' positioning his view "on the knife-edge between the intolerable extremes of simple realism and simple relativism" (1981, p.238).

his view 'instrumentalism.'[13] The clearest places that one may find the view described below are Dennett's essays, "Intentional Systems" and "True Believers: The Intentional Strategy and Why it Works".

As I said above, the basic idea at work in the instrumentalist picture is the notion of adopting 'stances' toward a system for the purposes of explaining and predicting that system's behavior. When one adopts a particular stance toward a system, that system can then be said to be a certain kind of system. So if, for example, one adopts the intentional stance toward a system, that system can then be said to *be* an intentional system. We will be looking at some further details here, of course, but it is important to keep this simple fact in mind: at the heart of instrumentalism is the view that being a system of a certain kind is nothing more than having someone take up the appropriate stance toward that system for the purposes of explanation and prediction. That said, what kind of stances are we talking about, and what does it really mean to 'take up a particular stance' toward a system? In the two essays by Dennett mentioned at the end of the previous paragraph, there are three stances that Dennett considers: the physical stance, the design stance, and the intentional stance. I see no reason in principle why there couldn't be more than these three, though in practice, these three seem to cover things pretty well—and because we don't really need any more examples for our purposes here, I will limit myself to explaining just these three stances.

Let's begin with the physical stance. What would it mean to 'adopt the physical stance' toward some system? Quite simply, it would mean choosing to describe that

---

[13] In fact, it took me quite a while to finally determine that Dennett's view basically *is* instrumentalist, in no small part due to the fact that Dennett constantly takes back with one hand what he has just given with the other. More on this below.

system in the language of the physical sciences. That is, we would choose to explain and predict that system's behavior in purely physical terms. We might explain, say, the behavior of an old grandfather clock by talking about how the various parts interact on a physical level: the way that the gears, etc. exert certain forces on each other causing this and that part of the clock to move in such and such ways. Or (here I borrow an example from Dennett) we might describe the behavior of a chess playing computer by talking about the flow of electrons through various circuits, etc. If we chose to limit our talk in this way, i.e., to talk only at the physical level of description, then we would be taking the 'physical stance' toward the clock, or the computer. The system in question then qualifies as a physical system.[14]

So far this seems pretty straightforward (not counting the complication noted in the last footnote). And the physical stance seems like a fairly effective stance to take when attempting to predict and explain the behavior of any number of real-world systems. Sometimes, though, the physical stance either seems impractical, or seems to miss what we might take to be some important features of the system in question (or both). Take the grandfather clock, for example. A description of its behavior in physical

---

[14] This may all seem well and good on the surface, but there is one thing here that has always bothered me. As will become more apparent below, Dennett seems to be suggesting with his 'stance' talk that *all there is* to being a system of a given kind is having someone take up the appropriate stance toward that system. But, while this may seem somewhat plausible with regard to the design and intentional stances, I confess that it has always sounded very odd to my ears to suggest that something's being a *physical* system is tied to the explanatory and predictive purposes of beings such as ourselves. This sounds odd, of course, because most (if not all) of us believe that the physical world would remain as it is regardless of whether we, or anything, were around to treat it as a physical system. And I have always had a hard time believing that Dennett would deny that, say, a clock was a physical system just because no one had chosen to take up the physical stance toward it. In many ways, this is the same kind of objection that will seem so intuitively obvious (though ultimately question-begging, as we'll see) when we reach the intentional stance. I do not know how to resolve this issue. It seems fairly clear that Dennett means for us to take his 'stance' talk seriously, and doing so seems to mean that a system is a physical system if and only if someone adopts the physical stance toward it. While this issue is puzzling, at the least, it is fortunately not ultimately relevant to the overall purpose of this discussion, and so this footnote will be all I have to say on the matter.

terms will explain well enough why certain gears turn and levers move—the physical stance will tell us why these things happen, in terms of the mechanics of the clock. But the physical stance will *not* tell us why these things happen in terms that are appropriate for discussing a timepiece. That is, the physical stance misses the fact that the clock has a particular purpose to its design. This is where the design stance comes in. When we move to the design stance, we attempt to predict and explain a system's behavior not in terms of the physical 'nuts and bolts' as it were, but rather in terms of what the system is meant to do. Rather than talk about gears and the like, we'll start talking about hands of the clock, pendulums, chimes, and numbers written on the face of the clock. Why does the clock chime so when both of its hands are pointed at the twelve? Because the clock is designed to mark the hour. There is, of course, a purely physical explanation for the chiming, but that explanation misses something important about the system: that it is a clock meant to allow us to keep track of the time.

Certainly, we gain no predictive power, and very little explanatory convenience, by adopting the design, rather than the physical, stance when dealing with the clock. We seem to gain an important new perspective concerning the clock's behavior, but that's about it. Consider, however, the other example: the chess playing computer. An expert in physics, electronics, and computer design might very well be able to explain exhaustively, and predict flawlessly, the behavior of the computer—but Oh! what a task. Explaining why the computer moved its queen in such-and-such a way after I moved my king the way I did would take, with any modern computer, a massive amount of time and knowledge, covering everything from the physical inputs to the system, the make up of the circuit boards, the movement of electrons through the circuits, and the resulting state

of the machine. Yet, there is a much simpler way to explain the behavior of the computer: it moved its queen because it was designed to play chess, and that is the move that the program, the algorithms the machine was built to calculate, came up with. This will vastly simplify what we need to know and say about the computer to explain and predict its behavior. Even at this level, though, the explanation will require some very complex details—the sorts of details that the computer programmers deal with when they build chess playing computers. Still, the design stance seems to recommend itself in these cases (the clock and computer), whether it simplifies our explanations or not. If the design stance leaves us with a barely manageable complexity concerning the computer (manageable for the programmers, say, but not for the ordinary chess player), as we begin to consider even more complex systems, even the design stance begins to fail us. Even if we had all the relevant information, it seems likely that taking the design stance toward the behavior of, say, a human being would quickly swamp us in complex explanatory acrobatics (since the behavior of human beings is vastly more complex than that of a chess playing computer—in fact, it's not even entirely clear what sorts of things we'd have to take into account about the 'design' of a human being to explain even the simplest of her behaviors).

This is where the intentional stance comes in. Like the physical stance and the design stance before it, the intentional stance is merely a matter of choosing to describe a system in a particular way—in this case, in intentional terms, in the language of folk psychology and the propositional attitudes. In short, rather than talk of gears and levers, or of functions and programs, we talk of beliefs and desires (etc.). The predictive and explanatory power that we gain, the instrumentalist tells us, is in many cases immense.

Let's start simply, and consider the chess playing computer again. As we noted, the design stance seems to offer quite a bit of convenience over the physical stance in this case, but we also noted that the amount of information, and the time it takes to convey it, are still very limiting in practice. Sure, the programmer could walk us through the code, revealing to us the sequence of events that led to the move we want explained, but there is a vastly easier way to go about it. The computer moved its queen because it *wants* to win, and it *believes* that moving its queen just so is the best way to achieve that goal. As Dennett has said, if you want to beat any modern chess playing computer in real time, your best bet is to treat it as an intentional system and figure out what it believes and what it's trying to do.

A chess playing computer is nothing, though, next to the complexity of an adult human being. If you want any chance at success in explaining what I've just done, or in predicting what I might do next, you had better adopt the intentional stance toward me and take into account my beliefs, desires, etc. FP isn't merely a convenience when dealing with people, says the instrumentalist, it's a practical necessity.

This is not to say that one couldn't, theoretically, adopt the physical or design stances toward human beings (Dennett has an argument about why doing so would be to miss something important about us ('real patterns' he calls them), but in theory he still believes that, as far as prediction and explanation go, one would lose only convenience and brevity by resorting to the design or physical stance). For the instrumentalist, beliefs and desires are not interpretation-independent properties of any system; talking about such things is merely a convenience. The world isn't divided into things that are really

intentional systems and things that aren't.[15] Being an intentional system just means being a system that someone has taken up the intentional stance toward. To illustrate with a famous (or infamous) example from Dennett: while we can perfectly well explain the behavior of a thermostat from the physical stance, and even more simply from the design stance, there is no reason in principle that we couldn't decide to adopt the intentional stance toward the thermostat. If we do so, we will ascribe to it a probably quite limited set of beliefs (e.g., the room is too cold, the room is too hot, or the room is just right)—but (and this is the kicker) these beliefs will be no different in kind from the beliefs that you and I have. The thermostat will be a very simple intentional system, with a very limited set of intentional states, but it will count as intentional in the very same way that a thinking adult human being does.

I think that there is a temptation, upon hearing this last bit, simply to reject instrumentalism outright, for the conclusion that thermostats have beliefs just like we do can seem so obviously absurd as to appear to provide a *reductio* of the view itself. I admit I have great sympathy for this response, but I think that it is ultimately mistaken—not because I believe that instrumentalism has got things right, but because I believe the supposed *reductio* to be question-begging. This gut response, while seemingly quite obvious on its surface, rests, upon further reflection, on the presupposition that while we *really* have beliefs, thermostats do not—that is, on the presupposition that there are real, concrete things in the world called 'beliefs', and thermostats just don't have them (however they (beliefs) may ultimately be constituted). Yet this is just what

[15] This will come up again below, but let me add a point of clarification here. Instrumentalism in the philosophy of mind is clearly a metaphysical, not an epistemological, position. That is, the instrumentalist is not simply agnostic about beliefs and desires, he claims that there simply are no such things, not really. I would like to thank Andrea Woody for calling my attention to the need to make this point explicit.

instrumentalism denies. Instrumentalism is quite clearly an anti-realist position with regard to the propositional attitudes (though, again, Dennett's view is less clearly anti-realist than the basic position described here[16]; I ultimately believe his view to be more instrumentalist than not, however). And we cannot argue against such a position by assuming what it denies.

Argue against the position we must, however, for I am trying to make a case for RTM and, as I've said before, all representationalists of whatever stripe are realists about the propositional attitudes. My argument against instrumentalism, though, will be less of an attack on the positive arguments for instrumentalism (since there really aren't any—instrumentalism is more like a bold hypothesis, the proof of which is supposed to be in the light it might shed on problems in the philosophy of mind[17]) and more of a defense of realism about the propositional attitudes. To be fair, though, I do wish to begin by looking at some of the things that would seem to recommend instrumentalism in the philosophy of mind. In many ways (despite what I take to be an obviously absurd result, as noted above), I think that instrumentalism is a more powerfully compelling view than eliminativism. At the very least, instrumentalism is a much more slippery foe.

I suppose that different philosophers will take different features to be virtues of instrumentalism (or any theory of mind, for that matter). For my part, though, the primary virtues of instrumentalism are its complete compatibility with naturalism, and the powerful simplicity with which it seems to solve the mystery of mind. I do not wish to

---

[16] Lycan has even suggested that Dennett is in fact moving in a realist direction because he has apparently come to be comfortable (as evidenced by comments in his "Real Patterns) with allowing propositional attitudes to be causes in some sense (Lycan, 1993, p.419, note 25).

[17] It is worth noting that I have no problem in principle with this way of arguing for a theory of mind. For the most part, I think, the LTT is also a bold hypothesis of this kind. It is the benefits, actual and projected, of such an hypothesis, as well as its resistance to criticisms, that determine its ultimate success. Instrumentalism, as I think we will see, ultimately fails because it cannot deliver on its promises.

dwell overlong on the supposed virtues of instrumentalism, however, since ultimately on my view the anti-realism of instrumentalism is fatal. Thus, I'll just say that as a competitor to representationalism, instrumentalism can seem very appealing to those who do not wish to countenance 'sentences in the head,' as it were. First, instrumentalism is fully compatible with a naturalistic sentiment; there is no need, on an instrumentalist view, to even attempt to fit FP into the modern scientific synthesis that Churchland mentions, because there are no actual properties of the system, no actual states of a minded being, to synthesize with science. All FP comes to is a convenient way of talking. (Of course, things aren't really this simple; there are complications that one could explore, but these would take too much space here, while remaining ultimately irrelevant.)

Eliminativism also has the virtue of being compatible with naturalism, of course—in fact, as we saw, an appeal to naturalism is one of the main arguments eliminativists turn to in defending their view. Eliminativism is anything but a simple solution to the problems of mind, though. Eliminating all talk of beliefs, desires, intentions, and other mental states in favor of the language of, say, cognitive neuroscience as Churchland recommends is hard even to imagine, especially with the current, rather limited state of neuroscience. Instrumentalism, on the other hand, fully embraces the powerful simplicity of FP. Yet, because things like beliefs, on an instrumentalist view, are not 'real' entities, but merely explanatory conveniences, many if not all of the classic problems about minds vanish. We no longer need worry about the relation between mind and body, about the causal powers of mental states, about how some systems can be intentional while others are not, how thoughts fit into the natural

order, what makes thought possible in the first place, nor about the actual constitution of mental states. Or so it would seem.

I think that the apparent simplicity of instrumentalism is ultimately its downfall (which may be one reason that Dennett's version is so hard to pin down: by his own admission Dennett's view rests upon a 'knife edge' between incompatible alternatives—a rather unstable position, I'd say). My objection to instrumentalism is quite straightforward, but completely fatal I think. Simply put, I believe that instrumentalism ultimately fails to explain anything. The problem is that the instrumentalist simultaneously denies the reality of the propositional attitudes and offers nothing to take their place. The eliminativist denies the reality of the propositional attitudes, but she offers up a complex scientific alternative, so that eliminativists still have plenty to say about why minded creatures behave as they do. The instrumentalist, on the other hand, insists on the continued use of the language of FP, but robs it of all content. When we explain someone's behavior by appeal to his beliefs and desires, on the instrumentalist picture, our 'explanation' has no explanatory power. What good is it to say that I behaved as I did because of what I believe if my 'having' beliefs in the first place is merely a matter of someone taking the intentional stance toward me, and not an interpretation-independent fact about me? Let me spell this worry out.

Suppose that Jane grabs an umbrella before going outside. The philosopher of mind wants to explain this behavior. If we endorse FP, the explanation goes like this: Jane believes that it is raining, and desires to stay dry. We might add that she also believes that the umbrella will keep her dry, etc., but to keep things simple we'll stick simply to her belief that it's raining and her desire to stay dry. So far, so good. Now we

ask, What are these things we're ascribing to Jane, this 'belief' and this 'desire'? For simplicity's sake, let's rule out the Cartesian-type answers—that is, the theories that are incompatible with naturalism. The important question then becomes one of the constitution and nature of the intentional states ascribed to Jane. The possible answers are, of course, myriad. The answer that I will be endorsing in the rest of this document, for example, says (very roughly) that the states ascribed to Jane are brain-state symbolic tokens of the language that Jane speaks—representations with functionally defined semantic contents tokened in certain specific ways. (I'm not going to worry here, of course, how much of an explanation this ends up being. We'll have plenty of opportunity to probe the details of the LTT later.) Right or not, this is at least an explanation that tells us something of the nature and constitution of Jane's belief and desire, something that helps us to understand how that belief and that desire work to cause the behavior in question.

What, then, is the instrumentalist's answer? What, on the instrumentalist view, is the nature of Jane's belief and desire, such that reference to those intentional states explains Jane's behavior? As we've noted, on the surface the instrumentalist's answer seems quite simple, and quite powerful. Being an intentional system is simply a matter of having someone take up the intentional stance toward that system. When we say that Jane has the belief that it's raining, we are choosing to take up the intentional stance toward Jane, for the purpose of explaining and predicting her behavior, and so we use the language of FP. So, what is the nature of Jane's belief? Well, the question makes no sense on the instrumentalist's view, because 'belief' is just a word that we use when

we've chosen to take up the intentional stance. It has no referent; thus there is nothing to which it refers, the nature of which could be subject to investigation.

Now, however, I think we're forced to ask: What, then, makes this an explanation of *why* Jane's belief contributes to the explanation of her behavior. The instrumentalist has told us that we can explain Jane's behavior by talking about her beliefs and desires, but when we ask what these 'beliefs' and 'desires' are, we are told that they aren't anything at all—that such talk is merely a convenient way of talking about why some system does what it does. But we already knew that; we were taking for granted that FP helped explain behavior. What we wanted to know was *why* FP helped explain behavior. Instrumentalism doesn't tell us that; it just reiterates that FP does help explain behavior. The further 'why' question is ruled out of bounds by the theory. Yet surely the emptiness of the resulting account is becoming apparent. How does instrumentalism propose to explain Jane's behavior? By claiming that Jane's behavior is explainable by adopting the intentional stance toward her? That's no answer at all. This isn't to say that instrumentalism says nothing at all—just in endorsing FP, instrumentalism says something about how we can explain Jane's behavior. What instrumentalism fails to do, though, is say anything about the question we're actually interested in.

This isn't, of course, an argument for realism with respect to the propositional attitudes—an argument of the sort that could be used (and has been used) against the eliminative materialists as well. Rather, all I'm saying here is that if one is going to deny the reality of the propositional attitudes, one must offer some alternative to take their place. Otherwise we bring explanation to an end too soon. The instrumentalist explains

Jane's behavior by appeal to beliefs and desires, adding the redundant claim that beliefs and desires are explanatory without telling us how or why they are explanatory.

If we had the space for it, we could of course look at independent arguments for realism about the propositional attitudes, but I think our time will be better spent at this point if I turn to an argument for RTM generally—to a positive argument for RTM itself, rather than an argument aimed at knocking down one of the main alternatives to RTM. Combined with the attack on instrumentalism in this section, and the attack on eliminativism in the previous section, the following discussion should give us a decent case for RTM—which is the primary purpose of this chapter.

1.13 Lycan's Deductive Argument for RTM

In "A Deductive Argument for the Representational Theory of Thinking" (1993), William Lycan presents a, in many ways quite straightforward, positive argument for representationalism. Lycan's argument may or may not work as advertised (there is a comment on it by Robert Stalnaker, "What is the Representational Theory of Thinking? A Comment on William G. Lycan" (1993), immediately following Lycan's piece that raises a few difficult and probing questions), but I am more interested in certain insights of Lycan's essay than in the ultimate success of his "deductive argument." So what I will do in this section is explore those insights and argue that they present some very strong points in favor of RTM. I should also note up front that I believe that the nature of Lycan's argument, if not the exact details, reveals something important about how many representationalists understand their view (which is not the same as saying that I think all,

or even most, representationalists would endorse Lycan's argument). Of course, I will also attempt (briefly) to respond to a few objections, as they arise.

In brief, then, the central insights (to my mind) of Lycan's essay are the following. First of all, Lycan draws attention to a feature of thoughts that he calls the 'unboundedness of thinking' which parallels a similar feature of language emphasized by Chomsky. The unboundedness of thinking is just the fact that human beings have the ability to think any one of an infinite number of thoughts. Lycan is hardly the first representationalist to draw attention to this feature; as a matter of fact, the unboundedness of thinking is quite often one of the main reasons cited by contemporary representationalists in favor of their theory. It is an important point, however, and so I will rehearse the discussion of it here. Secondly, and most importantly, Lycan presents RTM as a really fairly obvious theory, given a few assumptions. There is a minor point of interest to this (namely, being clear about, as Lycan puts it, the "epistemic status" that representationalists attach to their view), but there is also a much more important element to Lycan's claim about the supposed obviousness of RTM. Naturally there are those who find RTM anything but obvious; perhaps the primary thrust of Stalnaker's response to Lycan is to call into question this supposed obviousness. And by Lycan's own admission, RTM is obvious only given a few assumptions that, while not taken by him to be particularly controversial, are nevertheless possible sticking points for the unconverted. Yet I do think that Lycan is on to something important when he claims that RTM is a fairly obvious theory. While I personally find RTM perhaps less obvious than Lycan, I do agree with him when he approvingly quotes Fodor as referring to representationalism as the 'only game in town'—i.e., that there are really no sensible

alternatives to representationalism.[18] This claim, of course, is quite strong indeed, but it is also, to my mind, the most important claim in Lycan's essay, so I will spend the largest part of this section considering it from different angles.

To begin with, though, let me quickly present the "deductive argument" itself. It will help move the discussion along to have Lycan's argument on the table to work with, but I also do this in part so that it will be easier to separate the insights of Lycan's piece from the specific way in which he has presented those insights. Here then is the argument[19]:

(1) Thinking is unbounded.

(2) Physicalism (of the token-identity sort, for Lycan) is true.

(3) If thinking is unbounded, then we need a finite stock of primitive elements and a compositional system of recursive rules for combining those primitive elements into whole thoughts, all of which (given (2)) must be physically realized.

Therefore,

(4) To have a thought is to physically token, in a certain specified way, a representation with a given semantic content, the entire tokening being governed by a system of rules that allows for the combining of atomic mental parts into whole thoughts. This is representationalism.

---

[18] Lycan eventually characterizes his "deductive argument" as in fact a 'special form' of inference to the best explanation: what he calls, 'inference to the only explanation' (which is supposed to explain why he can claim both that representationalism is an empirical theory and that it is a theory one can arrive at from the armchair). See, 1993, p.412. I should note here also that since Lycan's argument seems best characterized as an instance of IBE, it would not ordinarily be considered *deductive* at all, but Lycan claims that he views *all* deductive arguments as instances of IBE (see n.12, p.412)—which, I suppose, helps explain the title of the essay, despite the 'rhetorical retreat' on p.412 to calling his argument a special instance of IBE.

[19] I have modified the exact statements here to more clearly fit the needs of our discussion, but I do not believe that I have thereby in any way misrepresented, or done any injustice to, Lycan's argument.

The first thing that I would like to make note of here is that the representationalism of Lycan's conclusion[20] is not quite the representationalism I have in mind when I speak of representationalism generally. Specifically, the physicalist requirement seems to me unnecessary.[21] Of course, *modern* representationalists are most likely all physicalists, but one needn't be. This isn't really a problem, though: simply drop premise (2), and the following references to it, or to any physicalist requirement, and you'll have the same basic argument for a now non-specific (with respect to the question of physical versus non-physical) representationalism. This point is an important one, however, since once we drop physicalism from the argument we see that its essential elements are this: first, the claim that thinking is unbounded, and second, the conditional claim that if thinking is unbounded, some system of the representationalist sort[22] is required. The first part of this simplified version of the argument I will deal with in the next couple of paragraphs. The second part, the conditional, however, is the real meat of the argument, which we will come to presently.

The 'unboundedness of thinking' is, as I've said, simply the idea that human beings seem to have the ability to think any one of an infinite number of thoughts. This, in itself, is not a particularly staggering observation—or, should I say, while the ability itself is quite amazing, noting that we have said ability isn't, since it's a fact that is

---

[20] Of all the statements in Lycan's argument, I have changed the wording of the conclusion the most; I have done so because I believe that my statement (4) is a clearer statement of the representationalist view than Lycan's statement (IV) (p.409). This is not to say that I think Lycan's (IV) is incorrect in any way; I just think that it isn't the clearest statement of what even Lycan himself takes to be representationalism. A much clearer statement of representationalism by Lycan occurs in the first paragraph of his essay (pp.404-405). My conclusion (4) is actually something of a hybrid of Lycan's statement of representationalism in his first paragraph, and his statement (IV) on p.409, which more obviously follows from his premises. Again, I do not believe I have done any disservice to Lycan's argument by adjusting the wording as I have here.

[21] Lycan seems to more-or-less admit as much in a footnote (n.9, p.408).

[22] Why specifically 'representationalist'? I will address that question below.

available to just about everyone upon some very simple reflection. Why, then, is the unboundedness of thinking of any importance to an argument in favor of representationalism? To answer that question, let me begin by looking at another 'unbounded' human ability: the ability of language users to, as Lycan puts it, "understand [and produce, it should be said] indefinitely long, totally novel sentences on the spot" (1993, p.406). Lycan then goes on to note:

> Linguists have been driven by this infinitary or at least unbounded ability to suppose that the sentences that are understood must be composed of atomic, semantically primitive elements drawn from a finite and finitely learnable stock of morphemes, and that the meanings of sentences are compositionally determined by productive, recursive rules for arranging morphemes into certain combinations and orders. (That, surely, is what *syntax* or grammars of natural languages are for.) (P.406).

That is, linguists have supposed that the only way to account for natural language speakers' ability to understand and produce any one of an infinite number of sentences is to suppose that languages are made up of finite sets of simple parts, and that there are rules for combining those parts such that, theoretically at least, one could produce an infinite number of (well-formed and meaningful) sentences. One of the most interesting things about this linguistic theory is that while (as Lycan is quick to point out) it is an empirical theory, and so might be false, it seems to many, if not most, people to be the only theory that makes sense. As Lycan says,

> One cannot now imagine any other explanation for the striking facts of speakers' near-universal understanding.... For well-known reasons, Behaviorism will never do.... Divine inspiration is always a possibility ... but not a very likely one. The Chomskyan argument from unbounded competence out of finite resources to recursive structure is familiar and overwhelmingly persuasive. How could the thing be otherwise, barring either magic or divine intervention? (Pp.406-407).

We will, of course, be returning to this theme (the idea that there is only one sensible explanation for the phenomenon under consideration) below. For now, however, I wish to move back from this consideration about language to our discussion of thought, for we are now in a position to see the significance of the unboundedness of thinking as a premise in the argument for representationalism.

The significance lies in the drawing of an analogy: just as the only sensible explanation for the unbounded ability to use language seems to require a finite set of atomic linguistic primitives and recursive rules for their combination into sentences, so too, representationalists suggest, the unboundedness of thinking would seem to require a finite set of atomic mental primitives and recursive rules for their combination into thoughts. As was the case with language use, the unboundedness of thinking *might* be explained by something other than the posited system of atomic mental primitives and rules for their combination into thoughts—it might be explained by magic, or by divine intervention. Since these are not generally considered real alternatives, though, the analogy with language seems like a strong reason to endorse the primitives-and-rules version of thought.[23] By itself, of course, this doesn't quite get us to representationalism, except perhaps in some very broad sense, but if there is any consideration that figures prominently in just about every modern representationalist's argument for his or her own version of representationalism, the unboundedness of thinking has got to be it. For most representationalists, the analogy between language and thought doesn't simply stop here,

[23] Let me be entirely clear here. I am not providing an argument for the primitive-elements-and-recursive-rules system based on some deep principle. It may be that we could get the unboundedness of thinking in some other way—but no other way that seems even remotely sensible has ever been suggested. When it comes to languages, the Chomskyan argument has seemed unassailable for the simple reason that there doesn't appear to be any other way to account for the phenomenon in question. Representationalists have been drawn to this argument with respect to thoughts because the phenomena (unbounded linguistic ability, unbounded thinking) are so similar.

either, but continues on in some way or another (hence Lycan's parody, near the beginning of his essay, of certain anti-representationalists: "'Of course I'm a physicalist and a realist about beliefs and desires, but I don't believe in little sentences written in brain matter or anything crazy like that'" (pp.405-406)—for Lycan, and Fodor, and any number of other representationalists, it is perhaps oversimplifying, but not wrong, to say that they believe in 'little sentences written in brain matter').[24] At this point, though, the potential further analogies between thought and language are irrelevant; all we need recognize here is that the Chomskyan argument about unbounded language use has been heartily adapted by representationalists into a reason for taking the unboundedness of thinking as evidence for representationalism. That is the significance of noting the unboundedness of thinking, as far as representationalism is concerned.

Now, however, I wish to turn to a far more important element of Lycan's essay: his justification for premise (3), *viz.*, the claim that representationalism is in fact the only sensible way to explain minds. Whatever one may think of Lycan's argument overall, one cannot deny that this claim is important not only to Lycan but to representationalists generally. We are required, therefore, to give this particular claim some very close attention.[25] As I said above, representationalism certainly does not seem obvious to

---

[24] Stalnaker, in his reply to Lycan, actually focuses a fair bit of his criticism on the representationalist tendency to analogize thought to language. We will thus return briefly to this point below.

[25] Indeed, in calling representationalism the 'only game in town' we may very well seem to be begging the question against the non-representationalist. And we would be, if we were to take this claim as a presupposition. Hence the need for a careful look at the supposed 'obviousness' of RTM: to avoid simply begging the question, we must give an argument for the claim that representationalism is the only sensible way to explain what appear to be facts about thinking. On a further note, it might seem as though the argument for calling representationalism the 'only game in town' just *is* an argument for representationalism—in which case the argument that Lycan actually gives would seem superfluous. As a matter of fact, though, the argument for representationalism's supposed obviousness isn't a complete argument for the view itself, because RTM will seem like the 'only game in town' only given certain details about thinking, most notably its unboundedness. We must combine these premises to arrive at RTM as conclusion. Nevertheless, I do not think it is wrong to say that nearly all of the work in a generic

everyone. So the claim that it really is obvious, that in fact it is the only sensible theory available, appears to reveal (as Stalnaker mentions) the presence of a deep misunderstanding between representationalists and their opponents. As I also said above, though, even if I do not think representationalism is as obvious as Lycan suggests, I do believe that one of the strongest reasons for ultimately accepting it is the fact that it is far and away the best theory available—it may be a bit of an oversimplification, but I quite like the suggestion that the best way to characterize the argument for RTM is as an inference to the *only* explanation. I think that there is some truth in that way of putting it (not to mention that putting it that way also captures somewhat nicely the fact that RTM is an empirical theory which nevertheless allows for much development by philosophers in their armchairs instead of scientists in their labs).

What, then, is going on? How can a representationalist like Lycan (or myself) take RTM to be the only sensible theory available while others (some of whom, it is worth mentioning, are themselves representationalists) take it to be a bold speculation, and even others see it as entirely false? The problem, I believe, lies in how one understands both the scope of, and justification for, premise (3) in the argument for RTM above. Dropping the physicalist requirement, as I advocated doing earlier, what (3) says is that a system of a finite number of mental primitives and recursive rules for their combination into whole thoughts is a necessary condition for the unboundedness of thinking. Problems arise, then, as I said, from two directions. First of all, I think that there is quite a lot of confusion, and potential confusion, over the nature of the system of primitives and rules that is claimed in the premise to be necessary to account for the

---

argument for representationalism is done by the argument that there is really no other way to explain the facts about thinking.

unboundedness of thinking. More specifically, I imagine that many non-representationalists, as well as, most likely, quite a few representationalists, ascribe a more determinate character to this system than is warranted by the justification offered for the premise itself. That leads me to the second problem, which is that, because of the confusion just mentioned, the justification for (3) will often seem to fall short of making (3) seem at all obvious. One of Stalnaker's more important points in his criticism of Lycan's essay is the observation that the more determinate a representationalist's conception of the system mentioned in the consequent of (3), the less obvious it is that (3) is true.

Let us start from the beginning, then, and work through just what (3) says, and entails, and what it does not. If we can be clear about that, I think we will be able to clear up the confusion that leads to such differing attitudes about representationalism's epistemic status. If we take the unboundedness of thinking to be uncontroversial (or, at least grant it here for argument's sake), the first question is simply, How do we account for this phenomenon? Lycan, and representationalists generally, respond by claiming that it must involve some sort of system of mental primitives (of finite number) and recursive rules for their combination into thoughts. We are trying to explain an infinite capacity in finite beings; in attempting to do so, representationalists have turned to the only other example of this kind of problem that anyone is aware of, *viz.*, the unboundedness of linguistic ability. This analogy with language use is probably responsible for a good deal of the confusion at issue here.[26] Yet that doesn't change the

[26] In his critique of Lycan, Stalnaker focuses a lot of attention on the representationalist tendency to analogize thought to language. While I do not think that Stalnaker diagnosis that problem correctly, he is at least on the right track.

fact that in attempting to explain the unboundedness of thinking, it has appeared to many philosophers that we simply have no other options. Are there other possible ways of accounting for the unboundedness of thinking? Yes—but none that are at all plausible. Of course, we might just assert, in a Searlean sort of way, that this unbounded ability is simply a brute fact about minds, about which nothing more can be said. This, however, is not an explanation—it is a call for an end to inquiry, and as such is not an alternative to the approach favored by the representationalists.

What, though, does this line of argument really get us in the end? And aren't I simply repeating the assertion that representationalism is obvious, and thereby ignoring what I said was a very real confusion between representationalists and their opponents? To the latter question I respond: Not at all. The reason is that, as I understand it, the argument so far really only gets us a very basic representationalism, of a sort that is likely only controversial to those who are anti-realist about the propositional attitudes (and neither Lycan's argument, nor my analysis of it here, is meant to convince those philosophers—that is the main reason I spent the time I did above attacking eliminative materialism and instrumentalism). So far all I've said is that a combinatorial system with finite atomic parts is the only way we know of for getting the infinite capacity for new thoughts out of finite beings like us.

The confusion comes in because in saying this we haven't really said much about the determinate character of this combinatorial system. To some it probably seems like nothing other than representations will work—even if those representations have no symbolic character (maybe they're just physical states of the brain). But I suppose it's not *obvious* that we need representations—at least, not in the very specific ways in which

most representationalists envision them ('little sentences written in brain matter'). And now we strike suddenly upon the heart of the problem: the analogy with language.[27] This analogy tends to get introduced in arguments for representationalism because speaking, like thinking, involves an infinite capacity in finite beings. But for many representationalists the analogy with language is taken quite a bit further, so that the mental primitives are imagined as parts of a "language of thought"; and the rules for their combination are thought to be much like the rules that govern the formation of sentences from morphemes. None of that, however, is obviously true. At the very least, a lot more needs to be said besides simply noting the unboundedness of thinking if we are going to go all the way to a language of thought, i.e., if we are going to treat the analogy with language as much stronger than is initially required.[28]

My point is that none of this (the further, stronger analogy to language, or the determinate character of the required combinatorial system) is essential to the above argument for representationalism. The combinatorial system that seems necessary for unbounded thinking is, in (3), rather indeterminate in character. A short bit of further brainstorming will probably convince us that representations (in a broad sense) are going to be necessary, but I will have more to say specifically about representations in Chapter Two, so I won't go through all that here. Beyond the simple need for a finite set of representations and rules for their combination, the argument above doesn't get us

---

[27] Stalnaker relies heavily upon attacking this analogy in his response to Lycan. I'm not addressing Stalnaker's worries here precisely because I think that, while they may be relevant to a critique of Lycan's essay, they are irrelevant to the basic argument for representationalism.

[28] To repeat and clarify: initially language use enters the picture only as an instance of another case where we need to get an unbounded ability out of finite beings. The analogy is therefore quite weak, in many respects—it does not even suggest that thinking involve anything like linguistic structure. All it says it that thinking requires a combinatorial system of which natural languages seem to be one kind.

anything else. But, of course, we don't *need* anything else. For we already have all we need for basic representationalism.

Lycan's essay ends up being somewhat misleading. He characterizes representationalism as the view that "for *S* to think or 'occurrently believe' that *P* is for there to be a state of *S*'s central nervous system that (a) plays a characteristic information-storing role in *S*'s behavioral economy and (b) bears the semantic or propositional content that *P*" (p.404). The mention of the central nervous system is not typical in stating the representationalist picture, but aside from that this is pretty much how representationalists usually characterize their view. To put it another way: 'Having the belief that-p is having a representation that means 'p' in one's 'belief box'—i.e., to token that representation in the 'believing' way.' But this way of putting things is heavily influenced by the linguistic metaphor (as Stalnaker calls it); that is, the picture presented by these statements of representationalism draws heavily upon how we understand language to work. Meaningful utterances are tokens of words and sentences with certain semantic contents, and the same contents can be uttered in different ways to produce different, though related, sentences.

Again, the point is that all this is simply added baggage, which I believe is weighing down the argument for representationalism. We do need to give any representationalist theory that is going to be even remotely helpful a great deal of determinate character—but that comes later (in my case, in the entirety of the remaining chapters). All we need to establish the *basic* representationalist picture is the appeal to a combinatorial system of the type discussed here, and perhaps the ever so small additional comment that (again, so far as we can see) the only way such a system is going to do

what we need it to do in order for it to be an explanation of the phenomenon we need it to explain is if the system involves representations—of some as-of-yet undetermined character.

That's really all there is to it. The only thing wrong with Lycan's argument, in my view, is that he tries to use this simple and powerful argument for representationalism to sell a much more specific version of the view. He therefore opens himself up to the kinds of attacks that Stalnaker mounts against him. None of this is too terribly surprising, though, since the kind of generic representationalism that I am here suggesting really is pretty obvious, i.e., really is the 'only game in town', is so general as to leave almost all the really pressing questions unanswered. Nevertheless, we have managed to establish something useful in this chapter. Namely, that if one is going to be an intentional realist, one really has no choice but to be a representationalist. We have carved out our first chunk of the logical space that the LTT resides in. It is a rather large chunk, yes, excluding only a few views, like eliminativism and instrumentalism, but this is an important step because, as we shall see, we'll have problems aplenty within the representationalist camp without having to worry simultaneously about anti-realist and non-representationalist views.

2. On Mental Representation

We are now working within the Representational Theory of Mind. For the rest of this project I will more or less take RTM for granted. As I hopefully made clear in my Introduction, however, RTM is a broad playing field: there is a wide variety to the views that could properly be considered representational.[29] So the first question we should ask when distinguishing different representational theories is: What are the natures of the representations of each theory? The answers generally have two parts, discussions along two different axes, which we might call form and content.[30]

In Section 2.1 below I will discuss some of the different types or forms that philosophers have thought mental representations might take. Representations can take many forms, from pictures to symbols. Of course, these representations can also be realized in various media—but the discussion below is concerned with the type of representations that seem most appropriate for mental representations, and is not concerned directly with the medium in which mental representations are realized. The object of Section 2.1 is to argue that some sort of symbolic form of representation is necessary for thought (though that claim alone is hardly contentious these days). I will then briefly discuss what seems to be the most promising form of symbolic representation for thought: language-like symbols. We've already seen part of the argument for this suggestion in §1.13 above. But, of course, the main purpose of this entire project is to argue for linguistic symbols as the form of thoughts, so what I say in Section 2.1 will only be a small part of the overall discussion.

---

[29] Indeed, as I noted in the Introduction, if one construes 'representational' broadly enough—though RTM is usually taken more narrowly—representationalism will embrace even Cartesian dualism.

[30] Robert Cummins refers to these two axes as the Problem of Representations (plural) and the Problem of Representation (singular), respectively. See Cummins, 1989, pp.1-10.

In Section 2.2 I will discuss the major theories of content. Once we have adopted the idea that mental representations work through some form of symbolic representation we are then faced with the further question of how these symbolic representations come to mean what they do, to be about whatever they're about. This is the search for a psychosemantics. There are many psychosemantic theories, of course, but the major ones divide fairly nicely into three main camps: causal-covariance theories, teleological theories, and functional role theories. I discuss a choice example or two of each type of psychosemantic theory in Section 2.2 below. Ultimately, I endorse a type of functional role semantics, but the main arguments for the specific version I have in mind form the material of Chapter Three. So what I am primarily doing in this chapter is mapping out the field within which the rest of our discussion is to take place.

2.1 Types of Representations

If our thoughts are constituted by representations, the most obvious question is: What type of representations are they? That is, what form do they take? We encounter all sorts of representations in the world: pictures (of varying degrees of realism), signs (e.g., stop signs and yield signs), words (like those on this page); then there are things like works of abstract art, metaphors in books and film, and symbols on maps. All of these things are representations, but not all of them are suited to serve as *mental* representations. The most serious contenders have generally been images and symbols of some kind. I will argue below that an imagistic conception of mental representation is inadequate, and that some form of symbolic representation is required by RTM. Furthermore, I think that our best examples of the kind of symbolic representational

system needed are the languages we speak. While I do not yet wish to argue about whether a natural language or some special 'mental language' (Mentalese) is the language of thought, I will begin to try to make the case that thinking takes place in *some* sort of language-like symbolic system.

2.11 Images vs. Symbols

Here is a story about how the mind might come to represent the world around it.[31] Everything that is, is a combination of matter and form. A red ball is just physical stuff informed by redness and sphericity. The idea of a red ball is just mental stuff informed by redness and sphericity. Thus, when we have an idea of some perceptual object, what we have is a bit of mental substance that has taken on the same properties (i.e., the same forms) as the object. My idea of a red ball represents the red ball itself because it has the exact same form: it is red and spherical, just like the ball. Representation, in this instance, is, Cummins says, "perfectly transparent: The idea represents the red ball, and it represents it as red and as spherical because the idea is red and spherical and the redness and sphericity come from the physical ball" (1989, p.4). Now, if we drop the (probably oversimplified, and even fictitious) ancient metaphysics, we are left with the basic idea behind the imagistic conception of mental representation: thoughts represent objects by resembling them. A mental representation is an image, a picture, that has (in some sense) the same properties as the object that it represents. The thought of a red ball is an image which is itself red and spherical. Of course, the idea of a red ball, that is, the mental image of a red ball, might not be red and spherical in *exactly* the same way as the

[31] I am borrowing the discussion of this paragraph from Cummins, 1989, pp.2-4.

physical object that it represents, but the resemblance is close enough to make the representation, as before, entirely transparent.

There might seem to be some intuitive plausibility to this idea, at least when we are considering the representation of simple perceptual objects or events. We should note that by 'image' here we do not mean only visual images; one might represent the sound of a hammer striking steel by imagining—having an auditory image—of how that would sound. That is, thinking about a hammer striking steel might consist in having a visual image, an auditory image, or both together. Again, I think that there might seem to be some intuitive plausibility to this conception of mental representation. When I think of a red ball, or of a blacksmith at work, or of my office at home, these thoughts often do seem to involve mental images. And given the close resemblance—the shared properties—between the object and the image of it, we have no trouble seeing why the image is a representation of that object. It is *obviously* a representation of that object. Since I won't have occasion to mention this later, we should also note that determining the content of an imagistic representation seems (at least in simple cases like that of a red ball) to be unproblematic. The image gets its intentional content from the object it resembles; that is, an image of a red ball is about a red ball because it *looks* like one.

We should not find it surprising, then, that the imagistic conception of mental representation has enjoyed some popularity in the past. The early empiricists, for example, seem to have been attracted to this view. Nevertheless, one can show, without much difficulty, that images just aren't robust enough to do all the representing required for thought. There isn't much of a problem when the representations in question are of simple physical objects: an image of a red ball seems to represent that ball perfectly well.

But what about abstract ideas? What image could possibly represent the number 35, or the concept of mental representation itself? The problem isn't just that we can't easily pick an image that might represent each of these ideas, it's that they are not the sorts of things that *could* be resembled by an image. A red ball is red and spherical, and an image of red ball can perhaps resemble the ball itself in these respects. But the concept of mental representation has no properties, visual, auditory, or otherwise, which could be formed into an image of that concept. Of course, the words 'mental representation' look and sound a particular way, but an image of the words is not an image of the concept. Whatever the intuitive plausibility of the imagistic conception when it comes to representing simple perceptual objects, I do not think there is any plausibility to the idea that we represent concepts and abstract ideas by forming mental images of them.

If not images, then what? What kind of representational system could be powerful enough to represent the gamut from simple perceptual objects and events to abstract ideas and concepts? The answer seems to be some system of abstract symbols. First of all, abstract symbols do not represent in virtue of resembling that which they represent, so concepts and abstract ideas are no more problematic than simple objects. Of course, since the ability of symbols to represent is not dependant upon resemblance it is also not obvious how a symbol comes to mean what it does. This is the problem of developing a psychosemantics, and, as mentioned earlier, will be dealt with in Section 2.2 below, as well as in Chapter Three. For now, I want to talk generally about the virtues of a symbolic system of representation.

In the first chapter of *Language, Thought, and Consciousness* Peter Carruthers discusses briefly three features of the propositional attitudes that any system of mental

representation would have to be able to accommodate. First of all, he says, "propositional attitudes are *systematic*, having contents that are systematically related to one another" such that anyone who can have a thought with a particular content "must be capable of believing or thinking a number of closely related contents" (1996, p.33). So, if I can think that the cup is on the book, I must also be capable of thinking that the book is on the cup. In order to be a system of *mental* representation, any candidate representational system will have to have this kind of systematic organization, so that it can both accommodate and perhaps explain the systematic nature of the propositional attitudes.

Secondly, "propositional attitudes are *productive*, in the sense that anyone capable of thinking at all must be capable of entertaining unlimitedly many thoughts" (p.34). We've already seen this idea in Chapter One when discussing Lycan's deductive argument for RTM: this is just the unboundedness of thinking again (Lycan, 1993, p.408). Given that this is one of the reasons to prefer RTM in the first place, it should be obvious that any candidate system of mental representation is going to have to be able to accommodate this feature of thought.

Lastly, "propositional attitudes interact causally with one another in ways which respect their semantic contents" (Carruthers, 1996, p.34). The desire for a soda and the belief that there is soda in the refrigerator combine with each other, and other beliefs, to cause me to go into the kitchen. And the subsequent belief that I am out of soda will, given the right desires and intentions (desires and intentions that have related content), cause a further belief that I ought to go to the grocery store to buy more soda. Whatever

system of representation constitutes thought must be able to explain these semantically informed causal interactions.

The right system of symbolic representation can do all this. But what kind of system is the right system? Imagine for a moment some simple examples of symbolic representation—say, the universal symbols warning of a nuclear danger or a biohazard. Now, symbolic representations like these do not rely in any way upon resemblance—it is symbolic representation, not imagistic representation. So there is no problem in supposing that we can have a representation of abstract concepts and other things which do not lend themselves to imagistic representation. Now imagine such a system used not just to represent certain specific dangers, but to represent everything about which human beings can form a thought. I might represent the fact that I have an appointment tomorrow at noon with one symbol, that I taught class last Friday with another, and so on.[32] As long as we had some way of specifying the content of any given symbol (that is, of saying how it came to be a representation of just *that* thought, and no other), we would have a system of representation that avoided the main problem with the imagistic conception.

So at this point we're imagining a system of symbolic cognitive representation in which every distinct thought consists of a symbol of some kind tokened in the brain. Coupled with a psychosemantic theory that assigns to each symbol a distinct content, we seem to have a candidate representational account of thought. But now we need to consider the three features of propositional attitudes that we just looked at. Can our

[32] Since we are going to need an infinite supply of symbols on a view such as this, it's clear that shapes (as with our warning symbols) aren't the sort of thing we're looking for. Perhaps every thought is a particular number encoded in binary form, as in a computer: with an infinite supply of numbers, we can encode an infinite number of thoughts.

candidate representational system of thought accommodate these three features? Apparently not.

First of all, a system in which every distinct thought is represented by a single distinct symbol will have a problem accounting for the systematic nature of the propositional attitudes. Suppose that we represent the thought that the cup is on the book with one symbol (say, α); then we represent the thought that the book is on the cup with another symbol (say, β). We know that the contents of each symbol are related: they're about the same objects in the world (the cup and the book), they're about the same physical relation (one object being on top of the other), and the only difference between them is that they transpose which object is on top of the other. But how much of this comes out in the relation between α and β? They are just distinct simple symbols; nothing in the relation between α and β reflects the relations between their contents. The problem is that the thoughts (the cup is on the book, the book is on the cup) are related in ways that break into the parts of each thought, but the symbols (α and β) have no parts.[33] Each thought is a different symbol, and the symbols have no systematic, content-reflective relation to one another.[34]

A similar problem arises from consideration of the third feature of propositional attitudes mentioned above: the fact that propositional attitudes interact causally in ways that respect their semantic content. Getting the symbols to interact causally isn't the issue (since getting *any* symbols to interact causally is just a matter of coding those

---

[33] This is like the problem of trying to represent the logical structure of 'All men are mortal; Socrates is a man; therefore Socrates is mortal' in propositional logic. The representational resources of propositional logic are simply not fine-grained enough to capture the logical structure of the argument.

[34] This is also similar to an objection that Fodor has to what he calls the "fusion story" of propositional attitudes. See Chapter 7 of his *Representations* (1981, pp.179-180).

symbols into some physical system or other—a computer or a human brain, say—such that the physical interactions constitute causal interactions between the symbols). But getting the symbols to interact causally *in ways that respect their semantic content* is much harder, for now we need not just interactions between the physical instantiations of the symbols, but interactions that mirror physically what goes on abstractly at the symbolic level. We've already seen, however, that in representing every distinct thought with a distinct and simple symbol we seem to *lose* the semantic relations between thoughts. If the symbols themselves don't mirror the semantic relations in question, there is no way that a physical instantiation of those symbols will do so. Hence, there is no way for the symbols to interact causally in ways that respect their content.

The unboundedness of thinking is also problematic for the type of representational system that we're considering. The system we've imagined might be theoretically unbounded (just so long as we have an infinite number of symbols to draw upon, we can theoretically represent an infinite number of thoughts), but it's hard to see how it won't suffer from severe limitations when put into practice. Given that each of us is capable of thinking any one of an infinite number of new and novel thoughts, we're going to need an infinite number of symbols, each with it's own meaning, ready to go at a moment's notice. The main issue here is probably going to be generating a psychosemantics that would work in such a case. Clearly it cannot be the case that we each have an infinite store of symbols waiting to be activated should we suddenly wish to think (or understand) something we'd never thought before. Our brains are powerful, but finite. So our brains would need some way of generating the required symbols on demand, and assigning to them the content that they were supposed to have. But how

would such a system determine what content went with each new symbol? That is, how would any content get assigned at all? The challenge for a psychosemantics here seems overwhelming. Combined with the other two problems above, I think we can safely say that no system of simple, unique symbolic representations for each represented is going to be able to accommodate the productive, systemically organized, and causally efficacious features of the propositional attitudes. Fortunately, we already have a perfect example of symbolic representational system that *can* accommodate these features: a language.

2.12 Linguistic Representation[35]

We saw in Chapter One that the best (perhaps only) way to account for the unboundedness of thinking is by having a system of representation that allows for the construction of an infinite number of unique wholes out of a finite stock of symbolic primitives along with rules for their combination (a syntax). What might such a system look like? We have to search no farther for an example than the language that we speak. Language, too, is unbounded, in that there are an infinite number of new and novel sentences that any speaker of a language can form once he has the basic and finite vocabulary, and knows (or at any rate, grasps for practical purposes, even if he cannot explicitly state them) the grammatical rules for their combination into meaningful sentences. But the words and sentences of a language are also systematically related in a way that respects content, and, as such, are the sorts of symbols that can be encoded into

[35] What follows is a discussion of how language as a representational system has the features required for a theory of mental representation, i.e., one that accounts for the features of propositional attitudes mentioned above. For a detailed, yet readable, description of how languages work, see Steven Pinker's *The Language Instinct* (1994), especially Chapter 4.

physical systems in such a way that the causal interactions of the system respect the contents of the symbols. That is, language has just the properties that Carruthers has suggested propositional attitudes have, and that any system of representation which we took to be a candidate for *mental* representation would need to have.

Sentences are built up from words, and the meanings of words and sentences are functionally related (a fact that we will explore in great detail in Chapter Three, and continue to work with in subsequent chapters). There is, therefore, a systematic relationship between sentences that respects the content of those sentences. The content of 'the cup is on the book' is related in certain specific ways to the content of 'the book is on the cup' in virtue of the fact that the words in each sentence, along with the rules of syntax, determine their meanings, and the two sentences use exactly the same words in a very similar order.[36] Thus, unlike the simple symbolic system imagined earlier, linguistic representation is fine-grained enough to capture the semantic relationship between the thought that the cup is on the book and the thought that the book is on the cup. If, then, thoughts were relations to internal sentences (of some language or other—perhaps Mentalese), we would have a way of accounting for the systematic organization of the propositional attitudes. That is, just as the sentences of a natural language are systematically related to one another, so too will the sentences of our mental language be systematically related to one another. Hence, our thoughts, being simply relations to these sentences, would also be related in just the way that we already know them to be.

Language is also productive, in the sense that an infinite number of sentences can be constructed (and understood) using only a finite set of components. We saw this in

---

[36] We will be discussing how those words become meaningful in the first place in Chapter Three.

Chapter One, as this is one of the features that recommends the Representational Theory of Mind in the first place: a recursive symbolic system is the only way that we can see of getting infinite ability out of finite beings—whether it's the ability to understand any one of an infinite number of sentences, or the ability to think any one of an infinite number of thoughts. So here, too, language recommends itself as the type of representational system needed for thought.

Finally, language can also be causally efficacious. That is, if sentences are coded into a physical system of some kind (like a brain) in such a way that the physical interactions mirror the syntactic and semantic relations that the sentences bear to one another, then the sentence tokens will be causally robust. As we noted earlier, the issue is not the coding of the symbols into a physical system, but rather whether or not those symbols have any syntactic and semantic relations to one another that can be coded for in the first place. We've already seen that the symbols of a language do bear the rights sorts of relations to one another. Thus, by instantiating those symbols in a physical system we enable tokens of them to be causes. So, again, a language of some kind seems like the perfect sort of system for representing thoughts.

As far as the form of mental representations goes, then, we seem to have some reason to think that they will be language-like symbols. If we're going to be realists about the propositional attitudes, and representationalists to boot, then we're going to need a sophisticated system of representation. And, again, the best example we have of the type of representational system that is needed is a language. At the very least, though, I think that we can say that thought must be some kind of symbolic process—that, I think, is something that almost all representationalists would agree with these days.

In the absence of alternate suggestions, then, we might as well use our best example of symbolic representation—language—as the model for mental representation.

While the issue of the form that mental representations take, at least broadly speaking, isn't likely to be overly contentious among contemporary representationalists, the issue of content is really the issue that truly divides modern theorists. Even among philosophers who agree that mental representations will take the form of internal symbols, there is some variety in the answers given to the question of how those symbols come to represent (i.e., mean) what they do. Thus, we now turn to a consideration of the second axis in our general discussion of RTM, a look at the three major forms of modern theories of content for mental representations.

2.2 Psychosemantics

I will have a lot to say about theories of meaning in general in Chapter Three. What I wish to do in this section is say a little bit about the major psychosemantic theories out there these days. I will try to make the case that while each of these theories has its strengths as a semantic theory for some forms of representation, none of them is appropriate as a *psycho*semantic theory. The problem, as I see it, is that thinking is essentially a norm-governed activity, and none of the theories below gives us a semantic theory that takes this into account. That is, all of these theories of representational content treat representation as a monolithic activity: the philosophers who propose these theories treat the representing that goes on in things like simple organisms and computers as of a piece with mental representation, and I think that is a mistake. A semantic theory which explains the content of the representations in a computer, or in a magnetotactic

bacterium, is not necessarily suited to explaining the content of the representations in the mind of a thinking being. When I say that thinking is a norm-governed activity, what I mean is that it is rule-governed in a way that involves agency, and a responsibility on the part of the representing system for obeying the rules. A thinker who thinks 'P' and 'If P, then Q', but who also thinks 'It's not the case that Q' has made a mistake for which she can be held responsible; a thinker who thinks 'Cat' when looking at a dog has made a mistake for which he should be held responsible. Conversely, when I think 'Dog' while looking at a dog, I've done something right, and it's an activity for which I am to be commended. On the other hand, a computer that processes '2 + 2' and comes up with '4' deserves no credit: the programmer deserves the credit (if any is to be given). And a computer that represents 'P' and 'If P, then Q' and also represents 'It is not the case that Q' deserves no blame; the whole notion of blame is out of order here. The machine has made a mistake, but it's not its *fault*.

Thus, I want a psychosemantic theory that allows for true ascriptions of praise and blame. And I will try to show below that the major theories out there today do not allow for this. Getting it right, and getting it wrong, on these theories, is a matter of pure mechanism. Of course, I think this was probably done on purpose, for all of the theories below are attempts at reductive, naturalistic theories of meaning. I, too, want ultimately to be something of a naturalist about meaning. That is, I take the semantic content of mental representations to be something which arises from, and supervenes upon, the complex behavior of certain complex physical systems, and something which arose, ultimately, via evolution. But I do not think that we can be quite as reductive about

meaning as the theories below are. That is, I do not think that meaning in the context of mental representation can be reduced to simple physical interactions and causes.[37]

Of course, I'm not supposing that it's in any way *obvious* that thinking is a norm-governed affair. I will take up that issue in Chapter Four, where I will also be critiquing the theories discussed below more generally. What I want to do here, though, is introduce these psychosemantic theories in individual detail (and give preliminary critiques of them), for the issue of content is perhaps *the* defining issue when it comes to different theories of mental representation, and we will need the background of this chapter to fully appreciate the arguments to come.

We've just taken a look at the forms that mental representations might take, and it seemed as though our choices were very limited. Indeed, one might be tempted to suppose that we really only have one choice when it comes to the form of mental representations: language-like symbols. But saying how representations come to have the content that they do, i.e., to mean what they do, is more difficult, and the possibilities here are more varied, and more evenly weighted. Of course, the need for a psychosemantics should be clear. According to RTM, to think is to token a representation of some kind; and the content of the thought is determined by the content of the representation. Now, if we're supposing (and for now I do want to suppose this) that the representations we're interested in are symbols of some form or other, we've got to be able to tell some story about how those symbols get their meaning. And since it's *mental* representation that we're interested in, we cannot suppose that those symbols get

[37] I say more about this issue of striking a balance between the naturalistic and the normative in §6.12.

their content from the content of our thoughts—it's the content of our thoughts that we're trying to explain in the first place. Hence the need for a psychosemantics.

As I said, there are a number of different theories that have been proposed over the years for how the contents of mental representations get determined. The most important theories generally fall into one of three categories: covariance, teleological, or functional.[38] For the rest of this section I will examine each of these in some detail. They all have certain strengths (some of which I might want to hold on to), but ultimately none of them seems to me to provide a satisfactory account of the meanings of mental representations—because none of them recognizes the inherently norm-governed nature of thinking.

2.21 Causal Covariance

In their simplest form, covariance theories are quite straight forward. According to what Fodor has called the "Crude Causal Theory"—a very basic covariance theory—the content of a mental representation is determined by whatever condition(s) in the world that representation causally covaries with. That is, "the symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings" (Fodor, 1987, p.99). So, if cats cause mental representation *R*, the Crude Causal Theory says that *R* means 'cat.' This causal relationship has to be reliable, of course. The claim is not that sometimes one mental representation will mean 'cat' and another time a different representation will mean 'cat' while the first does not. That

[38] Sometimes a theory will actually fall into more than one of these categories—though there is usually one category to which it belongs more clearly than the others, one category that dominates, that is. For example, Dretske's view is largely a covariance view, though it relies heavily on some teleological notions. And Block's theory is primarily a functional one, though his 'wide content' requirement encompasses some elements of covariance.

wouldn't fix content in the way that covariance theorists want. What we want to be able to say is that |CAT|[39] tokenings mean 'cat' because it's cats and only cats that cause such tokenings.

Of course, there are problems with the Crude Causal Theory right off the bat. For instance, the problem of misrepresentation (often also called the 'disjunction problem'). The worry goes like this: We know that sometimes representation misfires. That is, we think we're seeing a cat when actually it's a dog, say. In a case like this, |CAT| is tokened, but it is caused not by a cat, but by a dog. Hence, we misrepresent the world. But how, on the Crude Causal Theory, is such a thing possible? For if mental symbol tokens denote their causes then a |CAT| caused by a dog denotes a dog; and since |CAT|s express the properties of the instantiations that cause them, if dogs cause |CAT|s then part of what |CAT| expresses is the property of being a dog. Of course, |CAT|s also represent cats. Hence, it would seem that, on the Crude Causal Theory, we ought to say not that |CAT|s mean 'cat' but that |CAT|s mean 'cat or dog' (this is why the problem is often called the 'disjunction problem'). But now if |CAT|s mean 'cat or dog' then it turns out that a |CAT| caused by a dog does not, in fact, misrepresent the world after all. In general, simple covariance theories will seem to fail to be able to account for misrepresentation for just this reason. Every instance of apparent misrepresentation ends up being better understood as an instance in which the content of a given mental symbol token has been misdescribed (if dogs can sometimes cause |CAT|s, then in claiming that

[39] I will use capital letters and absolute values to refer to the mental symbols themselves. This should not, however, be taken as implying that the symbols have any particular linguistic features, or even any linguistic features whatsoever (though, of course, for Fodor they do). I thus use |CAT| to refer to the mental symbol token that means 'cat,' whatever its intrinsic characteristics may be, and however it manages to have the semantic content that it does.

|CAT| means 'cat,' we have merely misdescribed the content of |CAT|s—given what they causally covary with, |CAT|s must mean 'cat or dog').

Thus a simple covariance theory won't get us very far. More sophisticated covariance theories, though, might have better luck. I want to look at two of these more sophisticated covariance theories here: Fodor's and Dretske's. The best way to begin is by looking at how each of these theories addresses the problem of misrepresentation.

Fodor tries to solve the disjunction problem by appealing to what he calls *asymmetric dependence*. Basically, the story goes like this. The dependence of |CAT|s on cats and the dependence of |CAT|s on dogs are not the same. Dogs wouldn't cause |CAT| tokens if cats didn't; but cats would still cause |CAT| tokens even if dogs didn't. That is, the fact that dogs sometimes cause |CAT| tokens depends upon the fact that cats cause them, but the fact that cats cause |CAT| tokens does not depend upon dogs causing them. You might say, then, that the primary relation is between cats and |CAT| tokens; hence |CAT| means 'cat,' and a dog-caused |CAT| is a misrepresentation. We must note that this asymmetric dependence is supposed to be a purely causal story. That is, we can't appeal to any semantic or intentional elements to explain why the cat-to-|CAT| relation is the primary one. For example, we cannot say that cats would cause |CAT| tokens even if dogs didn't because |CAT| tokens are intended to be about cats and not dogs. Such a claim would make the covariance theory obviously circular. Rather, the primacy of the cat-to-|CAT| relation is supposed to be due to something like the fact that |CAT|s are *usually* caused by cats, and only sometimes by dogs.

Asymmetric dependence may seem promising as a solution to the disjunction problem, but there is a worry about it. Imagine (to adapt an example from Robert

Cummins) that I can't tell the difference between a mouse and a shrew. Suppose, however, that I have learned to call certain creatures 'mice' and that I have the mental representation |MOUSE| tokened in me whenever I think about such creatures. But in fact I've never seen a mouse, just lots of shrews. That is, all of my |MOUSE| tokens so far have been caused by shrews.[40] Now, the reason that shrews cause |MOUSE| tokens in me is that shrews look like mice and the people who taught me to talk about mice weren't any better than I am at distinguishing mice from shrews. That is, shrews only cause |MOUSE| tokens in me because mice would (if I were ever to see one). On the other hand, suppose I were now to suddenly see a mouse. It would, we can suppose, cause a |MOUSE| token in me. It would only do this, however, because it looks like a shrew and shrews causes |MOUSE| tokens in me. So, shrews wouldn't cause |MOUSE|s in me if mice didn't; and mice wouldn't cause |MOUSE|s in me if shrews didn't. Hence, there's

---

[40] I have a worry about whether such a thing is possible on a causal covariance view. We need to remember that the convention I've chosen (and that is similar to a convention that Fodor follows) to signify a mental representation using the English word that matches what we take the representation to mean is just that: a convention. That is, for example, writing '|MOUSE|' is a convenient way to mention a representation *R* that *means* 'mouse.' But the *fact* that *R* means 'mouse' isn't determined by this notational convenience, but rather by whatever theory of content we're currently concerned with.

So, here's my worry: We might reformulate the set-up for the problem in the text by saying: "All of my *R* tokens so far have been caused by shrews." Now, in order to set up this problem for Fodor's account of asymmetric dependence, we are supposing that *R* means 'mouse' even though *R* has always (in me) been caused by shrews. But why can we suppose that? Presumably it's because whenever I saw a shrew in the past, and had *R* occur in me, someone (my parents, say) pointed and said the word, 'mouse.' But this account can't be right, because it would suggest that the content of the representation *R* was determined by the word that was tokened along with the representation, rather than the thing (or, more precisely, the property of the thing) which caused *R* to be tokened in me. Even on Fodor's 'Language of Thought' view, it's not the *meanings of English words* that determine the content of mental representations. The content of a mental representation is determined by whatever that representation (asymmetrically) causally covaries with. So how would it be possible to 'learn' a representation entirely from non-instances? If *R* in me has always covaried with shrews, then it seems as though the covariance theorist ought to say that *R* means 'shrew' even though I *call* shrews 'mice.'

I am worried about this situation not because I find the 'learning entirely from non-instances' problem with Fodor's asymmetric dependence account all that important, but because Fodor himself suggests that problem. That is, Fodor himself worries about the case where someone 'learns' a mental representation "entirely from noninstances" (p.109). Thus, Fodor obviously thinks that it makes sense to suppose someone having *R* tokens that *mean* 'mouse' occur in them entirely from encounters with shrews, and I just don't see how such a thing would even be possible on his view.

no asymmetric dependence in a case like this. But that means that, in me, a shrew-caused |MOUSE| token *isn't* a misrepresentation—though, it seems that it should be.

Fodor's way out of this bit of trouble is to insist that the asymmetric dependence condition apply with "synchronic force" (p.109). Robert Cummins summarizes this requirement thus: "No matter how |mouse| and |shrew| are learned, current dispositions make the mouse-to-|mouse| connection primary" (1989, p.60). That is, regardless of past connections between shrews and |MOUSE|s, the only thing that matters for asymmetric dependence is the *current* connection between shrews and |MOUSE|s. And Fodor's contention is that, even if |MOUSE| was learned entirely from shrew-instances, the current shrew-to-|MOUSE| connection is asymmetrically dependent upon the mouse-to-|MOUSE| connection. Thus |MOUSE| means 'mouse' and a shrew-caused |MOUSE| is a *mis*representation. So we now have a more sophisticated covariance theory: the content of a representation is not determined just by whatever that representation covaries with, period, but rather by whatever that representation *asymmetrically* covaries with. Let me now turn briefly to Dretske's way of dealing with the disjunction problem, for the difference between his solution and Fodor's is, I think, instructive.

Dretske's solution to the problem of misrepresentation involves appealing to what he calls "functionally derived meaning" (1986, p.332). Functional meaning (M_f) is defined thus: "*d*'s being *G* means$_f$ that *w* is *F* = *d*'s function is to indicate the condition of *w*, and the way it performs this function is, in part, by indicating that *w* is *F* by its (*d*'s) being *G*" (p.332). A simple example: a gas gauge's function is to indicate the amount of fuel in the tank; it does this, in part, by the movement of its needle. That is, the gauge indicates that the tank is empty by the needle resting on 'E.' Now, if the tank were full of

water, the needle would rest on 'F' rather than 'E,' but since it is the function of the gauge to indicate the amount of fuel in the tank, we could say that the gauge *mis*represents the tank as full of fuel. Of course, if we disregard function, the position of the gauge's needle merely covaries causally with the amount of liquid in the tank—so in that sense, the gauge would not be misrepresenting anything. Taking function into account, though, allows for misrepresentation.

I mentioned earlier that some of the theories that we would look at crossed over the various divides that I had put in place. Here we have an instance in which a covariance theory turns in part on the notion of teleology. It is not a teleological theory because, on Dretske's view, the content of a mental representation, even when it involves functional meaning, is still determined by causal covariance. But the inclusion of purpose into the picture brings Dretske's view closer to a teleological theorist's view than, say, the Crude Causal Theory is. This becomes more apparent, I think, when we develop Dretske's view. For, of course, we can't leave it as it is: the notion of function that we appeal to in the gas gauge example already involves human intention and semantic evaluation. The gauge indicates the level of fuel in the tank because *we intend* it to indicate that—that's what it was designed for. But mental representations, if they are to help explain human thought, cannot presuppose intentions. So Dretske develops his theory further by including the notion of a *natural* function.

A natural function is a function of an organism (or, more precisely, a particular part of an organism) that derives from the basic biological/evolutionary/survival needs of the organism. The classic example is magnetotactic bacteria. These bacteria cannot survive in an oxygen-rich environment, and they have internal magnets called

magnetosomes that align each bacterium with the Earth's magnetic field. This internal system draws northern hemisphere bacteria toward geomagnetic north, which ends up being down toward the ocean's floor where the water is less oxygenated. It thus seems reasonable to suppose that the *natural function* of these magnetosomes is to guide the bacterium toward oxygen-free water. There are problems with this example, but to make a long story short, I'll cut to the chase: while simple organisms (like the bacteria) most dramatically illustrate the notion of a natural function, that notion doesn't really help to explain how *mis*representation is possible until you consider an organism with a sufficiently complex representational apparatus. With simple organisms (like the bacteria) it is too easy, when faced with a potential example of misrepresentation, to simply redefine the natural function in question. In order for natural function to 'stick' and make misrepresentation possible, you need a representational system of great complexity. In fact, on Dretske's view, the only kind of representational system that will work is one that is capable of associative learning.

Consider a system that has multiple ways of representing the same information. Dretske's example is the "way we might identify oak trees visually by either one of two ways: by the distinctive leaf pattern (in the summer) or by the characteristic texture and pattern of the bark (in winter)" (p.337). Let's call the mental representation of a visually detected oak tree *R*. Now, there are two distinct causal paths to the tokening, in me, say, of *R*. Either the distinctive leaf pattern (a feature of the oak tree, call it f_1) causes an internal state in my sensory system (call it i_1) that causes *R* to be tokened; or the characteristic appearance of the bark (call it f_2) causes an internal state in my sensory system (call it i_2) that causes *R* to be tokened. An oak tree, *O*, will thus cause a tokening

of *R* in either one of two ways ($O \rightarrow f_1 \rightarrow i_1 \rightarrow R$; or $O \rightarrow f_2 \rightarrow i_2 \rightarrow R$). We might then say that it is the natural function of *R* to indicate the presence of oak trees.[41] That is, we might say that *R means* 'oak tree.' Now, if there were only one causal path from *O* to *R* (through, say, f_1 and i_1) we wouldn't really be able to say that *R* represents the presence of oak trees. Maybe *R* merely represents the presence of feature f_1; or internal state i_1. (This is the predicament we get into with simple organisms like the bacteria and their magnetosomes.) But since there are *multiple* causal paths from *O* to *R*, the tokening of *R* doesn't indicate one way or the other between f_1 and f_2. Hence, *R* represents the presence of an oak tree.

That is, unless we are prepared to say that *R* has a disjunctive meaning, in which case we could say that *R* means, not 'oak tree' but 'distinctive-leaf-pattern *or* characteristic-appearance-of-bark.' Thus, the disjunction problem rears its ugly head again. Dretske solves this problem by appealing to associative learning. If a system is capable of being changed over time, so that new causal paths from, say, *O* to *R* can develop, then, in terms of everything but *O* itself (i.e., in terms of the various features, f_x, and sensory states, i_x, that lead to *R*) there will be "no *time-invariant* meaning$_n$ for *R*; hence, nothing that, through time, could be its function to indicate" (p.338). At time t_1, *R*

[41] You might wonder how it could possibly be the *natural* function of *R* to indicate the presence of *oak* trees in particular, since there doesn't seem to be any biological/evolutionary/survival value in my ability to detect these trees specifically. Dretske has a response to a worry such as this:

> [I]t seems clear that a cognitive system might develop so as to service, and hence have the natural function of servicing, some biological need without its representational (*and* misrepresentational) efforts being confined to these needs. In order to identify its natural predator, an organism might develop detectors of color, shape, and movement of considerable discriminative power. Equipped, then, with this capacity for differentiating various colors, shapes, and movements, the organism acquires, as a fringe benefit so to speak, the ability to identify (and, hence, misidentify) things for which it has no biological need. (P.335).

I find this response plausible enough for present purposes, so I'll leave it be.

might $mean_n$ (naturally mean, i.e., causally—but not functionally—indicate) f_1-or-f_2; at t_2, *R* might $mean_n$ f_1-or-f_2-or-f_3. Since the function, over time, of *R* cannot thus be to indicate any one disjunction of features or sensory states, the only natural function for *R* is to indicate the presence of oak trees.[42] *R* therefore $means_f$ *O*.

How does this allow for misrepresentation? Well, suppose that some feature of a maple tree causes an internal state of my sensory system that causes a tokening of *R*. Since it is the natural function of *R* to represent the presence of a visually detected oak tree, when a tokening of *R* is caused by something other than an oak tree it is not fulfilling its function. Hence, it is a *mis*representation. Again, then, finding a solution to the disjunction problem helps produce a more sophisticated covariance theory.

I think that there is an intuitive appeal to covariance views like Fodor's and Dretske's. Given that |CAT|s are (usually) caused by cats (on that, everyone should agree), we save ourselves a step or two by simply taking this causal relationship to *be* the meaning relationship. Covariance views also have the advantage of giving us a thoroughly 'natural' account of representational content. On the other hand, as I mentioned in the introduction to Section 2.2, I do not think that covariance views offer a promising account of *mental* representation, whatever their other virtues, because they cannot account for the sensitivity to norms that I take to be essential to thinking. In the quest for a naturalistic and reductive account of intentionality, I think that covariance views go too far.

Fodor begins Chapter Four of Psychosemantics with the following observation about the apparent need for a reductive account of intentionality:

---

[42] See "Misrepresentation," (Dretske, 1986, p.338) for a slightly more detailed explanation of this conclusion.

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else. (1987, p.97).

He continues on to say that one of the main motivations for not being a realist about intentionality is the worry that "the intentional can't be *naturalized*" (p.97). So the project he sets himself is finding a reductive naturalistic account, that avoids intentional and semantic talk, of intentionality. Since on his view the "intentionality of the [propositional] attitudes reduces to the content of mental representations" (p.98), his task begins with the challenge of developing a thoroughly naturalistic psychosemantics. Hence, his causal covariance approach. The same sort of consideration motivates Dretske's covariance account. I agree with the sentiment, here, if not the solution.

I think that Fodor is right that intentionality is not one of the ultimate and irreducible properties of the world, and I agree that we need to be able to give an account of intentionality in non-intentional, non-semantic terms if we have any hope of integrating our theory of mind with the rest of science. Covariance theories do this by analyzing meaning in terms of causation. It's probably true that most |CAT|s are caused by cats. There is thus a very obvious relationship between the representation and the thing it's supposedly about. If we then wish to explain, in naturalistic terms, *how* |CAT|s come to be about cats, the causal relationship between the animals and the representations is an obvious first place to look. Again, I think there is an initial plausibility to

covariance views: |CAT|s are about cats because they're *caused* by cats. There is a sense of natural meaning here that Dretske emphasizes. The northerly flow of a river naturally means that there is a downward gradient in that direction; those spots on Tommy's face naturally mean that he has measles (Dretske, 1986, p.330). I think that covariance theories take their cue from observations like these. If natural signs represent because of certain lawful, natural relationships (like causation), then why not suppose that mental representation works this way (minds being part of the natural order, of course)? If a causal relationship obtains between the represented and the representation, perhaps that just *is* the meaning relation. If this were true, we would have a straightforwardly naturalistic account of representational content. There are still problems to be overcome for an account like this, of course, but it might nevertheless seem like a good start.

As I said, I agree with the sentiment that drives the covariance theorists, but I feel that they take the reduction of the intentional a bit too far. In reducing meaning to causation, Fodor and Dretske bypass what I take to be the crucial level of description for the semantic content of mental representations—the social, functional level—and wind up with a view that cannot explain the normative element of thinking. This, too, is illustrated most clearly in their solutions to the disjunction problem. Both Fodor and Dretske make misrepresenting a matter of malfunction. That is, on both of their views, to misrepresent something is to suffer a mechanical misfire of sorts: having told us how misrepresentation is possible, both philosophers take it for granted that occurrences of this sort are failings of the representing *mechanism*. In both cases what happens is that something which should cause tokens of one mental representation in fact causes tokens of another mental representation. This seems simply to be a brute fact about such

cases—the causal mechanism breaks down, and the chain that should go from, say, a cat to a |CAT|, instead goes from a dog to a |CAT|. Even if we can explain this failure at some deeper level (e.g., as the misfiring of certain neurons in the brain) the mistake is of the same kind: a simple physical error. This is, of course, a direct consequence of the psychosemantic theory both philosophers endorse. If *correct* representation is simply a matter of causation (no matter how complex the causal situation: asymmetric dependence, or different possible chains of causation), then *mis*representation will also be simply a matter of causation. Again, what gets left aside is the fact that at least one kind of representing (*mental* representing) is normative in nature, such that getting it right or wrong is not simply a matter of mechanism. When I mistakenly think |CAT| in the presence of a dog, I am open to the reproach that I ought not do such things; I ought to know better. (Though, of course, I may defend myself: It looked, for a moment, like a cat; now that I look closer I see my error.) On a covariance view the proper response of my peers to a case where I misrepresent the world is not to tell me that I ought not do such things, but rather to fix (if they are able) the causal mechanism that led to the misrepresentation in the first place. We don't need remonstration, we need engineering. After all, unless my mental representing is sensitive to some set of rules that govern thought, fixing the problem can't be a matter of trying to get me to follow such rules.

Again, I save the development and defense of the claim that thought is essentially norm-governed until Chapter Four, so a full analysis of covariance theories will have to wait until later. We should move on, then, to the second of the three main categories of psychosemantic theories that I mentioned above: the teleological approach.

2.22 Teleology

Teleological theories differ from covariance theories in that, while the latter appeal to the cause of a representation to determine its content, the former appeal instead to the biological or evolutionary function of the representation. Thus, for example, in the magnetotactic bacteria that we looked at above, the evolutionary function of the bacterium's magnetosome is, given that the bacterium will die in oxygen-rich water, to guide the bacterium to oxygen-free water—the biological *purpose*, if you will, of the magnetosome is to represent the direction of oxygen-free water. We might therefore say that a certain state of the magnetotactic system *means* 'oxygen-free water that way.' Dretske's use of the notion of a natural function is, of course, an appeal to teleology—but Dretske's theory is a covariance theory, not a teleological theory, because (ultimately) Dretske appeals to causation to determine representational content. A teleological theorist would not be happy with such an appeal, because according to such a theorist the content of a representation is determined by whatever biological purpose that representation is meant to serve, whether it is regularly *caused* by whatever it is meant to represent or not. In his book, *The Representational Theory of Mind*, Kim Sterelny puts it this way:

> [I]t can be the biological function of a device to represent **f** even when it usually misrepresents.... Predator representation is the usual example. A rabbit that thinks '**fox**' when there is no fox loses little; the rabbit that fails to think '**fox**' when there is one loses all. So it can be the function of a rabbitish state to *represent* foxes even though it does not *indicate* foxes. Its firings might be usually caused by wombats. It would thus indicate wombats, not indicate foxes, yet represent foxes. (1990, pp.123-124).

According to a teleological theory what is important is not what a representation causally covaries with (i.e., what it indicates), but rather what it is meant (biologically/evolutionarily) to represent.

Fodor and Dretske each attempt to solve the disjunction problem, but their success in doing so, while also providing a plausible account of representational content, is at least debatable (though we did not take the time to debate it). Teleological theories, on the other hand, overcome the disjunction problem easily; or, rather, that problem does not even arise for such theories, and this is at least one thing that speaks strongly in their favor. Yet it is the disjunction problem, I think, that helped to motivate the insight (if such it was) that produced the teleological approach to representational content, for the disjunction problem manifests itself in what is a very common occurrence: a representation that means one thing is tokened in a context inappropriate to its production. We would be hard pressed to argue that this does not happen all the time, in many different representational contexts. It may even be the case, as Sterelny suggests, that *mis*representation occurs more frequently (in some cases, at least) than correct representation. Even when the elaborations introduced to get a covariance theory around the disjunction problem seem to be somewhat successful, constant misrepresentation seems to remain very problematic on such views. But misrepresentation is a problem for covariance theorists because they rest meaning on simple (or perhaps, sometimes, not-so-simple) causation. That is, they take a representation to *mean* what it *indicates*, i.e., what caused it to occur, and it's very often the case that representations don't mean what they indicate—which is just to say, they misrepresent. So the insight of the teleological approach is to notice that *indication* is not the same as *representation*. (The needle's

resting on 'F' *represents* the tank as full of gas, even when what it *indicates* is that the tank is full of some liquid or other (not necessarily gas), or that there is a short in the fuel gauge system, or that the needle is stuck.)

The main motivation behind covariance views is the need for a naturalistic theory of intentionality, and I think the same motivation lies behind the teleological approach. The question is: How do you separate indication and representation without presupposing intentionality? After all, the reason (as we saw when we looked at Dretske's view) that the fuel gauge's needle's resting on 'F' represents the tank as full of gas is that *we intend* for it to represent just that fact, and no other. That is, we designed and deployed such gauges to represent the amount of fuel in the tank, not to represent the needle's being stuck, or any other such thing. It therefore *represents* the amount of fuel in the tank, even when it does not *indicate* the amount of fuel in the tank, only as a consequence of our intentions. A story like that, though, will just not do when what we are after is an account of representational content in non-intentional, non-semantic terms. So perhaps we've cast the net too narrowly when all we look at is brute causation. Perhaps we need to take a broader naturalistic approach. To develop this idea, let us turn to what is probably the most well known of the teleological theories, Ruth Garrett Millikan's view in 'Biosemantics.'

Of central importance to Millikan's view is a shift from representation *production* to representation *consumption*. As she puts it, "It is the devices that *use* representations which determine these to be representations and, at the same time (contra Fodor), determine their content" (1989, pp.402-403). One obvious effect of this shift away from representation production is that the *cause* of any particular representation becomes

irrelevant. What is important, for Millikan, is the *purpose* that the representation serves for the system in which it is a representation. Another way to put this would be to say that Millikan isn't concerned with how a representation comes to be in a system, she is only concerned with how that system *uses* the representation *as* a representation. She would not say, however, that any use is as good as any other. Millikan is concerned with the biological/evolutionary role of a representation, and that means that she is only interested in the cases in which a representation performs its *proper function*.

The notion of a 'proper function' is a technical one for Millikan, but Robert Cummins summarizes it nicely: "Something x performs a Proper Function in a system S when it does the sort of thing the doing of which has been, historically, responsible for the replication of things of x's type" (1989, p.76). For example, the magnetosome in a magnetotactic bacterium performs its proper function only when it indicates the direction of oxygen-free water, for it is only when it indicates the direction of oxygen-free water that the bacterium survives. The content of a representation is then determined by whatever the proper function of that representation is. We must notice, however, that proper function is an historical notion: a representation functions properly when it functions as it was (evolutionarily, for natural cases) designed to function. This means that it is important to pay attention to what Millikan calls the 'normal explanation' for a function as well as the 'normal conditions' for that function's performance.

A 'normal explanation' "explains the performance of a particular function, telling how it was (typically) historically performed on those (perhaps rare) occasions when it was properly performed" (1989, p.403). A 'normal condition' for the performance of a function is "a condition, the presence of which must be mentioned in giving a full normal

explanation for performance of that function" (p.403). So, the presence of oxygen-free water in the indicated direction is a normal condition for the proper functioning of the magnetotactic bacterium's magnetosome, for, again, it is only when there is oxygen-free water in the direction indicated that the bacterium survives. We must notice, too, that the sense of 'normal' used here by Millikan is not a statistical one: "'normal conditions' must not be read as having anything to do with what is typical or average or even, in many cases, at all common" (p.403). Many functions will only be properly performed under rare or uncommon conditions. Consider one of Millikan's examples: the tail-splashing behavior of beavers. When a beaver splashes the water in a certain way with its tail, this is meant to indicate to fellow beavers in the area that some sort of danger is present. That is, the proper function of the splashing behavior is to indicate the presence of danger, and a normal condition for the behavior's performing this function is the actual presence of some sort of danger. But it does not follow from this, says Millikan, that any sort of danger is *usually* present when beavers splash the water in this way: "Beavers being skittish, most beaver splashes possibly occur in response to things not in fact endangering the beaver" (p.405). That is, it may be the case that tail-splashing behavior by beavers only rarely performs its proper function. Nevertheless, such splashes *mean* danger, because that is what they are *supposed* to mean, what they *must* mean to help the beavers survive.

Consider again Sterelny's rabbit. The proper function of its |FOX| representation is to represent the presence of a fox nearby, for it is the performing of that function which makes the |FOX| evolutionarily advantageous to the rabbit. That is, the normal explanation for the proper functioning of a rabbit's |FOX|s is the nearby presence of a

fox, since only then does the tokening of |FOX| keep the rabbit from being eaten.[43] On most occasions, though, the tokening of |FOX| may not actually indicate the presence of a fox. Millikan separates representing and indicating by appeal to the teleological function of the representation.

Misrepresentation, then (as I've said) is not a problem for teleological theories. Just so long as we can say what the biological/evolutionary function of a representation is, we can say when its occurrence is a misrepresentation. This seems like it might be an advantage of teleological theories over covariance approaches. But, of course, teleological theories have their own problems. One of the main concerns with a teleological account of at least human representational content is the fact that many, if not most, of the representations that human beings have are original—that is, they *have no* evolutionary history to which we could appeal to determine their proper function. Sterelny raises this worry in *The Representational Theory of Mind*:

> Most of my beliefs, no doubt, are had for the very first time in human history by me. This is no tribute to my extraordinary and original genius. Rather it's due to the egocentric preoccupations of human belief. A large chunk of my beliefs are first person beliefs: beliefs that I have been and done such and such.... Furthermore, many of my other beliefs have only recently been possible: for they are about recent events. Very few human beliefs have been available to the ancestral population long enough to be the subjects of an evolutionary history. For the overwhelming majority of human beliefs, it's just not true that we have them because, in certain historically typical circumstances, the having of them improved our ancestor's fitness. (1990, p.129).

How can we say that it is the proper function of one of my representations to mean 'microwave oven'? That is, how could the presence of microwave ovens be a

---

[43] Perhaps in characterizing the representation as one that means 'fox' rather than just, say, 'predator' or even 'danger' we have assigned the representation too fine a content. Maybe rabbits aren't that discriminatory—after all, other creatures could eat rabbits, too. But the example serves *its* purpose even so.

(historically) normal condition for the proper functioning of such a relatively new representation? This worry is related to Dretske's worry that the notion of natural function cannot account for the complexity of human representations, and his solution to that problem is similar to the kind of solution one might pose to this current concern: appeal to the complexity and capacity for learning of the human representational system. Millikan writes:

> Unlike evolutionary adaptation, learning is not accomplished by *random* generate-and-test procedures. Even when learning involves trial and error ... there are principles in accordance with which responses are selected by the system to try, and there are specific principles of generalization and discrimination, etc., which have been built into the system by natural selection. How these principles normally work, that is, how they work given normal (i.e., historically optimal) environments, to produce changes in the learner's nervous system which will effect the furthering of ends of the system has, of course, an explanation—the normal explanation for proper performance of the learning mechanism and of the states of the nervous system it produces. (1989, p.407).

The point, I take it, is that in sufficiently complex representational systems, evolution is replaced (to a certain extent) by learning. Thus, it can be the proper function of a representation of mine to mean 'microwave oven' because the normal (historically optimal *during the learning process*) condition for the occurrence of this representation was the presence of microwave ovens. Only when the tokening of said representation corresponds to the presence of microwave ovens will I be able to heat my food. Of course, the proper functioning of the learning mechanism itself was (presumably) determined by natural selection.

Like Dretske, Millikan believes that the ability of human beings to represent new or novel states of affairs has to do with the complexity of our representational system. Dretske, recall, suggested that in order to identify predators, a system might develop in

such a way as to be able to discriminate color, shape, and movement. The ability to represent things other than predators, then, would be a happy side-effect of this development. Similarly, Millikan suggests that the representational capacities of human beings might have started out as a solution to some simple survival problem; but the solution was so powerful that we were subsequently able to represent all sorts of things completely unrelated to that original problem. She writes:

> Indeed, it is reasonable to suppose that the brain structures we have recently been using in developing space technology and elementary particle physics have been operating in accordance with the very same general principles as when prehistoric man used them for more primitive ventures. They are no more performing new and different functions or operating in accordance with new and different principles nowadays than are the eyes when what they see is television screens and space shuttles. (1989, pp.407-408).

So, again, according to Millikan, the proper function of the human cognitive mechanism may be such that it is able to represent all kinds of new things even though these representations have no evolutionary history.

Notice, however, that once again we have a theory of representational content that makes representation a matter of (for lack of a better word) mechanism. The change from focus on representation production to focus on representation consumption is, in my view, an important one. But it does not, by itself, move us out of the realm of simple mechanistic representing. Importantly, for my purposes, it is still the case that misrepresenting is a matter of simple misfire. On a teleological view, it's not even the case that misrepresentations are malfunctions of the representing system, since (a point we dwelled upon heavily) in some cases the system might function best when it misrepresents often (as, for example, when a rabbit survives because it tokens |FOX|

frequently, even though there are seldom any foxes around). We are still not in the realm of normative rules for thought. Though I often find myself attracted to the teleological view when it comes to the representational activity of simple organisms (like the magnetotactic bacteria) or animals (like rabbits), the teleological approach seems to me to still fall short of what is needed for *mental* representation—for *thinking*. Including the notion of function in our psychosemantics is, I think, a crucial step in getting to the level of mental representation—it's just that philosophers like Dretske and Millikan don't go far enough with it. Hence we turn to the last of the three major categories of psychosemantic theory: functional role semantics.

2.23 Functional Role

A functional role theory of mental representation is, in the words of Robert Cummins, "just functionalism applied to mental representations rather than to mental states and events generally" (1989, p.114). The basic idea behind functional role theories is, again (as with covariance and teleological theories), quite simple. According to functional role theories, the content of a mental representation is determined by the role it plays in the representational system—that is, by the position (as Cummins puts it) that it occupies in the cognitive network (p.114). Typically, this role is specified causally: a given representation is identified by its causes and its effects, i.e., by its being caused by certain states or events and its causing certain other states or events.[44] For example, a given representation in me might be a |CAT| because it is caused by seeing cats and thinking about my neighbor's pet, and it causes me to wonder if they really do taste like

[44] The 'states' in question could be states of the world (e.g., the color of the sky) or states of the representational system (e.g., representing the color of the sky). Similarly, the 'events' could be events in the world or mental events.

chicken. The content of that representation is determined by its function in my representational system. To flesh this idea out, let's look at a specific functional role theory: Ned Block's Conceptual Role Semantics.

Block's theory is what he calls a 'two-factor' theory. The basic idea behind a two-factor conceptual role theory is that

> there are two components to meaning, a conceptual role component that is entirely "in the head" (this is narrow meaning) and an external component that has to do with the relations between the representations in the head (with their internal conceptual roles) and the referents and/or truth-conditions of these representations in the world. (1994, p.93).

Part of the point, I take it, of having the second factor (a relation between the representation and the world) is to make sure that meaning and reference don't come completely apart. (According to Block, the two-factor approach "derives from Putnam's argument ... that meaning could not both be 'in the head' and also determine reference" (p.93)). Representational content is then determined by the total functional role (including both factors, both narrow and wide content) of the representation in the system.

Like functional role theories generally, Block's Conceptual Role Semantics (CRS) specifies the function of a representation causally, i.e., in terms of what states and events cause and what states and events are caused by the representation. For CRS, though, what defines a representation's conceptual role is not just *any* functional role that the representation might play; rather, it is the *conceptual* role that the representation plays that is important—that is, the role it plays in processes like inference. Block writes:

> The internal factor, the conceptual role, is a matter of the causal role of the expression in reasoning and deliberation and, in general, in the way the expression combines and interacts with other expressions so as to mediate between sensory inputs and behavioral outputs. A crucial component of a sentence's conceptual role is a matter of how it participates in inductive and deductive inferences. (P.93).

In many cases distinguishing between conceptual role (as just defined by Block) and functional role generally may not be important. As Robert Cummins points out, however, there is a possibly significant consequence of focusing on *conceptual* role: an explanation of representational content in such terms makes "no provision for representation in noncognitive systems" (Cummins, 1989, p.117). Depending upon how narrowly we construe the notion of a 'cognitive system' conceptual role semantics may end up having a rather limited scope. We can say that it applies to human beings for sure, but anything else is open to question.[45]

CRS and functional role theories generally are not subject to the same kinds of worries that face covariance and teleological theories. Misrepresentation isn't an issue: something is a misrepresentation according to CRS when it doesn't play the conceptual role that it is supposed to play. If being a |CAT|, in me, is in part mediating between my awareness of a cat in my visual field and my decision to make shooing noises, a |CAT| tokening in the presence of a dog (and the absence of a cat) is a misrepresentation. Nor does CRS have any trouble explaining new or novel representations (i.e., representations without any adaptive history). Something is a |MICROWAVE OVEN|, in me, if it, in part, mediates between (say) my desire to reheat some pizza and my decision to go into

---

[45] Cummins points this out because he is looking for a theory of representation that can serve an explanatory role in the computational theory of cognition (CTC), and he thinks that it is "perverse" from the standpoint of the CTC to tie representation essentially to cognition, since the CTC "takes cognitive representation to be of a piece with representation in computational systems generally" (1989, p.118).

the kitchen. That is, if it is caused (in part) by that desire and causes that decision. There are, however, other problems that shape CRS and functional role theories. I will mention one of them.

According to functional role theories, what makes a given representation a |CAT| is that it plays the right sort of functional role in the representational system: the |CAT| role. The problem is that the role a particular representation plays is defined by its place in the representational system: by its causes and its effects; but there is very good reason to suppose (indeed, it seems obvious) that no two representational systems will have all the same causal connections. In me a |CAT| caused by the presence of a cat in my visual field may cause me to make shooing noises, while a similarly caused |CAT| in you may cause you to make cooing noises. In me it may cause a desire to rid the world of the pesky beasts, while in you it may cause warm feelings of affection. Still, we would ordinarily want to say, the content of our representations is nevertheless the same: they both *mean* 'cat.' Yet, if content is determined by functional role, and your representation doesn't play the same role in your representational system as my representation plays in mine, then it would seem that our representations do not have the same content after all. Indeed, it would seem that no two thinking beings could ever be thinking about the same things. As Kim Sterelny puts it:

> Holistic theories of content make content idiosyncratic; people never act the same way because they believe the same thing, because they never do believe the same thing. Indeed, it is hard to see how I (or anyone) could explain an agent's behaviour by reference to belief content at all. For to do so I have to attribute intentional states to the agent. To do that, if inferential role determines content, I have to know the inferential role a belief has in the mind of the agent in whom I am interested. That, manifestly, is something that I am not within a bull's roar of knowing about any agent but myself. (1990, p.136).

One way to fix this problem might simply be to bring in the second of Block's two factors and appeal to the referential connection between the representation in question and the world. That is, part of the functional role of a representation is determined by what things in the world cause it to be tokened. And it's reasonable to suppose that these will be the same things for everyone (e.g., despite the different ways we might feel about the animals, actual cats are a cause of |CAT|s in both you and me). But we have to be careful, here. If, in determining representational content, we ignore everything in the functional role of a representation except for whatever it is in the world that causes that representation to be tokened, then we will have abandoned functional role semantics in favor of a covariance theory.

Another way to address this problem would be to define representations globally, rather than individually. That is, rather than looking at the functional role of a particular representation in an individual system, we could look at the functional role of that representation type in all the individuals in whom it is, or can be, tokened. |CAT| is thus defined in terms of all of the possible inputs and outputs it might mediate between in all the representational systems in which |CAT|s occur. When we then compare token |CAT|s across individuals, some of the inputs (e.g., being caused by the presence of a cat in one's visual field) will likely be the same for everyone, while others (e.g., compiling a mental list of animals that one can't stand) will differ. The same will be true of the possible outputs as well. What makes all these representations across individuals |CAT|s is the globally defined functional role. This avoids making content idiosyncratic, though at the cost of making it even more abstract and general.

I think it is appropriate at this juncture, however, to confess here that the theory of content I will defend for the LTT is, at root, a functional role theory. I will therefore have more to say about such theories in Chapters Three and Five. The view that I ultimately endorse bears a lot of resemblance to Block's CRS, though there are differences enough, I think, to set my own functional role theory apart from Block's. Nevertheless I find CRS quite attractive in many respects, including its wide content approach, and the fact that CRS recognizes, in its own way, that mental representations are something of a special case. The primary difference between Block's functional role semantics and my own, perhaps, is that Block never explicitly cashes out the claim that the psychosemantics of mental representation is determined by functional role with any kind of specific mapping of the functional network. I, on the other hand, will cash out the functional roles of the symbols that constitute thought in a very specific way: in terms of the function of the symbols of the language of the thinker. That is, on my view, the semantics of linguistic types is to be understood in terms of functional role (this is the topic of Chapter Three), and the mental representations that make up thought are constituted by symbols of the language, or languages, that one speaks (Chapters Four and Five are primarily concerned with making this case, and this is obviously the very heart of the idea that I am putting forward in this project).

Since we will have ample opportunity to discuss functional role semantics in the chapters to come, I will say no more about it here. Instead I propose to turn to a discussion of theories of meaning generally, and arguments for the theory of meaning that will form the cornerstone of the psychosemantics of the LTT. That theory is the

'functional classification' theory argued for by Sellars, and we must take some time to make a solid case for it. This I do in the next chapter.

3. Meaning

As we saw in the last chapter, a very important part of any representational theory of mind is its theory of semantic content, or psychosemantics. On the traditional view of the relationship between thought and language, a theory linguistic meaning will usually cite the (prior) meaningfulness of thoughts. (Whether any account of the meaningfulness of thoughts is given—or presumed needed—I'm not sure. It might simply be believed that it is the *nature* of thoughts to be meaningful—perhaps, even, that thoughts just *are* meanings.) This will not do, however, as a theory of meaning for the LTT, since the meaningfulness of thoughts is to be cashed out in terms of the meaningfulness of the linguistic symbols that constitute them. That is, for the LTT, the theories of linguistic meaning and of the meaningfulness of thoughts will be one and the same. So I'm going to need to develop a general theory of meaning that can pull double-duty, as it were, explaining both the meaningfulness of linguistic tokens as well as the meaningfulness of thoughts.

I will begin by taking a look at a theory of linguistic meaning that takes up the traditional view about the relationship between thought and language, and indeed appeals to that relationship to explain linguistic meaning. The theory I have in mind is Grice's. It is what Strawson has called a 'communication-intention' theory of meaning. According to such a theory the meaning of a linguistic token is determined by what a speaker intends to communicate to his listeners by his use of that token. (The theory is a bit more complicated than this, of course, as we will see below—but this is the basic idea.) Obviously, such a view takes for granted the idea that a speaker can have meaningful thoughts prior to putting those thoughts into words. In fact, it's the

meaningfulness of these prior thoughts that explains the meaningfulness of linguistic tokens. I will explore this view briefly to serve as a foil to the theory of meaning that I want to develop for the LTT.

Next I will introduce the basis for my own view of the meaningfulness of linguistic tokens (and, ultimately, thoughts): Sellars' claim that meaning is really just a matter of the functional classification of linguistic tokens. On this view to say that *x* means *y* is just to say that the token 'x' plays a role in the linguistic economy of which it is a part that is the same as the role played by 'y' in the linguistic economy of which *it* is a part. (Again, the actual view is more complicated than this, but this description serves well enough as a one-sentence summary of the position that Sellars develops.) In contrast to Grice, Sellars is not going to appeal to the communicative intentions of a speaker to determine the meaning of linguistic token because he does not treat meaning as a relation between a speaker and a linguistic token. Rather, he treats 'means' as a special form of the copula (as I will explain in more detail below). The immediate advantage for me, of course, with a view such as this is that it does not appeal to the prior meaningfulness of thoughts to explain the meaningfulness of linguistic tokens—and that is precisely what I need to avoid if the LTT is going to be successful.

This will lead me into the final section of this chapter where I will attempt to develop and adapt Sellars' theory of meaning to the needs of the Linguistic Theory of Thought. As we will see, while Sellars develops his theory of meaning as a rough approximation of how thoughts can be meaningful there is some question about how this is supposed to work within his overall philosophy of mind. I will, therefore, try to put the right pieces together, modifying Sellars' view when necessary to develop a view wherein

the meanings of the linguistic tokens that constitute our thoughts are fixed by the functions of those tokens in the language in question. That is, I will attempt to develop a theory of linguistic meaning that doesn't require an appeal to the prior meaningfulness of thoughts, which I will then be able to use to explain the meaningfulness of our thoughts according to the LTT.

3.1 Gricean Semantics

The most simple and straightforward statement of Grice's view comes from his 1957 piece "Meaning."[46] Grice opens his paper by distinguishing between what he calls natural meaning and non-natural meaning ($meaning_{NN}$). That is, he distinguishes between two different senses of 'mean.' Natural meaning is exemplified by the use of 'mean' in the following sentence:

(1) Those spots mean measles.

Here 'mean' is used to signify that the spots are a natural indication of one's having measles (or something like that). Those spots *mean* measles because they are caused by one's *having* measles. It's probably the case that one has those spots if and only if one has measles. Presumably most (if not all) cases of natural meaning are tied closely to causation like this, and have nothing whatever to do with convention. As Grice says, "I cannot argue from 'Those spots meant measles' to any conclusion to the effect that somebody meant by those spots so-and-so" (1957, p.92). The person who has measles did not intend, by having those spots, to let others know that he has measles. Nor did

[46] He develops the view further in his 1968 "Utterer's Meaning, Sentence-Meaning, and Word-Meaning", but I find the 1957 essay clearer and more straightforward.

anyone else intend the spots to mean that he had measles.[47] Intentions just have nothing to do with natural meaning. Non-natural meaning, on the other hand, seems to depend crucially on an agent's intentions. $Meaning_{NN}$ is exemplified by the use of 'mean' in the following sentence:

> (2) Those three rings on the bell (of the bus) mean that the bus is full.

Here 'mean' seems to make essential reference to the intentions of some individual (in this case, the bus driver). It's clear that the bus driver means by the three rings on the bell that the bus is full. Of course, he might be mistaken about this, so we could say "Those three rings on the bell (of the bus) mean that the bus is full, but as a matter of fact it isn't full."[48] Nevertheless, the rings on the bell mean that the bus is full in part, at least, because the bus driver intends to communicate just that idea to those who hear the bell. This is the idea that forms the heart of Grice's 'communication-intention' theory of meaning. Strawson characterizes the view succinctly. According to a communication-intention theory of meaning,

> it is impossible to give an adequate account of the concept of meaning without reference to the possession by speakers of audience-directed intentions of a certain complex kind. The particular meanings of words and sentences are, no doubt, largely a matter of rule and convention; but the general nature of such rules and conventions can be ultimately

---

[47] Obviously, neither the person who has measles, nor anyone else, could mean anything by those spots, since neither the person who has measles, nor anyone else, has any control over the occurrence of those spots. They appear if one has measles, and do not appear if one does not have measles. The possible intentions of sentient beings are simply irrelevant to the natural meaning of the spots. This does not, of course, preclude the possibility of someone *using* the spots (or the person who has them) to mean something else—to serve as a signal of some type, say (e.g., the presence of the person with the spots at a particular place and time meaning that it's time to start a certain activity). Even if the spots were used in a manner such as this, however, they would *still* have their natural meaning—that he has measles—and no one's intentions would have anything to do with the spots having *this* meaning.

[48] Contrast this, as Grice does, with the equivalent change to (1): "Those spots mean measles, but as a matter of fact he hasn't got measles." Grice rightly claims that we just can't say this. If he hasn't *got* measles, then those spots can't *mean* measles.

> understood only by reference to the concept of communication-intention.... For any theorist who follows this path, the fundamental concept in the theory of meaning is that of a speaker's, or, generally, an utterer's, *meaning something by* an audience-directed utterance on a particular occasion.... What an utterer means by his utterance is incidentally specified in specifying the complex intention with which he produces the utterance. (1970, pp.110-111).

This seems to me to be a correct characterization of Grice's theory in "Meaning." Grice puts his conclusions this way:

> (1) "*A* meant_{NN} something by *x*" is (roughly) equivalent to "*A* intended the utterance of *x* to produce some effect in an audience by means of the recognition of this intention"; and we may add to that to ask what *A* meant is to ask for a specification of the intended effect ....
> (2) "*x* meant something" is (roughly) equivalent to "Somebody meant_{NN} something by *x*." ...
> (3) "*x* means_{NN} (timeless) that so-and-so" might as a first shot be equated with some statement or disjunction of statements about what "people" (vague) intend (with qualifications about "recognition") to effect by *x*. (1957, p.96).

The fundamental idea should be clear. A (meaningful_{NN}) linguistic token gets its meaning_{NN} by appeal to the intentions of the utterer who produces that token. Thus, e.g., if I say "Snow is white" that sentence means_{NN} that snow is white because that is what I intended to communicate by uttering that sentence.

Why would anyone think that this could be the correct way to analyze meaning? I will take a brief look, here, at Grice's arguments for this position, though I should note that as far as the actual position goes, it may be that no one still thinks that such an approach to meaning will work.[49] I'm interested in discussing this view not because it seems to many (or to anyone) to be a plausible theory of linguistic meaning, but rather because it exemplifies a particular attitude towards the relationship between thought and

[49] See Carruthers (1996), pp. 76-82; Fodor (1998), p. 66.

language that I *do* think is still widely accepted. With that caveat out of the way, let me begin with Grice's arguments for adopting a communication-intention view.

The goal, of course, is to explain meaning$_{NN}$. Grice moves in small steps toward his final position (in "Meaning"), beginning with the idea that meaning$_{NN}$ has simply to do with the result (commonly) produced by the sign in question. (Grice calls this the 'causal' view, and it bears at least some resemblance to the causal semantics of Fodor and Dretske discussed in Chapter Two above.) The idea is simply that "for *x* to mean$_{NN}$ something, *x* must have (roughly) a tendency to produce in an audience some attitude (cognitive or otherwise) and a tendency, in the case of a speaker, to *be* produced *by* that attitude" (Grice, 1957, p.93). But many things can be said to have a tendency to produce in an audience some attitude without our being tempted to say that they are thereby meaningful$_{NN}$. Consider Grice's (somewhat outdated) example:

> It is no doubt the case that many people have a tendency to put on a tail coat when they think they are about to go to a dance, and it is also no doubt the case that many people, on seeing someone put on a tail coat, would conclude that the person in question was about to go to a dance. Does this satisfy us that putting on a tail coat means$_{NN}$ that one is about to go to a dance (or indeed means$_{NN}$ anything at all)? Obviously not. (P.93).

Consider a similar (but less dated) example. People probably have a tendency to put on sunscreen when it is hot and sunny outside, and they are going out. And seeing someone put on sunscreen might very well have the tendency to cause an audience to think that it must be hot and sunny outside, and that the person putting on the sunscreen is going out. But does putting on sunscreen mean$_{NN}$ that it's hot and sunny outside? To suggest that it does sounds (to my ears at least) rather absurd. Putting on sunscreen is something people do, but it doesn't *mean* anything at all. It's simply an action in response to a particular

state of the environment. So the mere tendency of *x* to produce in an audience a particular attitude does not seem sufficient to grant *x* a meaning$_{NN}$.

The next suggestion that Grice entertains is "that '*x* meant$_{NN}$ something' would be true if *x* was intended by its utterer to induce a belief in some 'audience' and that to say what the belief was would be to say what *x* meant$_{NN}$" (p.94). While Grice will go on to reject this formulation (for reasons that we will see in a moment) I want to call attention to the fact that it looks like he is already supposing the traditional view of the relationship between thought and language. Here he suggests that an utterance, '*x*', might get its meaning from the *belief* that it expresses. That is, he is appealing to a (somehow) meaningful belief to explain the meaning of an utterance. And that is precisely the kind of view that won't work for the LTT.[50]

Grice uses counterexamples to show that this formulation will not do any better than the 'causal' view. For example:

> I might leave *B*'s handkerchief near the scene of a murder in order to induce the detective to believe that *B* was the murderer; but we should not

[50] It may be worth mentioning that Gilbert Harman wrote an article in 1971 entitled "Three Levels of Meaning" in which he distinguishes between theories of (1) the meaningfulness of thoughts, (2) the meaningfulness of linguistic expressions, and (3) speech acts (promising, christening, etc.). He does so because he wants to claim that all the various 'theories of meaning' are not all attempts to answer one and the same question. Some theories (e.g., Grice's) attempt to define level 2 meaning; some (e.g. Sellars') attempt to define level 1 meaning; still others (e.g., Austin's) attempt to define level 3 meaning; and when these theories appear to come into conflict they are not really, according to Harman, in conflict at all. Rather, they are simply answers to different questions. I should say here that I am somewhat sympathetic to Harman's analysis (as is Sellars, quite explicitly, in "Meaning as Functional Classification"). But it seems clear that to make room for the LTT I cannot simply adopt Harman's three levels of meaning, claim to be working on a theory of meaning of level 1, and hence simply go around Grice's view (and others like it). For, according to my view, a theory of meaning of level 1 *just is* a theory of meaning of level 2 (more or less). I would like to thank S. Marc Cohen for first pointing out to me that I'd get myself in trouble if I tried to embrace Harman's distinction wholesale. Still, I have some affinity for the suggestion that when we ask what something means our question can have at least a couple of different senses. I will embrace this suggestion more fully in Chapter Four.

> want to say that the handkerchief (or my leaving it there) meant$_{NN}$ anything or that I had meant$_{NN}$ by leaving it that *B* was the murderer. (p.94).[51]

Again, I agree with Grice: leaving the handkerchief at the scene of the crime seems to fail to mean$_{NN}$ anything. There is an intention here (to cause a detective to come to a certain conclusion), but that doesn't seem like enough to give the leaving of the handkerchief at the scene of the crime a meaning$_{NN}$.

The step from here (a simple appeal to an utterer's intentions) to Grice's more-or-less final position is fairly short and rather than carefully trace out the thoughts that Grice goes through I'll just summarize the crucial point so that we can get on to reflecting on just what Grice's position seems to presuppose. The difference between the above position as an attempt to analyze meaning$_{NN}$ and Grice's final position is this: in the final analysis Grice takes it as crucially important that the intention with which someone makes an utterance be recognized by the audience. That is, " '*A* meant$_{NN}$ something by *x*' is (roughly) equivalent to '*A* intended the utterance of *x* to produce some effect in an audience *by means of the recognition of this intention*' " (p.96, emphasis mine). The above leaving of the handkerchief at the scene of the crime is, therefore, excluded from having meaning$_{NN}$ because, while I might have had the intention to produce a belief in the detective by my action, it's actually crucial to the production of the desired effect in this case that the 'audience' (the detective) *not* recognize my intention. On the other hand, if I say to you, "It's raining outside; you'd better take your umbrella with you," my

[51] It's probably worth noting here how broadly Grice is interpreting the idea of an 'utterance'—since he is apparently willing to count leaving a handkerchief somewhere (at least, leaving it there with a certain intention) as an utterance. While I will interpret 'linguistic meaning' quite broadly (to include symbol tokenings of all kinds: spoken, written, gestured, or thought), I think including intentionally leaving false evidence at the scene of a crime is a little too broad—for the simple reason that it does not seem to me that there is a *symbol tokening* in this instance. (Or, rather, while there are many symbol tokenings going on in the vicinity, the *act of leaving the handkerchief* is not, itself, a symbol tokening.) This difference between Grice and me, though, I doubt really comes to much in the end. Nevertheless, it does seem worth noting.

utterance means$_{NN}$ what it does (on Grice's final analysis) because I intend to produce in you a certain effect (the belief that it's raining and that you ought to take an umbrella with you, say) and I intend for you to recognize that my words are intended to produce just that effect in you.

Now, as an analysis of meaning, Grice himself admits that this is rather rough. And, as I said above, it has been strongly criticized over the years, and it may be that no one holds such a view anymore. However, there is one thing about this view that it seems to me most people *have* continued to hold on to (both in philosophical and in more everyday contexts), and that is the idea that the meaning of a linguistic utterance (of any kind) is determined by appeal to the antecedent *thoughts* of the utterer. That this idea is at the heart of Grice's theory of meaning should be clear, for he claims that we fix meaning by appeal to the *intention* of the utterer in making the utterance and the *recognition* of that intention in the mind(s) of the audience. What could be more obvious, then, than that Grice believes linguistic meaning (as the most common type of utterance, we might suppose) to be determined by the prior meaningfulness of thoughts?

The purpose of this discussion is not, however, to simply call attention to what, upon even casual reflection, seems to be a central presupposition of Gricean semantics. Rather, the purpose I have here is to illustrate, by means of a look at Grice's analysis of meaning, how *unquestioned* this presupposition seems to be. Grice himself never once seems to consider that meaning (that is meaning$_{NN}$) could be determined by anything *other than* an appeal to the thoughts of an utterer. That is, he seems to presuppose, without question, that thoughts are meaningful and that any attempt to explain the meaningfulness of linguistic tokenings will have to appeal to the (apparently

foundational) meaningfulness of thoughts. Nor, I contend, is this some idiosyncratic feature of Grice's analysis. Rather, it seems to me that the presupposition that thoughts are meaningful prior to, and explain the meaningfulness of, words is so completely unquestioned in Grice's theory precisely because it seems to so many to be the obvious way to understand the relationship between thought and language.

It is this presupposition, as I have now said many times, that is my real target here. Of course, the fact that a particular assumption shows up, and is taken for granted, in one philosophical analysis of meaning by no means demonstrates that that presupposition is everywhere prevalent. But a few moments reflection on the part of the reader will, I am sure, make it clear that I am not jumping to conclusions in asserting that Grice's assumption is one common to most people. If this happens not to be a presupposition that you share yourself, so much the better for me in what is about to follow.

3.2 Sellars' "Meaning as Functional Classification"

As I mentioned earlier, in his essay, "Meaning as Functional Classification," Sellars explicitly endorses Gilbert Harman's distinction between three levels of meaning: a level pertaining to thoughts, a level pertaining to linguistic tokens, and a level pertaining to full-blown speech acts. Sellars then proceeds immediately to make clear that what he is up to is the development of a level one theory of meaning. This actually turns out to be somewhat problematic, as Sellars then proceeds to spend most of his time talking about the meaning of linguistic tokens (which would make the theory he is developing a level two theory, on Harman's view). For our purposes, however, we can

ignore this point of potential trouble; as I have also already said, the theory of meaning that I am here going to develop for the LTT is at once *both* a level one and a level two theory of meaning. It therefore makes something of a mess of Harman's distinction right from the start. What is important, though, is not how we classify the theory, but what that theory says. And in this regard the thing to focus on is this: in "Meaning as Functional Classification" Sellars is developing a theory of meaning that, quite unlike Grice's, or most others', does not require us to presuppose the prior meaningfulness of thoughts. The theory actually attempts to explain the meaningfulness of thoughts themselves. What I'm going to do here is take a look at Sellars' arguments; in the final section of this chapter, then, I will adapt those arguments into a theory of meaning that will work for the LTT.

Sellars' approach to thoughts in general is to treat them much like I do in the LTT: he espouses what he calls Verbal Behaviorism (VB). According to this view, "thinking 'that-*p*,' where this means 'having the thought occur to one that-*p*,' has as its *primary* sense *saying* '*p*'; and a *secondary* sense in which it stands for a short term proximate propensity to say '*p*' " (1974, pp.418-419). There are differences between Sellars' VB and the LTT, but they are not important to us at this stage. At a basic and oversimplified level, Verbal Behaviorism treats *thinking* as *saying* (or, as Sellars often puts it, *thinking-out-loud*). The question then becomes how to specify the meaning of what is *said*, i.e., thought-out-loud. I think that it is extremely important to recognize that the 'sayings' that the verbal behaviorist is concerned with are thinkings-out-loud, that is, *thoughts*. Their meaning, then, is not primarily a matter of communication at all (in stark contrast to Gricean semantics). There is a way in which the social aspect of language is

built into the use of language in thinking-out-loud, but the sayings, or more generically, as Sellars sometimes puts it, 'languagings' (since *saying* is only one mode of linguistic expression) that constitute thoughts are not produced for the purposes of communicating thoughts or engaging in complex social rituals, etc. They are themselves simply thoughts. Hence, as with the LTT, no account of meaning which (like a CI theory of meaning) appeals to the intentions of agents could satisfy a verbal behaviorist's needs.

The core of Sellars' position, then, is this. Meaning is not a relation between a linguistic token and a non-linguistic item of some kind (including speaker's intentions), but rather a way of classifying the function played by that token in the language of which it is a part. That is, to give the meaning of a linguistic token is to specify the role that it plays in the language game. This

> functional classification involves a special (illustrating) use of expressions with which the addressee is presumed to be familiar, i.e. which are, so to speak, in his background language. Some of the functions with respect to which utterances are classified are purely intra-linguistic (syntactical), and, in simple cases, are correlated with formation and transformation rules as described in classical logical syntax. Others concern language as a response to sensory stimulation by environmental objects—thus, candidly saying, or having the short term propensity to say, 'Here is a penny', or 'This table is red'. Still others concern the connection of practical thinking with behavior. All these dimensions of functioning recur at the metalinguistic level in the language in which we respond to verbal behavior, draw inferences about verbal behavior and engage in practical thinking about verbal behavior—i.e. practical thinking-out-loud (or propensities to think-out-loud) about thinking-out-loud (or propensities to think-out-loud). (1974, p.421).

We specify the meaning of, e.g., 'Here is a penny' by specifying the role that this expression plays in our language as a response to stimuli of a certain sort. Similarly, we specify the meaning of, e.g., 'I was thinking about my penny' by specifying the role that this expression plays in our language as a bit of thinking (i.e., thinking-out-loud) about

thinking (i.e., thinking-out-loud). *Thinking*, 'Here is a penny' is, in a verbal behaviorist sense, just like *saying* 'Here is a penny.' And to describe someone as saying, 'Here is a penny' is to describe his linguistic behavior as playing the role played in English by the signs within the quotes while specifying speech as the modality of that behavior. Let's slow down and see how Sellars spells this out.

In Section IV of "Meaning as Functional Classification" Sellars asks the following question: "How does 'that-*fa*' function in 'Jones says that-*fa*' (where 'says' is used in the sense of 'thinks-out-loud')?" (p.426). In order to begin to answer this question Sellars asks another: "How does ' "*fa*" ' function is 'Jones says "*fa*" '?" (p.426). The answer, according to Sellars, is that " '*fa*' " functions

> as an adverbial modifier of the verb 'says'. Language can be written, spoken, gesticulated, etc., and 'says' serves to pin down the modality of a languaging to utterances. If speech were the only modality, or we abstract from a difference of modality, we could replace
>
> Jones says '*fa*'
>
> by
>
> Jones '*fa*'s
>
> i.e. use the expression-cum-quotes as a verb. (P.426).

Sellars then claims that the

> expressions " '*f* ", " '*a*' ", " '*fa*' " ... are sortal predicates which classify linguistic tokens. The classification is partly *descriptive*, thus in terms of shape (or sound) and arrangement. It is also and, for our purposes, more importantly *functional*. Above all, the sortal predicates are 'illustrating'. Thus
>
> *t* is an '*f*'
>
> tells us that *t*, belonging to a certain language L, is of a *descriptive* character falling within a certain range of which the design of the item

> within the single quotes is a representative sample, and also tells us that (if *t* is in a primary sense an '*f*', i.e. is produced by a thinking-in-writing [since we're now dealing explicitly with inscriptions]) it is functioning as do items having such designs in language L. (P.427).

Obviously, the sortal predicates themselves are considered to belong to a background language "the ability to use which is presupposed" (p.427). It will be easier to see just how this works if we (a) contrast classifying words of a foreign language with classifying words of our own language, and (b) introduce Sellars 'dot-quote' convention.

Consider the following two sentences, which might be thought of as 'definitions' in a classical sort of linguistic context.

(3) '*Oder*' (in German) means *or*.

(4) 'Triangle' (in English) means *three sided polygon*.

In both cases the word in single quotes is said to mean what the italicized words at the end of each sentence are already understood to mean. That is, to borrow a metaphor from Quine, the word in single quotes is said to be a label for the same thing (object or mental entity) which the word(s) in italics also label(s). This, I think, is the classical and intuitive sense of meaning: meaning is a relation between a word and an object (physical, or abstract, or mental).[52] Sellars, however, denies this account of meaning, and would claim instead that what we're really doing in (3) and (4) is giving an illustrating functional classification of the words in single quotes. That is, what (3) says is not that '*oder*' and 'or' pick out, or label, the same thing (whatever that might be), but rather that '*oder*' functions in German as 'or' does in English. Sameness of meaning is being traded for sameness of function.

---

[52] I've borrowed the metaphor here from Quine, but it is of course a picture of meaning that Quine rejects. He uses the 'label' metaphor about meaning pejoratively. I shouldn't want anyone to think me confused on this point.

Consider another sentence:

(5) '*Und*' (in German) means *and*.[53]

Sellars notes that (5) involves

> an atypical use of the word 'and', for it is clearly not functioning as a sentential connective. A natural move is to construe the context as a quoting one. This idea may tempt one to rewrite [(5)] as
>
> [(5a)] '*Und*' (in German) means 'and'
>
> but quoting contexts are often such that to leave them unchanged while adding quotes to the quoted item changes the sense. And it is clear that [(5)] doesn't merely tell us that 'und' and 'and' *have the same meaning*, it in some sense *gives* the meaning. (P.431).

It gives the meaning by telling us that '*und*' functions in German much as 'and' functions in English. In order to more clearly represent this idea, however, Sellars has introduced the convention of using 'dot-quotes'. That is, the correct analysis of (5) is

(5b) '*Und*'s (in German) are ·and·s

"where to be an ·and· is to be an item *in any language* which functions as 'and' does *in our language*" (p.431, emphasis added). The proper analysis (on Sellars' view), then, of (3) is

(3a) '*Oder*'s (in German) are ·or·s

which says of '*oder*'s that they function in German as ·or·s, where the "criteria which an item must satisfy to be an ·or· are a matter of its functioning, in respects deemed relevant as do 'or's in the illustrating language, in the present case a professional dialect of English" (p.428).

---

[53] Though (3) and (5) are more-or-less Sellars' own examples, they are not numbered '(3)' and '(5)' in "Meaning as Functional Classification." I have changed the numbering (and the subsequent modifications, like '(3a)' and '(5b)' below) to fit into this Chapter of my own work. Furthermore, sentence (4) is not an example that Sellars himself considers (though I think he'd agree with my treatment of it).

If, though, (3a) and (5b) give the correct analyses of 'means' in (3) and (5) respectively, then the correct analysis of (4) is

(4a) 'Triangle's (in English) are ·three sided polygon·s

which says that 'triangle' functions, in English, much as 'three sided polygon' functions, *in English*. While (3a) and (5b) may not immediately sound strange (unless we have strong previous commitments concerning the analysis of 'means'), I am fairly sure that (4a) sounds very strange indeed. After all, (4a) seems simply to tell us that one English expression functions in the same way as another English expression. But, it might be objected, this is useful as a definition (and, after all, isn't that what 'means' statements are primarily for?) only if I *already* know what one of the expressions *means*—otherwise I'm simply being informed that two English expressions serve the same function while being left in the dark as to what that function is. (And notice that once this objection is on the table, it would seem to apply equally well to (3a) and (5b), too.)

This objection is in the spirit of an objection that Sellars considers to Verbal Behaviorism generally. He writes:

> Surely, it will be said, thinking that-*p* isn't just saying that-*p*—even candidly saying that-*p* as you have characterized it. For thinking-out-loud that-*p* involves *knowing the meaning* of what one says, and surely this is no matter of producing sound! (P.429).

In response, Sellars writes:

> To this the obvious answer is that there is all the difference in the world between parroting words and thinking-out-loud in terms of words. The difference however, is not that the latter involves a non-linguistic 'knowing the meaning' of what one utters. It is rather that the utterances one makes cohere with each other and with the context in which they occur in a way which is absent in mere parroting. Furthermore, the relevant sense of 'knowing the meaning of words' ... must be carefully distinguished from knowing the meaning of words in the sense of being

> able to talk about them as a lexicographer might—thus, defining them. Mastery of the language involves the latter as well as the former ability. Indeed they are *both* forms of *know how*, but at different levels—one at the 'object language' level, the other at the 'meta-language' level. (Pp.429-430).

The solution, that is, rests on distinguishing between merely reproducing sounds in a parroting manner, and producing them in accordance with (and sometimes in recognition of) the rules for the correct use of those words. This distinction will be central to the next chapter, and will be discussed in greater detail there, but I will summarize it here, briefly.

Sellars begins Section II of "Meaning as Functional Classification" with a clear illustration of just what he means by the difference between parroting and thinking-out-loud. He writes:

> One can imagine a child to learn a rudimentary language in terms of which he can perceive, draw inferences, and act. In doing so, he begins by uttering noises which *sound like* words and sentences and ends by uttering noises which *are* words and sentences. We might use quoted words to describe what he is doing at both stages, but in the earlier stage we are classifying his utterances as *sounds* and only by courtesy and anticipation as *words*. Only when the child has got the hang of how his utterances function in the language can he be properly characterized as saying 'This is a book' or 'It is not raining' or 'Lightning, so shortly thunder'. (P.421).

Crucial to understanding what Sellars is envisioning are the concepts of pattern-governed linguistic behavior, and what he calls 'ought to be' and 'ought to do' rules of behavior. Sellars thinks of all linguistic behavior as pattern-governed, by which he means that the behavior "exhibits a pattern, not because it is brought about by the intention that it exhibit this pattern, but because the propensity to emit behavior of the pattern has been selectively reinforced, and the propensity to emit behavior which does not conform to this pattern selectively extinguished" (p.423).

There are many examples of pattern-governed behavior in nature, such as the tail-slapping reaction of beavers to danger, the behavior of digger wasps when laying their eggs, or (Sellars' example here) the "so-called language of bees" (p.423). Any non-intentional behavior that has been naturally selected for exhibits the pattern-governed quality that Sellars is talking about.

His interesting claim, though, is that linguistic behavior is also pattern-governed. That's not to say, however, that linguistic behavior is always selected for 'naturally': If pattern-governed behavior "can arise by 'natural' selection, it can also arise by purposive selection on the part of trainers" (p.423). The pattern-governed linguistic behavior of a child learning language is not 'naturally' selected for, but rather selected for by the child's trainers. The trainers follow what Sellars calls 'ought to do' rules to bring it about that the child conforms to 'ought to be' rules of behavior.

To understand the distinction, imagine the trainers reasoning something like this: "Patterned-behavior of such and such a kind *ought to be* exhibited by trainees, hence we, the trainers, *ought to do* this and that, as likely to bring it about that it *is* exhibited" (Sellars, 1974, p.423). For example, it ought to be that speakers of English respond (*ceteris paribus*) to red objects with 'This is red'; therefore, full-fledged speakers of English ought to do whatever they can to make it the case that fledgling speakers *do* respond to red objects with 'This is red.'

As we can see, the distinction is quite simple (at least on its face): linguistic behavior is governed by ought to be rules, and the training of fledgling language users is governed by corresponding ought to do rules. We should, at this point I think, now have enough of an understanding of where Sellars is coming from to be able to see the

important aspects of the difference between merely parroting language and thinking-out-loud with language. (There are, of course, a number of issues that arise from Sellars' talk of 'ought to be's and 'ought to do's (and I deal with some of them in the next chapter), but we needn't worry about them here. All we need at this stage is the basic idea, which we now should have.)

The difference between mere parroting and thinking-out-loud in terms of words turns on the level of mastery that a language learner has gained over his own pattern-governed linguistic behavior. When a child is first learning language he is merely parroting, producing noises that are only by courtesy considered (by his trainers) to be words and sentences. The child is conforming to ought to be rules of linguistic behavior, but his conformity is still somewhat sketchy (not as reliable, perhaps, as it needs to be for us to consider the child to be using words and sentences, as opposed to simply making mere noises); and (perhaps importantly) the child is conforming to the ought to be rules without knowing the ought to *do* rules (i.e., he is not yet in a position to be a trainer himself).

Once the child has got the hang of the language, though—once his patterned behavior reaches a certain level of complexity (just where to draw the line, or whether there is a line versus perhaps something like a gradually more shaded area, is one of the issues that I try to deal with in Chapter Four)—his utterances become truly meaningful; he now not only conforms to the ought to be rules but has some mastery of the ought to do rules as well. Mastery of the ought to do rules is important not primarily so that the new full-fledged speaker can train *other* language learners, but so that he can now train (i.e., correct) *himself.*

Thus, while (4a) does just tell us that one English expression functions in much the same way as another English expression, this is not as empty as it first seems. For to 'know the meaning' of *any* English expression just is to know how it functions, i.e., to have been conditioned to produce (in a thinking-out-loud sense) that expression in just the right circumstances (e.g., to think 'There is a three sided polygon' or 'There is a triangle' when looking at one). The full-fledged speaker of English who knows the meaning of 'three sided polygon' conforms to the ought to be rule that (*ceteris paribus*) three sided polygons ought to be responded to with 'This is a three sided polygon'; and he can also follow the ought to do rule that one should bring it about that English speakers *do* respond (*ceteris paribus*) to three sided polygons with 'This is a three sided polygon.' But if I 'know the meaning,' in this Sellarsian sense, of 'three sided polygon' then what (4a) tells me is just as useful as one might first believe (4) to be. (4a) tells me that 'triangle' plays the same role as 'three sided polygon'; hence, it ought to be that English speakers respond to three sided polygons (thus, triangles) with 'This is a triangle', and I ought to do whatever I can to bring it about that English speakers *do* respond to three sided polygons (thus, triangles) with 'This is a triangle.' The objection that (4a) is unhelpful as a guide to 'meaning,' and hence that (3a) and (5b) are similarly unhelpful, is bypassed once one understands Sellars' approach to language learning.[54]

I do not, of course, mean to suggest that I think that there are no issues left surrounding Sellars' theory of meaning once the objection that (4a) is only informative if

[54] I realize that this response is probably somewhat unsatisfying, since I have really brushed by all the issues surrounding Sellars' position on language learning, ought to be and ought to do rules, and the like. Satisfaction, I'm afraid, will have to wait until Chapter Four. In lieu of the actual discussion, I insert this promissory note that I do deal with a thorough defense of this part of Sellars' philosophy in the next chapter. I ask the reader for leniency in allowing these issues to pass undefended at this point.

one already 'knows the meaning' of 'triangle' or of 'three sided polygon' has been defeated. On the contrary, once we see how Sellars responds to this objection (by bringing in the difference between merely parroting and thinking-out-loud) we can return to the main topic of discussion: Sellars' suggestion that "to say what an expression means is to classify it functionally by means of an illustrating sortal" (p.431). Sellars writes:

> According to this analysis, *meaning is not a relation* for the very simple reason that 'means' is a *specialized form of the copula.* Again, the meaning of an expression is its 'use' (in the sense of function), in that to say what an expression means is to classify it by means of an illustrating functional sortal. (P.431).

To say that

(3) '*Oder*' (in German) means *or*

is to say that

(3a) '*Oder*'s (in German) are ·or·s.

To say that

(4) 'Triangle' (in English) means *three sided polygon*

is to say that

(4a) 'Triangle's (in English) are ·three sided polygon·s.

Sellars rejects the classical, and perhaps intuitive, sense of meaning that meaning is a relation between a word and an object (physical, or abstract, or mental). Rather, on his view, to give the meaning of a word is to say what kind of piece that word *is* in the language game of which it's a part; i.e., to say how it functions by referring to it with an illustrating functional sortal.

Where does this leave us? We began by noting that what Sellars explicitly claims to be doing is developing a theory that can explain the meanings of the linguistic symbols

that constitute our thoughts (within the 'course-grained' framework of Verbal Behaviorism, for Sellars; within the context of the LTT, for my purposes) without appealing to antecedent *thoughts*. What we have now is the suggestion that this task can be accomplished by treating meaning as a matter of functional classification (i.e., as a matter of designating the role played by a word in the linguistic system of which it is a part), and 'means' as a special form of the copula. In explicating this last point, we've been looking at translational and definitional sentences like (3) and (4) above. But it is important to notice, as Sellars points out, that talking about meaning through the use of illustrating functional sortals is not the only way to 'give' the meaning of a word or expression in functional terms. Sellars writes:

> Notice that instead of 'giving' the complex function of 'und' (in German) by using an illustrating functional sortal [·and·], we could, instead, have listed the syntactical rules which govern the word 'und' in the German language. In general the rule governed uniformities which constitute a language (including our own) can be exhaustively described without the use of meaning statements.... In practice, the use of meaning statements (translation) is indispensable, for it provides a way of mobilizing our linguistic intuitions to classify expressions in terms of functions which we would find it difficult if not (practically) impossible to spell out in terms of explicit rules. (Pp.431-432).

The syntactic, ought to be, ought to do, etc., rules that govern a linguistic system and determine what role any piece of linguistic behavior plays in that system also give the 'meanings' of the pieces in question. Since the verbal behavior (e.g., saying) just is thinking (e.g., thinking-out-loud), and the meaning of those thoughts is determined by the functional role they play in the overall linguistic system, we have found a way to specify the meanings of thoughts that does not presuppose any antecedent intentions on the part of the language user. That is, the words in which we think-out-loud are not produced as a

by-product of our thoughts, they *are* our thoughts; and their meaning comes not from the intentions of a speaker trying to communicate, but simply from the position they occupy in the functional system that forms the language with which we think.[55]

3.21 Reasons to Endorse a Sellarsian Semantics

Sellars' view is, I think, interesting in its own right, but the question still arises: Given that Sellars' view will work (and I admit, of course, that I have not yet secured the claim that it *will* work—much of that is still to come in Chapter Four), why should we *prefer* this view to Grice's (or to any other, for that matter)? After all, just because a theory of meaning can be made to go through doesn't mean that it's the *right* theory. So, having some idea, at least, of what the theory says, what reasons do we have to endorse a Sellarsian semantics? I have two responses to this question.

My first response is quite simple, though many may find it less than compelling. Why do *I* prefer a Sellarsian semantics? Simple: I need a theory of meaning that will work with the LTT. That means that I need a theory that can explain the meaningfulness of linguistic tokens without appeal to the meaningfulness of antecedent thoughts. Grice's theory, as we have seen, does not meet this requirement. And I think that the majority of theories are like Grice's in this respect: they presuppose that thoughts come first and words second (even if they don't take a communication-intention approach to meaning). That is, most theories of meaning, as far as I am aware, reify 'meanings' as *language-independent* abstract entities of some sort which the mind must first grasp before one can learn any language at all—and no view that approaches meaning in this way can serve the needs of the Linguistic Theory of Thought.

---

[55] Which language we think in is still an open question at this point, of course.

Sellars' view (or, at the very least, a slight modification of Sellars' view such as I develop below in Section 3.3), on the other hand, *can* meet the needs of the LTT. So I can say this for sure in favor of Sellars' theory of meaning: *if* one wishes to develop a viable linguistic conception of thought, *then* one is going to need Sellars' view, or something very much like it. That's reason enough to endorse Sellars' theory for anyone, like me, *already* interested in developing a linguistic conception of thought.

But what of all those others who are not yet sure that a linguistic conception of thought is even a viable position? What of those who might even take the consequences of such an approach to thinking (such as its need for a Sellarsian semantics) to be a *reductio* of the approach itself? What can we say against such critics as these? Why should *they* prefer Sellars' theory to Grice's, or to any other's?

My second response, then, is more complicated, and aims at providing reasons for a Sellarsian semantics to those not already sold on some linguistic conception of thought or other. The gist of my argument here (and I am taking my cue largely from Sellars) is a nominalistic approach to ontology generally, and meaning in particular. And the upshot of that position is that 'meanings' are not abstract entities, mental states, or whatever, to which 'meaners,' whether people or sentences, are somehow related.

But what are some of the reasons that people would think 'meanings' *are* things of some sort to which meaners are somehow related? As far as I know, most theories of meaning treat 'meanings' in just this way. There must, then (one would think) be strong reasons for such a view. Let's call all such views (abstracting from their differences with each other, and speaking only of the very generic position against which I am concerned to be arguing) 'realistic' (as opposed to 'nominalistic'). They are, as I mentioned above,

the views that reify 'meanings' as things; meanings then become objects which the mind must grasp if one is to use language at all. Again, 'realism' about meanings is probably the most prevalent view today, and has certainly been the most prevalent throughout history.

There are a great many issues of ontology and metaphysics that get tangled up together when one begins to talk about whether meanings are, or are not, actual entities. For example, if one supposes that meanings, as entities, must, given that they are not concrete entities of any kind, be abstract entities, then we face the general issue of whether or not there are *any* abstract entities at all. Or, for another example, when we talk of meanings we can often face questions of predication generally, and more specifically of the reference of predicate terms. We haven't the space here to go into all of the related ontological and metaphysical issues , but I do want to discuss generally the temptation to treat meanings as abstract entities, for it is this temptation that Sellars (and I agree with him) argues is misguided (and, though many of Sellars' arguments for this do delve into the more general issues surrounding abstract entities, I will attempt to steer away from those arguments not directly applicable to issues of theories of meaning as well as I can without losing the force of Sellars' more specific argument).

Now, then, let us return to the question of the motivation behind realism with respect to meanings. I can think of a number of reasons that one might be tempted to suppose that there are things in the world called 'meanings.' In general these reasons seem to arise from considerations about how language is actually used, and they range from simple observations about the grammar of 'means' statements, to more probing questions about synonymy and translation. Many philosophers, I think, have supposed

that the only way to explain such phenomena is by supposing that there really are meanings, as some kinds of entities. So what I hope to show here is simply that this is not the case—i.e., that we can explain the phenomena in question without invoking a realist approach to meaning. I also happen to think that the nominalist approach is favorable because of its simplified ontology, but I do not wish nor intend to debate this claim.

To begin, then, let's start with a conjunction of two obvious facts (obvious to everyone, not just to philosophers upon reflection): (1) the people of this world speak a multitude of different languages, and yet (2) we are still able to communicate with one another. We call this phenomenon 'translation.' I speak English. Another man may speak French. Nevertheless, it is possible to translate between the two languages. This is, as I've said, boringly obvious.[56] But then we might ask ourselves, given that translation between languages is possible, just what is it that one is doing when one translates, say from English to French? Or perhaps we could put the question another

[56] I should make an important note right up front here. On the traditional view of the relationship between thought and language, the *fact* of translation (as opposed to the actual execution, which may be quite complicated) is fairly boring. Speakers of different languages quite regularly communicate with each other because (says the traditionalist) we can, in fact, translate from one language into another. On a functional classification theory of meaning, however, things aren't nearly this simple. From what I have already said, and from what we will see in subsequent chapters (particularly Chapter Five), it should be clear that once we tie meaning to functional role (in a linguistic community), (1) what goes on when one 'translates' from one language to another is quite different from the way the traditionalists would see it, and (2) it is entirely within the realm of possibility that we could come across situations in which it were completely impossible to translate from one language to another. (I happen to think that (2) is highly unlikely to occur among individuals or communities of the same species, but the idea that we could one day encounter intelligent beings from, say, another planet with whom communication was difficult, if not impossible, in virtue of an inability to translate from their language to ours is one that I have no trouble imagining. Furthermore, it is, at least in principle, possible on my view that translation between different human languages could at times break down, becoming only partial at best.) What I am trying to do here is actually to show how the shift to a functional role semantics of the kind that I endorse can still explain phenomena, like translation between languages, that would seem to be quite clearly part of the landscape. I would like to thank Andrea Woody for calling my attention to the need for this important point of clarification.

way: How is it possible that something said in one language could be 'translated' into another? What goes on when we do that?

The answer to these questions is, on the traditionalist's view, as obvious as the fact of translation itself: translation is possible because two different expressions (one in English, perhaps, and the other in French) can *mean the same thing*.[57] This answer is often thought to be synonymous with saying that the two expressions *have the same meaning*.[58] This second way of putting the answer, though, lends itself quite easily to the generation of a further question: What, exactly, are these 'meanings' that expressions can be said to share? That, of course, is just the topic of this entire Chapter. But the traditional response has gone something like this: Given that expressions in *different* languages can one and all 'have the same meaning,' 'meanings' must be *language-independent* phenomena. Often it is said that sentences in different languages simply *express* these language-independent meanings (the logician might call them 'propositions')—indeed, it is usually supposed that that is how sentences in a language become meaningful in the first place.

Once we notice all this we should also note that *translation* isn't really a necessary part of the equation—for just as two sentences in *different* languages can be said to 'have the same meaning,' so too can two sentences in the *same* language.[59] That is, two linguistic types in, say, English can be synonymous. This, again, raises the

---

[57] Interestingly enough, I don't disagree with this claim. I think that translation is possible because two expressions can mean the same thing. But, of course, I will cash out just what 'meaning the same thing' comes to quite differently than most of those who would give this answer. I'll say more on this below.

[58] This way of putting the answer is more problematic, for, as we'll see, I think that it lends itself (as a matter of grammar) to the generation of a confusion: the confusion of 'being meaningful' with 'possessing some*thing* that is a 'meaning'.'

[59] I am reminded here of a similar turn of the argument in Quine's "Ontological Relativity." See Quine, 1969, p.48.

question of what, exactly, these 'meanings' that are shared by synonymous expressions might actually be. Synonymy might not suggest the need for language-independent entities, in the same sense that translation seems to, but if we phrase our definition of 'synonymy' as two expressions 'having the same meaning' we still end up with 'meanings' as things to which words and sentences are somehow related, for 'having' is a relation between one *thing* and another. Hence, it might easily be supposed, 'having a meaning' is a relation between one thing (a sentence in a particular language, say) and another thing (a 'meaning'). Hence 'meanings' must be entities of some sort.[60]

Ultimately, I believe that it is considerations along these lines that have often given rise to the 'realistic' view that *meanings* are *things*. For how else are we to explain the inter-translatability of sentences of different languages or the synonymy relation between different sentences of the same language? Thus, goes this line of thinking, to explain sameness of meaning we are forced to admit 'meanings' into our ontology.

To respond to this line of thinking from the standpoint of Sellarsian semantics all we need to do is show that the phenomena in question are still explicable without invoking meanings as entities of some kind to which linguistic expression are then related. But, really, we've already shown this. Take, for example, sentences (3) and (3a) from the previous section:

(3) '*Oder*' (in German) means *or*

(3a) '*Oder*'s (in German) are ·or·s.

---

[60] I suppose I should note here as well that the idea that meaning is a relation between linguistic expressions on the one hand and abstract, mental, or whatever, entities on the other hand sometimes occurs to people from the simple grammatical similarity between, for example, "'inert' means *unable to move*" and "Jane loves John." As the latter expresses a relation (between Jane and John) so too, it might be supposed, the former expresses a relation as well (between the linguistic expression 'inert' and its meaning).

Sentence (3) is an ordinary means statement in a translational context. Sentence (3a) gives the Sellarsian 'functional classification' way of understanding (3). To spell things out: the realist about meanings is impressed enough by (3)—or, rather, the phenomenon of translation of which (3) is an example—to suppose that there must be *something* to which both the German word 'oder' and the English word 'or' are related such that they can both mean the same thing. Sellars, though, as we've seen, has given us a different way to explain this fact (the fact that 'oder' and 'or' mean the same thing): *viz.*, that both 'oder' and 'or' are ·or·s. That is, that 'oder' functions in the language of which it is a part (German) much as 'or' functions in the language of which *it* is a part (English). So (just to belabor the point), how is it that 'oder' (in German) and 'or' (in English) can mean the same thing? Simply by the fact that they play roughly the same functional roles in their respective languages. So far from supposing that there must be an entity of some sort called a 'meaning' to which both 'oder' and 'or' are related, meaning is not even a relation on this view. Nevertheless, we have no trouble explaining the phenomenon of translation.[61]

To say that Sellarsian semantics can explain relevant linguistic phenomena as well as a realist theory of meaning is not, of course, to give any reason for *preferring* the former. My preference for a nominalistic theory of meaning arises, at least in part, from an application of Ockham's razor. If we can explain the phenomena of language without invoking any additional entities, then I feel that we should. Yet, we've only looked briefly at the issues of translation and synonymy (seen at root to be the same issue,

---

[61] As we have seen, roughly the same solution will also explain the phenomenon of synonymy. As for the grammar of means statements: if a superficial similarity in grammatical structure between statements expressing a relation (e.g., "Jane loves John") and means statements is all we have to hang our hats on, I don't think we have any reason at all to inflate our ontology to include 'meanings' as entities.

really), so the question naturally arises, Can Sellarsian semantics explain all linguistic phenomena as well as a realist theory? Somewhat paradoxically, there is one issue that might seem to suggest that nominalism *won't* work as well as realism: the issue of how words come to be meaningful *in the first place.*[62]

The worry can be approached like this. On the functional classification theory of meaning, which we are in the midst of developing in this chapter, words are meaningful in virtue of the functional role that they play in the language of which they are a part. Yet how do these words come by that function to begin with? If we imagine a language already up and running, there is no worry. But if we try to imagine how language use began we might start to feel a tug of doubt. This is an issue not just of language learning, but of the genesis of language. Imagine some proto-humans that have not yet developed a language.[63] How could such a development take place? If the sounds that they produce are only meaningful linguistic tokens in virtue of playing a role in a language, none of these proto-humans' noises will count as meaningful. But if their noises aren't meaningful, how can they ever constitute a language? Here the realist seems to have the upper hand. If we ask him, How does language originally become meaningful?, he can respond that linguistic tokens are meaningful because they express thoughts (or propositions, or some such). If we then ask how those thoughts originally become meaningful, though, we are told that the question is out of bounds—that it makes no sense. It is the *job* of thoughts (propositions, whatever) to be meaningful. They don't

[62] This might appear paradoxical because this 'further' issue of how words become meaningful might seem to be just the very thing that's been at issue from the start. This isn't quite right, though; we'll see what I mean in the next paragraph.

[63] I will be dealing, to some degree, with the issue of language learning (within the context of a pre-established language) in Chapters Four and Five. We will revisit the issue of language genesis very briefly there, but as a matter of fact I do not propose to attempt a full solution to this problem within the confines of this project.

'get' their meaning from somewhere, they simply *are* meaningful. The proponent of a Sellarsian semantics, obviously, has no such recourse. So perhaps we have a reason to prefer a realist theory of meaning after all.

Obviously, though, I do not find the matter so easily resolved. To begin with, I am unmoved by the suggestion that thoughts simply are meaningful. This suggestion solves the issue of language genesis merely by pushing the important question (of where meaning comes from) back onto thoughts; when we pursue the question there, however, we are told that we can't ask it. So the realist uses a delaying tactic, followed by a simple refusal to engage the issue at all. Still, the question of language genesis is a puzzling one, and we can perhaps forgive the realist for seeking the easy way out. Nor am I suggesting, in expressing dissatisfaction with the realist's solution, that I have a fully worked out and viable nominalist alternative on hand. As may already be apparent, however, such a solution requires more than just a discussion of linguistic meaning—it requires also a theory of mind. In fact, in a sense, the solution that I envision finds a grain of truth in the realist's tactic. I do think that the issue of language genesis requires us to look to the meaningfulness of thoughts. As we shall see in subsequent chapters, though, when I say this I mean something quite different from what the realist means. Since we have not yet developed the foundation for the theory of thoughts that I have in mind, however, I will end the discussion of this section here and turn to the final part of our discussion of theories of meaning: an adaptation of Sellarsian semantics for use with the Linguistic Theory of Thought.

3.3 Meaning in the LTT: Adapting Sellars' View

Now that we have seen Sellars' theory of meaning, and looked briefly at some reasons for preferring it, it's time to see how we can use that theory within the context of the Linguistic Theory of Thought. In this section what I intend to do is outline how treating meaning as a matter of functional classification can be used to explain the meaningfulness not only of linguistic tokens but of the thoughts that are constituted by those tokens according to the LTT, and how doing so will allow me to sidestep some potentially hazardous issues.

By now the advantage of a Sellarsian semantics for the LTT should be almost obvious: since Sellars does not analyze or define the meaning of linguistic tokens in terms of antecedent thoughts it is at least possible that we could use this semantic theory to explain the meaningfulness of thoughts themselves. That is, assuming that there are good reasons to believe that thoughts are constituted by linguistic symbols (and the main arguments for this position were begun in Chapters One and Two, and are continued below in Chapter Four), a Sellarsian semantics seems almost necessary, for it is one of the few theories of meaning I know of that explains the meaning of linguistic tokens without invoking the prior meaningfulness of intentional mental states. And if thoughts are to be constituted by linguistic tokens, then the meaningfulness of linguistic tokens has to be explicable without appeal to thoughts, otherwise the account would be viciously circular. So let me say a bit about how I think this all fits together.

To begin, let's revisit an example and an objection from above. Take the claim that

(4a) 'Triangle's (in English) are ·three sided polygon·s.

On our analysis above, what (4a) tells us is that the English expression 'triangle' serves the same functional role in the English language as does 'three sided polygon.' Now, by itself (we noted) this seems to fail to tell us anything about meaning at all, for knowing that one bit of language works in the same way as another piece of language seems to be empty unless we know what way that is—i.e., what one or the other of the expressions actually *means*. But we also saw that we could blunt this objection by an appeal to the difference between the mere parroting of words and the use of words in accordance with the rules of correct usage; and this distinction led us to issues of language learning that had to be postponed until Chapter Four.

Not everything about this issue needs to be postponed, however. It's time to make fully explicit just what it means to talk about the 'function' of a linguistic token. As Sellars himself points out, using a word in the functionally correct manner means that "the utterances one makes cohere with each other *and with the context in which they occur*" (p.429, my italics). For example, using the term 'triangle' appropriately involves not only using it with the correct intra-linguistic connections between, e.g., 'triangle' and 'three sided polygon', but also using it, say, as a verbal response to the sensory perception of actual triangles. That is, the function of a word is determined by its functional connections to the world as well as its functional connections to other words and sentences.

All of this was at least implicit in our discussion of Sellars' view in Section 3.2 above. I think, however, that it's important to emphasize this point. The most important reason for calling attention to the fact that the 'function' of a linguistic token will often involve mention of world-word connections is a familiar empiricist one: Without this

connection our words become mere symbols spinning frictionless in the void. And it's very hard to see how meaning could be grounded in something as unstable as a (perhaps internally coherent) system of symbols and various transformations to and from some sets of these symbols to others that nevertheless made no contact with a reality that was (in some sense) language-independent. In the philosophy of mind, this issue is often framed as a distinction between the 'narrow' and 'wide' contents of a given mental state, where narrow content is merely the internal relations of one intentional state to any others, while wide content encompasses all those relations but adds to them relations between the mental state and states of the world.

With respect to a Sellarsian semantics, then, the point is this: An analysis of meaning as a matter of the functional classification of linguistic tokens should not be mistaken (as it might be if one looks only at dot-quoted examples of translations like (3a) and (5b) in 3.2 above) as the mere exchanging of one set of symbols for another set. In fact, it's important to remember that Sellars' analysis in those cases is of the meaning of 'means'; 'means' is not, according to Sellars, a relational word. Rather, it's a special form of the copula, telling us that one set of symbols has the same function as another set of symbols. But the meaning of any particular linguistic token is then only given fully when we look at the function in question.

The danger here is misunderstanding Sellars in such a way that his theory of meaning looks incomplete—misunderstanding him, in fact, in such a way that it might seem that his theory is open to the worry above about the 'frictionless spinning' of symbols that make no contact with the world. For, if all it means to say that 'triangle' means ·three sided polygon· is that 'triangle' and 'three sided polygon' have the same

function—without specifying which function that is—then we haven't really learned what we want to know.

I think, however, that it's fairly clear that Sellars intends for us to include a specification of the function of an expression in fully explicating its meaning. What he's doing with the dot-quoting convention, and sentences like (3a), (4a), and (5b) is giving us an analysis of the *grammar* of 'means' statements. He is *not* giving the meaning of any actual expressions.[64] Again, in order to do *that* we'd have to actually specify the function of a particular linguistic token. Of course, in certain contexts (particularly in contexts of translation) there is a sense in which we *are* giving the meaning of an expression when we use a 'means' statement. Take, for example, (3) from above:

(3) '*Oder*' (in German) means *or*.

On Sellars' analysis, this tells us that 'oder' functions, in the German language, as does 'or' in the English language. Now, assuming that we *already know* the function of 'or' in the English language, then (3) does, in fact, tell us the meaning of 'oder' in German. That is, since (3) tells us that 'oder' means, i.e., functions as does, 'or'—and the function of 'or' is plain to us (that is, already specified, in practice at least)—then (3) does specify the function of 'oder', at least in a sense.

All of this is intended by, and included in, I think, Sellars' analysis of meaning as functional classification. The task now is to apply this to the idea that thoughts are constituted by linguistic tokens (of some form or other—more on this below). Take any thought—say, the thought that today is Thursday. That thought has a certain semantic

---

[64] In truth, this isn't completely correct. For, as a matter of fact, Sellars *is* giving the meaning of one expression: 'means'. For he has just told us that the function of 'means' is to serve as a form of the copula indicating that one expression serves the same function as another. And that *is* to specify the function of 'means'—and, hence, to give its meaning, on Sellars' view.

content which we would ordinarily specify by using the English words 'Today is Thursday.' On the traditional view of thought, however, the *fact* that we would specify the semantic content of the thought in question with those particular English words reflects merely the fact that those words *express* the thought. Nor, it is supposed by the traditionalist, are those the only words that could be used to express that thought. There are other combinations of English words that would do the job, perhaps, and of course there are all the various other languages of the world to consider: the *very same* thought can be expressed using words in French, Italian, Russian, Japanese, etc.

Now, from one angle the question is how a particular linguistic token (e.g., 'Today is Thursday') comes to mean what it does, and we have seen the traditionalist's answer and Sellars' alternative. But from another angle the question is this: How does that *thought* come to be meaningful in the first place? This is the question that Sellars explicitly takes himself to be answering (in "Meaning as Functional Classification"),[65] though I hazard to say that it's not entirely obvious just *how* Sellars is addressing this question since much of what he says is aimed at the explication of the grammar of 'means' statements, especially in translational contexts. But the answer is there just the same: the *thought* that today is Thursday means what it does because *it's a linguistic token* that functions as does the English token 'Today is Thursday.'

Here I find myself faced with something that strikes me as somewhat of an oddity in Sellars' view. From the general corpus of Sellars' work it is clear that he believes thought to take place in (be constituted by) a universal 'language of thought' commonly

[65] Again, the explicitness comes in the form of an endorsement of Gilbert Harman's 'three levels of meaning' analysis of the problem of meaning coupled with the claim that he [Sellars] is engaged in "the construction of a 'level 1 theory of meaning' in Harman's sense of this phrase" (Sellars, 1974, p.18).

called Mentalese. Combined with his above analysis of 'means', and the explicit claim that this analysis is to be applied to the meaningfulness of *thoughts*, it clearly follows that Sellars intends the meaningfulness of thoughts to be a matter of the functional classification of Mentalese tokens. So far there seems to be no problem. But now what if we attempt to apply the Sellarsian dot-quote analysis of meaning to these Mentalese tokens? At the very least we run into this problem: there is not (on Sellars' view, or anyone else's that I know of) any way of actually *inscribing* or otherwise publicly tokening Mentalese symbols. Thus, we cannot construct sentences like (3a) when it comes to Mentalese, for we have nothing to put within the single quotes.

We might not think this such a problem, however, if we remember that 'means' statements only specify the function of a linguistic token by using a sophisticated sort of pointing device. We can, of course (at least in principle), specify the function of a linguistic token without recourse to the dot-quoting convention by simply laying out the actual function of the token. But then again, we still have no direct way of specifying *which* token we're concerned with when it comes to Mentalese. We seem to find ourselves in the position of having to go at it indirectly by saying something like,

(6) The Mentalese token that would ordinarily be expressed by the English phrase 'Today is Thursday' is a ·Today is Thursday·.

That seems rather roundabout to me, and I can't help but think that the inclusion of Mentalese here just adds a level of obscurity that might otherwise be avoided. How could one avoid this? By taking the English sentence *itself* to simply *be* the thought; i.e., by supposing that in thinking that today is Thursday, one is simply tokening (in the mental way—and I'll have to say more about what way that is) *that very phrase*: 'Today

is Thursday.' I will come back to this suggestion in a moment. For, while ultimately this will be the path I wish to take with the LTT, it is not the path that Sellars takes, and I feel that I should say at least a little more about why I find it a bit odd that Sellars embraces Mentalese as he does.

We need to recall here something that is at least implicit in the traditionalist's view of the relationship between thought and language. While words and sentences are supposed to get their meaning from the thoughts that they express, thoughts are generally treated as being by their very nature meaningful. Indeed, the traditional view might just be that thoughts simply *are* meanings. In that case, the question of *how* thoughts become meaningful fails to arise. On Sellars' analysis of meaning, however (again, especially considering his explicit claim to be developing a theory of the meaning of *thoughts*), this is the very question at issue. Thus, we are at the very least required to give *some* explanation of the meaningfulness of thoughts. This leads back to the first problem mentioned above that taking the language of thought to be Mentalese generates: How do we refer to the tokens of this universal, but publicly unobservable, language in the first place? Suppose, however, that we could get around this problem—or, perhaps, that we could learn to live with the rather awkward sentence (6). I think that there is a more serious problem to consider.

Unfortunately, the problem that I think here arises is a consequence in large part of Sellars' take on how a non-thinking thing becomes a thinking thing (e.g., the infant becomes a thinker), and that involves at least in part Sellars' approach to language learning. I have delayed a full exploration of these issues until Chapter Four. I cannot, therefore, fully explicate where the problem lies. I can say, though, what the problem *is*.

Simply, the problem is this: According to Sellars, while the language of thought is Mentalese, one is not simply born fluent in it. One learns to think, on Sellars' view, only after much behavioral conditioning involving the parroting of some natural language or other. Yet if this is the case, Mentalese seems even more unnecessary to me than it did when looking at the first objection above. For if one does not learn to think without learning first to *talk*, then what is gained by supposing that in addition to learning one's natural language, one simultaneously becomes proficient in the universal (though at least in some sense private) language of Mentalese? Again, why not simply suppose that the 'thinkings-out-loud' of Sellars' picture are just overt (speech) tokens of those very same symbols tokened 'covertly' as it were? Again, as far as I can see, Mentalese is just unnecessary baggage. In essence I'm suggesting that Sellars has failed to shave his ontology as closely with Ockham's razor as he could have.

The alternative, as I conceive of it, is of course what I am calling the Linguistic Theory of Thought. And according to the LTT, one's language of thought is simply the language that one speaks.[66] The question of how to specify the meaning of a particular linguistic token (in English, say) and the question of how to specify the meaning of a particular thought token simply collapse into one and the same question, to be answered, of course, by the functional classification analysis of 'means' statements offered by Sellars. I do, then, in a sense simply see my own position as a logical extension of Sellars' view. In order, however, to show this, I must finish my presentation of Sellars' position, and more clearly develop the reasons for which I find myself needing to depart from Sellars. These are the tasks of Chapters Four and Five.

---

[66] I'm also perfectly happy to suppose that multilingual people are capable of thinking in more than one language.

4. Thinking in Language

The stage is now set for us to look at the core arguments for the Linguistic Theory of Thought, and for me to begin to make good on the promissory notes that I issued in earlier chapters. I have, therefore, broadly speaking, two main aims for this chapter. First of all, I will argue that thought is essentially norm-governed. I claimed in Chapter Two that recognition of the normativity of thought is what separates the LTT from nearly every other representational theory of mind; I intend here to show not only how that is the case, but also why the norm-governed nature of thoughts is so essential to making them thoughts in the first place. Secondly, I want to lay out the basic Sellarsian framework of 'Psychological Nominalism'. Sellars' arguments will both buttress my argument for the normativity of thought and serve as a general foundation upon which I will build the LTT. With respect to Chapter Three, here I will discuss in detail Sellars' conception of the phenomenon of language learning, demonstrating how it shores up the psychosemantic theory developed in that chapter.

Within these two broad aims for this chapter we will have the opportunity to look at a number of additional issues that arise in discussions of the main arguments. For example, in laying out Sellars' Psychological Nominalism I will want to discuss, and interpret, Sellars' claim that language is 'first in the order of conceiving' though 'second in the order of being' when compared to thoughts (taken on analogy with theoretical entities). We will also look at the distinction between 'ought to be' and 'ought to do' rules of language, Sellars' arguments concerning 'rule-governed behavior' as opposed to merely 'tied behavior' (and 'tied symbol behavior'), and the difficult issues surrounding the questions of just when and how a language learner transitions into a language speaker

(and, hence, a thinker). Much of what I have to say about these, and other, topics will be interrelated, though I will try to make each issue as clear as possible.

When we have the argument for the claim that thought is essentially norm-governed, the Sellarsian foundation of Psychological Nominalism, and the psychosemantics from Chapter Three all firmly in place, we will at last be able to turn to developing the final superstructure of the LTT, adapting Sellars' view into a modern representationalist picture of thought (Chapter Five). It will remain then merely to reply to a few of the main objections that might be raised against this view (Chapter Six).

4.1 The Normativity of Thought

The question arose in Chapter Two: What separates the Linguistic Theory of Thought from its representationalist peers such that one would prefer the LTT to the views discussed in Chapter Two, or others like them? Given that I largely agree with much of what is said by people like Fodor, Dretske, Millikan, Cummins, Carruthers, etc., why do I find myself ultimately wishing to depart fairly dramatically from their various theories? In addition to the criticisms offered in Chapter Two I now wish to add perhaps the most important criticism of all: I believe that one very important requirement for considering anything to be a thought is routinely overlooked by nearly every modern proponent of representationalism. The overlooked requirement that I have in mind, of course, is the one mentioned in Chapter Two: that thought is essentially normative in nature.[67] I should mention, though, that while I take most representationalist views to be lacking with regard to this requirement, I have never wanted to suggest that they are

[67] There are other philosophers who endorse the general claim that thought is essentially normative, of course. Most obviously Sellars and his followers (among whom I clearly count myself).

completely wrong-headed—in fact, far from it. Here I must introduce a distinction that will prove to be of supreme importance in what follows. Many of the representationalist accounts in Chapter Two seem to me promising as accounts of the phenomenon of non-conceptual representation (what we might call 'natural representation'). Where they fail, however, is as accounts of *conceptual* representation. This distinction may look familiar, as though it is just, say, Dretske's distinction between natural signs and functional meaning, or Grice's distinction between natural and non-natural meaning. This appearance of familiarity is, however, I think misleading. Let me explain.

First of all, the difference between what I'm calling natural representation and conceptual representation is not simply the distinction between mental states and non-mental states. What I am interested in are those *conceptual* mental states that we call 'thoughts'. I thus classify, e.g., possibly, an ant's olfactory perception of a dead comrade as a non-conceptual mental state[68]; but the non-conceptual will also include representations that are not *mental* states at all, e.g., the states of a modern computer, or (to borrow from Dretske) the direction of pull of a magnetotactic bacterium's magnetosome. I then use the terms 'conceptual representation' and 'thought' interchangeably. I am thus drawing my line of distinction between one *subset* of mental

---

[68] I consider the ant's perception to be a mental state because it is a *perception*—a state which, it seems to me, requires that the system have at least some of what we would ordinarily consider to be mental faculties, e.g., consciousness. Contrast this with the examples that follow: the states of a modern computer, or the pull of a magnetotactic bacterium's magnetosome: these states do not seem to require that the systems in question have any sort of mental faculties. Though I will have occasion to discuss this in more detail later, it might be helpful to be explicit here about the fact that I obviously consider the LTT to be only one part of a full-fledged theory of mind—which is to say that *thoughts* are not the only kinds of mental states. Sensory states (e.g., pain), and affective states (e.g., anger) are clearly mental states (systems without minds do not have such states), but they aren't thoughts. (As an aside: desire is almost always used as an example of a propositional attitude, though desire is actually an affective state; thus, while affective states generally aren't thoughts, there is a sense in which many affective states in thinking beings either are thoughts or at least are combined with thoughts, insofar as they are conceptual representational states of the thinking being.)

representations and all other representations (mental or otherwise). It is only the former that I claim are essentially normative in nature. My category of 'natural representation' therefore obviously includes more than Dretske's 'natural signs' or Grice's 'natural meaning.' Similarly, when I speak of 'conceptual representations' I am clearly talking about a more narrow set of representations than would be included in Dretske's list of representations with 'functional meaning', or in Grice's list of representations with 'non-natural meaning.' We must keep this new distinction clear in our minds, for it will, as I said, underlie all that follows.

On a further note, one should not take my use of the term 'conceptual' as invoking the classic division between concepts and sensations, for as I said above I will be using 'conceptual representation' as interchangeable with 'thought', and there will be a class of representations that are non-conceptual, but are nevertheless representations and not merely sensations. For example, it will ultimately be my contention that while a dog may have a sensation of a ball, and that sensation may trigger a representation that serves some of the same functions as our *concept* of a ball, the dog's representation is *not* conceptual, i.e., the dog does not have the *concept* 'ball'. Unlike most representationalists, who want to make conceptual representation of a piece with all other forms of representation, I believe that conceptual representation is importantly unique.

4.11 The Unity and Continuum Claims

I've mentioned before that most, if not all, modern representationalists have a general commitment to naturalism. In addition to that, many of these representationalists (those discussed in Chapter Two can be considered a representative sample) seem to me

to endorse two related claims: to the unity of the phenomenon of representation (we'll call this the unity claim), and to a continuum from simple representational systems to complex ones (we'll call this the continuum claim). Let me explain.

All sorts of things are representations: the pull of a magnetotactic bacterium's magnetosome represents the direction of magnetic north; the 'dance' of a bee represents the location of pollen; the physical processes of a pocket calculator represent functions like adding and multiplying numbers. Now, as we've seen, while many representationalists will talk about the difference between, say, 'natural signs' and the symbol tokens that constitute cognitive representations, they are all nevertheless committed to the idea that cognitive representation (and importantly, for our purposes, our new category of conceptual representation) differs only in degree and not in kind from what I'm calling natural representation. That is, they believe that the phenomenon of representation is a singular one. (This is the unity claim). In light of this, modern representationalists also believe that there is no sharp divide between those representational systems that have, and those that do not have, thoughts. There is, these representationalists believe, a continuum from the simple representers (e.g., the magnetotactic bacteria) to the complex representers (e.g., thinking human beings). (This is the continuum claim.)

While I agree that it would be preferable to have only one kind of phenomenon of representation, and that we should preserve as much of the continuity between simple and complex representational systems as possible, I also believe that the unity and continuum claims are ultimately mistaken. Before I say *why* I think the unity and continuum claims are mistaken, though, let me explain their bases a little more.

The unity claim, as should be apparent, doesn't follow from a basic naturalism, though it is always combined with naturalism in such a way as to take a particular shape. The basic argument for the unity claim goes like this: treating the phenomenon of representation as a single, unified phenomenon is a good idea because if we can do with only one type of representation, then there's really no sense in having two.[69] Given that parsimony with respect to our ontology is a good thing, this seems like a reasonable position to take. When we combine the unity claim with the general commitment to naturalism we get the view that all representations are of the 'natural' type.

Now, as we noted earlier, the treatment of the phenomenon of representing as a singular, unified phenomenon leads to the idea that there must exist a continuum from those organisms that have only simple representations to those organisms that have complex representations. Materialist philosophers of mind in general tend to hold the existence of such a continuum as an almost necessary condition for success in the philosophy of mind. The motivations behind wanting such a continuum, however, are more complex than a simple desire for a parsimonious ontology—and I do not think that materialists have always been boldly explicit about some of these motivations.[70] One belief that I think motivates materialists in this direction, but which is seldom, if ever, explicitly acknowledged is the belief that the rejection of Cartesian dualism requires a rejection of the idea that there is a strong divide between the thinkers and the non-thinkers. Another possible motivation behind the continuum claim, which is actually

---

[69] I don't know that anyone actually gives this argument (or even states their commitment to the unity claim) explicitly. That such an argument (and such a commitment) is *implicit* in modern representationalist views, however, is beyond reasonable doubt.

[70] Unlike with the related unity claim, however, most representationalists are more-or-less explicit in their acceptance of the continuum claim (though none have, to my knowledge, called it that).

more likely to be acknowledged by materialists, is the belief that if humans have evolved from other forms of life it is implausible (at best) to suppose that at some point humans gained a *unique* ability to represent in a special way called 'thinking'. This motivation is, of course, an off-shoot of a general commitment to naturalism.[71] Motivations aside, though, the basic sentiment behind the continuum claim is that all representing systems fall somewhere within a range of representational capacities, so that complex representing (e.g., thinking) is just a more advanced version of what goes on in simple representational systems.

In the next section I argue that both the unity claim and the continuum claim are mistaken. Before I get into the details of my argument, though, let me make a couple of preliminary comments right up front. First of all, as we will see below, I do not believe that a unified account of the phenomenon of representation will be able to explain everything about thoughts that needs explaining. The unity claim will therefore be the first to fall apart, and without the unity claim, the continuum claim will not immediately come into the picture. Secondly, with respect to the anti-Cartesian motivation for the continuum claim, while I understand and share the motivation to avoid Cartesian dualism, I think it drives materialists too far. There is no reason, that I can see, to suppose that if there is a sharp divide between the thinkers and the non-thinkers we are thereby forced into some form of dualism (let alone *Cartesian* dualism); the divide may be (and on my

[71] Even absent a commitment to evolution (e.g., for philosophers writing before the development of evolutionary theory), the continuum claim can look appealing, especially once one rejects the notion that humans (and perhaps other animals) have immaterial souls—for once that notion is discarded, it might seem intuitively quite plausible to suppose more continuity than not between us and (at least many) other animals.

view is) a perfectly natural one, entirely compatible with a materialist metaphysics.[72] And finally, with regard to the worry about evolution, again while I agree with my representationalist peers that human thought must have an evolutionary explanation, I do not believe that this prevents there being a sharp divide between those beings that think, and those that do not. The evolutionary explanation of this divide may be more complicated than explanations of other skills possessed by some creatures and lacked by others, but I do not see any reason to suppose that it will be impossible to develop such an explanation. I think that most people these days would agree that thought has something to do with the brain, and the brain is an organ that we are very far from understanding fully. In time, as our understanding of the brain grows, in concert with the right theoretical framework for dealing with minds, I believe that an evolutionary explanation of the origins of thought will be found. That such an explanation is not yet available is not by itself reason to oversimplify our concept of thinking as those who endorse the unity and continuum claims seem to me to be doing. These preliminary comments out of the way, I turn now to my detailed arguments for rejecting the unity and continuum claims, to key to which lies in my contention that thought is essentially norm-governed.

---

[72] There is a related motivation, for ethically-minded philosophers: some have supposed that if animals don't *think* then they cannot have a moral status that would prevent us from mistreating them. Though I have no interest in engaging in an ethical debate here, this claim (like the one concerning a looming dualism) seems to me misguided. Nothing in the claim that there is a sharp divide between the thinkers and the non-thinkers suggests to me that we cannot have moral duties toward animals. This is particularly true in my case since I wish to separate *thinking* (*conceptual* representing) from other types of mental states, e.g., experiencing pain. But that is another issue.

4.12 Error and Normativity

When I talk about the normativity of thought I mean primarily to be speaking of its rule-governed nature. Nothing will count as a thought unless it occurs, to borrow from Sellars, within the 'logical space of reasons' ("Empiricism and the Philosophy of Mind," §36, 1997, p.76). To attribute to someone a thought, the tokening of a conceptual representation, is to treat him as a rational being, and the tokening as governed by the norms of rationality.[73] What this comes to, in my view, is the idea that the *way* a representation is tokened plays a crucial (perhaps *the* crucial) role in determining whether that representation is a thought or merely a natural representation. It is, of course, my contention that most modern representationalists are unable to account for thoughts thus conceived. To see this we revisit the notion of error (from Chapter Two).

Consider again what happens when a representational system gets things *wrong*, i.e., when it commits an error, or when it *mis*represents. As we saw in Chapter Two, the problem of misrepresentation is a famous one in the literature on RTM, and as we also saw, I believe it is safe to say that this problem has been the primary driving force behind the development of many of the various differing representationalist views, since they are all to some extent attempts to solve, or at least avoid, the problem of misrepresentation.[74] Recall briefly the problem: A system can misrepresent only when a given representation of that system can be said to have a determinate content. So, for example, if we suppose that a certain representation in the brain of a frog has the content 'fly', then if that representation occurs in response to a BB, the frog's brain has misrepresented the world

[73] For a brief argument to this effect, see, e.g., Jaegwon Kim, "What is 'Naturalized Epistemology'?", 1988, pp.392-394.

[74] See again the discussions of Causal Covariance, Teleology, and Functional Role Semantics in §§ 2.21-2.23 above.

(for a BB is not a fly). But suppose that we have no way of fixing the content of a given representation. What, then, would allow us to say that a system had misrepresented? If I cannot say with certainty that *that* representation in the frog's brain means 'fly', then when that representation is caused by a BB, why should I suppose that anything has gone wrong? If I cannot fix the content of the representation beforehand, then I cannot identify cases of error, for it is always open to me to re-describe the content of the representation so as to make it a correct representing after all.

What, though, does this have to do with the supposed normative nature of thought? The answer is that, for most representationalists, we can fix the content of a given representation even in systems that are not ordinarily considered cognitive—let alone possessed of *conceptual* representations. For example, we can come up with some story to tell that will allow us to say that a given representation in a computer represents the number four. If that representation occurs in the computer as a result of the computer adding two and three, then, it has made an error, i.e., misrepresented. However, and this is the crucial point, *when such a computer misrepresents, the error is one of pure malfunction, not one for which the computer is itself responsible.* But, as I will argue below, a defining feature of conceptual representations (perhaps *the* defining feature, when comparing them with non-conceptual representations) is that they are rule-governed in such a way that the representing system can be held responsible for its misrepresentations. When I add two and three and get four, I've made an error—but it is an error for which I can be criticized: I know how to add, so I shouldn't have produced the answer that I did. My representational mistakes are not *mere* malfunction, like some sort of mechanical failure—they are a violation, on my part, of the rules that govern the

use of the representations in question. So, when I say that conceptual representation is normative in nature, I mean that it is rule-governed in a way that involves agency, and a responsibility on the part of the representing system for obeying the rules (and this will involve, as Sellars has emphasized, a *recognition*, in some sense, *of* those rules on the part of the representing system itself[75]). To be a thought, I will argue below, a representation must be essentially rule-governed, because it is only when representations are so governed that they can play the roles that we intuitively take thoughts to play.

Assuming, for the moment, that I am right about thoughts requiring this notion of responsibility or agency (and I have not, yet, argued for this claim), how do other representationalists get themselves into the situation of not being able to account for this feature of thoughts? The answer has to do with the two claims (the unity and continuum claims) that we discussed in the previous section. These two claims force representationalists who endorse them into the position of having to treat *all* error, *all* misrepresentation in the same way. This should really be quite obvious: for if there is no difference (except in degree of complexity) between non-conceptual and conceptual representations that get things *right*, then there can be no difference (except in degree of complexity) between non-conceptual and conceptual representations that get things *wrong*. If representing is a unified, singular phenomenon, and there exists a continuum

---

[75] See his discussion of the 'second hurdle' in §35 of "Empiricism and the Philosophy of Mind" (1997, pp.73-75); though Sellars is here talking about recognition of one's authority in making a report, the point—and the 'hurdle'—are much the same. This issue of one 'recognizing' the rules that govern one's representations in order that one's representations count as *conceptual* is complicated to say the least. I will discuss it more fully below (in Section 4.2), though even there I will have to leave the issue less than completely settled, as we shall see. Please note, however, that I will be making liberal use of phrases such as 'recognizing the rules' that govern a representational system, 'awareness of one's authority' to token certain representations, and 'recognition of the appropriateness' of one's representational tokens; I find these phrases convenient, but I do not endorse any rationalist, Platonist, etc. implications that these phrases might seem to carry.

from those beings with only simple (natural, non-conceptual) representations to those beings with complex (conceptual) representations, then any account that a representationalist gives of misrepresentation in one case will have to be more-or-less the account she would give in any other case. I am not suggesting that this only causes problems (or is only apparent) when one looks at the issue of error, but the problem of misrepresentation is famous for bringing out the differences between simple and complex representations, and thus is a familiar problem with which to illustrate the failure of representationalist views that endorse the unity and continuum claims to account for the rule-governed nature of thoughts.

Again, to put it simply and boldly: error when dealing with a non-conceptual representation is to be accounted for as mere malfunction; error when dealing with conceptual representations requires agency and recognition of thoughts as taking place within the space of reasons. And here we find a deep divide between what I claim are in fact two *different* types of representational capacities: there are two phenomena, not one; there is a break in the continuum from simple to complex representational systems—a break that occurs when a representing system becomes part of a norm-governed enterprise.

Before I turn to the two examples that I will use to drive my argument for the uniqueness of conceptual representations, I should probably say a few words about terminology, specifically with regard to the various phrases I use having to do with representations. Sometimes I speak of representations, by which I mean the actual symbol tokens (and sometime types) themselves; other times I speak of representational systems, by which I mean anything in which such symbol tokens occur; still other times I

speak of 'representings' or 'tokenings of representations,' which are of course the actions[76] of representational systems in producing representations. I don't think that I've been unclear in the way that I've used these various phrases, or that the phrases themselves are particularly ambiguous. One might wonder, however, when we bring in the notions of error and responsibility whether it is most appropriate to talk about the representations, the representational systems, or the acts of representing.

This is a reasonable question, and it may seem as though the answer ought to be that, when we are speaking of the notions of agency and responsibility, the proper focus is the systems in which representations occur. There is an element of truth to this, of course (for it makes little sense to talk about the 'agency' of the symbol tokens themselves), yet I don't think that the matter is so simple. More precisely, I think that while the concepts of agency and responsibility are ones that can only be applied to representational systems (and representational systems of a certain type, given what I've set out here to argue, *viz.*, that some representing is conceptual and rule governed, while some is not), to truly understand the point that I'm making in this section we need to keep all three phrases on the table. For, as we'll see, while error and responsibility separate conceptual representers (i.e., thinkers) from other representational systems, these ideas only come into play when we look at both the nature of the representations themselves, and the ways in which they are produced. I think it will be most beneficial to keep this in mind in the succeeding discussion.

---

[76] I do not mean to imply, by using the term 'action' here, that the production is intentional (as in, "I meant to produce that token"), that the system is an 'agent', or even that the system is a cognitive one.

4.13 Two Examples

I have used the notion of error as the entering wedge, one might say, for my criticism of modern representationalist accounts of thought. In doing so I have relied largely upon an appeal to intuition: it seems quite obvious that when a computer miscalculates the machine is not itself at fault; but when a thinking being employs, say, faulty reasoning that being *is* itself at fault. There is a difference, to put it bluntly, between machine error and human error.[77] But if one endorses the unity and continuum claims, one cannot explain (without explaining away) this intuitive difference.[78] I will now try to flesh out the intuition that there is a difference between natural and conceptual representation, and argue that thought requires agency and sensitivity to the norms that govern the representations that constitute those thoughts—and, hence, that the unity and continuum claims must be mistaken.

Let's begin by comparing two representations, one simple and natural, the other complex and conceptual. To avoid as much as possible being contentious at this stage, I will pick for the former the representation in a bee's brain of a visually detected flower. I could just as easily have chosen to use the frog of our earlier example, or a dog, but I

---

[77] I juxtapose machines and humans when putting the point bluntly, because I think that everyone can agree that there is a sharp divide in that case. But I should note that in what follows I will be initially less concerned with *where* one draws the line that separates conceptual from non-conceptual representation as I will be with *that* and *why* one draws such a line. I am not, that is, simply *presupposing* that only humans have conceptual representations.

Furthermore, as an aside, I should note that the problem of error encompasses two different types of error: (1) the kind of mistake that is made when we can characterize a system's representation as simply false (for example, a false perceptual belief), and (2) the kind of mistake that is made when a system is unjustified in tokening a particular representation (for example, when making a faulty inference). Recognizing this distinction here does not change anything that has thus far been said.

[78] I want us to recognize here that the issue is *not* the classic one in the literature on misrepresentation of how to account for 'genuine' misrepresentation. Everyone from Fodor to Cummins addresses this issue in one way or another—but all of their accounts allow for 'genuine' misrepresentation in systems where error, when it occurs, is just malfunction. *That* is precisely the problem, for then these theorists are forced to say that error in *human thought* is just malfunction as well. And *that* is what seems intuitively wrong.

wish to claim that the representation in question is non-conceptual, and there are many who would find it immediately implausible to assert that, say, a dog's representation of its environment isn't conceptual (though they may agree that it isn't as richly conceptual as a human being's representing of her environment). I take it, though, that most of us would agree that bees simply lack the sophistication required for the formation of conceptual representations. (If you disagree with this, simply move further down the phylogenic table until you reach an organism that you're comfortable asserting has representations, but lacks concepts. It doesn't matter for what I'm about to argue where you choose to draw your line between those systems that are conceptual and those that are not.) For the latter example, of a conceptual representation, let's use my belief that today is Thursday. At this point it is not important for the representations being compared to have similar features (e.g., to both be visually-triggered representations of the environment).

Since the bee's representation is *ex hypothesi* non-conceptual it cannot, of course, really *mean* 'flower' (at least, not for the bee, since the bee lacks the concept *flower*); but I wish to stipulate that we've managed to solve the misrepresentation problem in this case. That is, we can assert that this representation is a representation of the presence of a visually detected flower, and if triggered in response to, say, my big toe would constitute a misrepresenting.[79] On the other hand, my belief that today is Thursday is

[79] It may be that in characterizing this representation as having determinate content (it represents the presence of a flower and not my big toe) we are simply interpreting the representation from our point of view, but there are stories that can be told such that this needn't be the case. For example, Millikan's biosemantic approach to representations would allow us to assert that the representation in question represented the presence of a visually detected flower for the bee itself, independent of our interpretive enterprises, for the representation will only serve its evolutionary function when tokened in response to flowers and not when tokened in response to my big toe. Either way, though, the representation is not a

conceptual, and has the semantic content that I believe that today is Thursday.[80] Of course, since it has a determinate semantic content, my belief can easily be a misrepresenting—if, for example, today is actually Friday. So, we have two representations, both with determinate content (so that the misrepresentation problem isn't going to distinguish them).

Now, for a representationalist who accepts the unity and continuum claims, the only difference between these two representations, by which we could account for the fact that one is conceptual and the other is not, lies in the complexity of the representations themselves. That is, for those who endorse the unity and continuum claims, these two instances of representation differ not in kind, but only in degree. And if at some point degree of complexity moves us across the divide between the non-conceptual and conceptual—well, this isn't really a divide, since all representing exists on a continuum. It's more likely that the 'conceptual' is just a fancy way of referring to appropriately complex representations.

I contend, however, that there is a sharp divide between these two examples of representations—I contend, as I've said, that they are not of the same kind at all, i.e., that they differ not only in degree of complexity (there's no contesting that), but that they differ dramatically as *types* of representations. How shall I make my case?

First of all, let me say that I do not think that the *contents* (whatever they may be) of the representations are the right things to focus on. Rather, what we need to focus on

---

conceptual one. (Well, *we* attach concepts to the representation, of course, but there is no claim that the representation constitutes the possession *on the part of the bee* of the concept *flower*.)

[80] Or, if you like, the representation has the semantic content 'Today is Thursday', and I am standing in the 'believing' relation to that representation—but see my earlier footnote in which I expressed my distaste for this way of thinking about the propositional attitudes.

is the *way* in which each representation is tokened. This is not to say that content is irrelevant, of course (nor should I be understood as suggesting that the 'way in which a representation is tokened' is a simple matter), but (as we'll see when I turn to a more contentious set of examples below) even when the contents of the representations being compared bear a greater similarity to each other than in the present case, the difference between them as *kinds* of representations will still turn more on the manner of their tokening than on their content (at least as far as content is separate from the way in which the representation is tokened—a separation that is not, of course, total).

So, what do I mean when I say that we need to focus on the *way* that each of these representations is tokened? Let's discard a few possible red herrings first: the difference does not lie in the fact that one representation is tokened in the small, relatively simple nervous system of a bee, while the other is tokened in the (presumably) relatively complex brain of a human being, nor does the difference lie in the fact that the bee's representation is a direct response to a visual stimuli, while my representation is a response to other conceptual states of my mind. I do not mean that the bee's representation is non-conceptual while mine is conceptual—though I have stipulated this difference, this is nevertheless the difference that I'm trying to explain, so it would be circular to suppose that the *manner* of the tokenings already included (rather than explained) this difference. I am also not about to appeal to a mental vs. non-mental distinction[81], nor am I going to claim that my representation is part of a functional system

[81] If I wanted to emphasize such a difference as that, I would have used a computer rather than a bee for my first example, though even in that case the mental vs. non-mental distinction would be largely irrelevant—I say 'largely' because obviously non-mental representational states (again, e.g., the states of a modern computer) certainly cannot be *thoughts*, so the distinction is relevant to that extent.

of representations while the bee's is not—both of our representations, I'm happy to grant, are functionally defined at some general level.

Having safely placed those potential distractions aside, what I do mean when emphasizing the *way* in which each representation is tokened is that *the system of representations of which each token is a part, and the context* (the *representational* context, not the physical context) *of the tokening of each representation*, separate the bee's visual perception of a flower, and my belief that today is Thursday. These are the things that make the bee's representation non-conceptual while mine is conceptual. These are the things that make my representation a *thought*, while the bee's is not.

We start with the bee's representation of a visually detected flower. What is the manner of its tokening? First of all, the representation itself, while part (we might suppose) of a functional network of representations, is primarily a causally determined response to the stimuli[82] in question. The functional network of which it is a part, that is, is a *causal*-functional network, much like, say, the inner workings of a computer, or some other complex machine. The story that we will tell to explain its having the content that it does is one that makes reference only to its place as a node in this causal network.[83] In simplistic, if nevertheless illustrative, terms: the system of which this representation is a part is simply a matter of *wiring* (so to speak).[84]

---

[82] Again, it doesn't matter that the stimuli is external rather than another representational state of the bee's internal network. What matters is the nature of its causation.

[83] This will be true, I think, even if our story is, say, a teleosemantic one. Granted that the representation gets its content from the evolutionary role that it must play—nevertheless, it plays that role because of its causal relations to the world and other representations of the system. Modern representationalists are all functionalists first.

[84] We might even suppose that bee representations are hard-wired. But that is an unnecessary distinction to draw, and ultimately also a red herring, since I will later bring in an example in which I believe the representational system in question is *not* hard-wired, but fails to be conceptual all the same.

If the system itself is simply a causal-functional one, the context of the tokening is also quite basic. The representation is tokened both as a simple response to stimuli, and, most importantly, in the *absence* of any *recognition* on the part of the bee that the token is *appropriate* given that stimuli. This latter point is the most crucial for my argument, for as I will claim in just a moment, my representation *is* tokened in recognition, on my part, of the appropriateness of its tokening. But with respect to the bee's representation, can there really be any question that the bee itself does *not* recognize, does not *understand*, that just *this* and no other representation *ought* to be tokened in its present context? In order for it to recognize the correctness of the representation, too many things would have to be the case: the bee would need to have the ability to represent to itself (a second-order representational capacity) *that* it was currently tokening a particular representation, which would of course require a sense of self (of how robust a nature is open to dispute, of course), so that it could claim its representations as its own; it would need to have the capacity to represent separately from the individual token under scrutiny the *context* in which the token was occurring (that, e.g., the perceptual conditions were normal, such that what looks to be a flower really is a flower); it would have to be able to represent to itself the *correctness* or *incorrectness* of tokening just this representation under these circumstances (a rather complex combining of the earlier representational abilities that the bee would need to possess); it would, in short, require a massive amount of additional representational capacities the vast majority of which, at least, it seems nearly absurd to suppose a simple organism like a bee could possess.[85]

---

[85] Again, if you grant to bees greater powers of cognition than I am doing here, simply move down the ladder to a simpler organism and my point will stand. At this stage it makes no difference *where* the line is drawn, only that it *is* drawn. Note, too: all I've said so far is that the bee lacks the ability to recognize the

What, then, of the manner in which *my* representation (the belief that today is Thursday) is tokened? For each point made above about the bee's representation, there is a related counter-point to be made about my own representation. First, my representation is not *merely* part of a causal-functional network; it is not simply a matter of wiring. Rather, my representation is a rule-governed token that places it within the space of reasons. To be a belief, the representation must be able to enter into certain relationships with other representations, especially that of justifying and being justified by those other representations, for that is a great deal of what beliefs *do*. The most important point to keep in mind, here, though, is this: conceptual representations are, unlike natural, non-conceptual representations, not primarily characterized by their physical, or even *causal*-functional, features; rather, they are characterized by those functional features whereby they can be considered to be part of a system of *rule-governed* symbol tokenings. Recall that in looking at the problem of error, and using it as an entering wedge for my criticism of those representationalists committed to the unity and continuum claims I noted that the key issue seemed to be one of *responsibility*. A computer is not responsible for its representational errors, but a thinking human being is. But being 'responsible' implies having a responsibility to someone (or something). That is, to make sense of our intuitive notion of responsibility when it comes to misrepresentation (and also correct representation, of course) we must posit a community of representers. Thus, I contend, to be rule-governed, a symbol tokening must be part of a *practice* of giving and asking for

---

correctness of its tokening of the representation in question; I have not yet argued that this ability is requisite for conceptual representation. But that argument is soon in coming.

reasons that determines under what circumstances a symbol should, and should not, be tokened.[86]

If being rule-governed involves a complex practice of giving and asking for reasons, however, then the context of my tokening of the belief that today is Thursday is immediately quite different from the context of the bee's tokening of a representation of a visually-detected flower. For one thing, my belief cannot *merely* be a simple response to some stimuli. No doubt it is caused (I do not deny that conceptual representations are broadly to be construed as causal-functional states), but the cause is not the important part in making my representation a conceptual one. What makes my representation conceptual is the fact that it is part of a particular rule-governed system (the exact character of which I have not yet specified), and (even more importantly—in fact, crucially) that my tokening of that representation is (in some sense) *recognized* by me as tokened *appropriately*. That is, it cannot simply be that my tokening occurs at the right time in accordance with the rules governing it; I have to be capable of recognizing my authority to token just that representation under the circumstances—I must (in some sense) *know* that the token is part of a rule-governed system of representations.

Notice how this notion works with the problem of error. One might suppose that there is a fundamental tension in the view I'm putting forward, for how can one recognize the appropriateness of a representational token that is, in fact, *inappropriate* (e.g., false, or unjustified)? Take my belief that today is Thursday. Now suppose that today is actually Friday. My belief is thus a misrepresenting, and is therefore an inappropriate (in

[86] I am, of course, only broadly sketching at this point the arguments that I will be developing later. All this talk of rules and public practices will be fleshed out primarily in Section 4.2 below. For now, I only want the bare structure to be clear.

the sense of false) representational tokening. Yet surely it's still conceptual. Having just tied a representation's status as conceptual or not to the representer's recognition of the token's appropriateness, though, how can I claim that an inappropriate representational token is nevertheless conceptual? The gist of my answer is fairly straightforward. Whatever else recognizing the appropriateness of my representational tokens comes to, it is at the very least an awareness on my part, as I said in the previous paragraph, of my authority to token such representations. Such authority, however, does not guarantee infallibility, nor does recognizing my general authority to token these representations make those tokens authoritative (i.e., true, or even epistemically justified). What makes the representation conceptual isn't that I've 'got it right,' but rather that I, again, (in some sense) *know* that it's part of a rule-governed representational system, and that I am at least implicitly obeying those rules when I token it. That I might be wrong and in fact violating those rules means only that I have made an error. My token is conceptual, though, so it is an error for which I can be held responsible (which is exactly the result we're after).

Recall, of course, that I am intentionally leaving what it means for one to 'recognize' that a representation is tokened 'appropriately' somewhat vague for now. This issue is complicated—and part, I think, of a larger collection of very difficult issues that even Sellars has not dealt with thoroughly. Yet I do not think the indistinctness of the idea of 'recognizing the appropriateness' of one's representations is a problem here. The skeleton of the argument thus far will stand, it seems to me, even if we do not give a determinate character to the requirement that conceptual representations be recognized by the representer as in fact appropriate or warranted.

This is not to say that we can't yet paint at least a broad picture of what is required for conceptual representing. The central argument of this section is that conceptual representing requires a sense of responsibility—which, as I noted above, seems to require a community, and a practice of giving and asking for reasons. As will become clearer later, I think that the ability to recognize as warranted (appropriate, in keeping with the rules, etc.) a representation that I token *now* is built on a long history of tokenings of a related sort that while *in fact* appropriate[87] (given the rules governing the symbolic system) were not *then* known *by me* to be appropriate (and hence were not *then* conceptual representations). Of course, it follows fairly closely upon this supposition that there must have been others (my trainers) who *could* recognize the appropriateness (or inappropriateness) of my tokenings, and conditioned me to token the right representations in the right contexts (and here we find ourselves back to the notion of a community). But eventually, we may suppose, I 'got the hang of it' myself, and became, as it were, self-correcting—and hence able to recognize the rightness or wrongness of my own tokenings.[88] The idea that coming to have conceptual representations involves a training process plays an important role in the LTT.

---

[87] In addition to the fact that I am leaving the notion of '*recognizing* the appropriateness' of one's representations intentionally vague, we should also note that the very term 'appropriate' is itself vague. It might mean (in the context of following rules) simply that the rules are in fact being obeyed. Yet, 'appropriate' might mean other things as well: for example, it might mean that what one does not only conforms to the rules, but is also the polite (e.g., saying 'Thank you' when given a gift—the 'rules' here being those of etiquette) or wise (e.g., making a particular move in chess—the 'rules' here being those that govern the game) thing to do. I think we can say that, at the very least, the appropriateness of one's conceptual representations will have to do with the representations conforming to the rules that govern the representational system. But I do not want to suggest that in all cases this is *all* that 'appropriate' comes to.

[88] Just how this transition in fact gets made is a great question, and perhaps one of the most difficult to address of all the questions within the nexus of difficult problems mentioned earlier. Sellars himself never answers this question very clearly, instead taking brief stabs at it here and there (he—and I agree—sees it as a probably gradual process, with growth and learning in many dimensions taking place simultaneously). I address this issue in §4.22 below, but I do no better than Sellars does, taking only a small stab at the problem. Wrestling with this problem will, I think, make a good project for the future.

Let's look now at another set of examples that I believe will more clearly illustrate the importance of the *manner* of a representation's tokening in making it conceptual. This set of examples will differ in two important ways from the set of examples that we have just looked at. First, unlike the above examples, the two examples below will involve representations that are as similar as possible (while still, I will contend, remaining separated by the fact that one is conceptual and the other is not). That is, the representational context of the tokenings will be nearly identical, and the causal-functional roles of the representations will be quite similar (in a way in which the representation of a visually detected flower and the belief that today is Thursday are not). Second, in this new set of examples I will try to bring the representational systems much closer together on the phylogenic table than a bee and human being are.[89] The importance of this difference will be clearer as we proceed, but to put it simply: I want to have two organisms that share roughly similar perceptual organs (at least as concerns the examples), and have much closer brain structures than a bee and human do.

Here, then, are the two representations that I wish for us to consider. Imagine a dog and its owner (let's call her Jane) both looking at a black ball. I want to consider the representations that occur in each case; i.e., I want to look at the representation in the dog that is the perception of a black ball. And I want to look at the representation in Jane that is the perception of a black ball.[90] As I noted in the previous paragraph, in this set of examples the subjects are physically more similar (dogs have eyes much like ours, and

---

[89] As I will mention later, I could bring the organisms under consideration much, much closer—in fact, I could make them both *human beings*—and still make the point that I wish to make. But that would introduce misleading complications that it is unnecessary to address at this juncture.

[90] We could choose to talk about Jane's perceptual belief that there is black ball over there, but the two representations will be much closer if we leave belief—and all propositional attitudes—out of the picture for the time being and simply imagine that Jane has *noticed*, thus *seen*, the black ball without (yet) forming any particular attitudes about it as a consequence.

their brains and central nervous systems are fairly similar to ours), and the context of each representation's tokening is roughly identical (both Jane and her dog are looking at the same physical object, having their retinas stimulated by similar light waves, etc.). Nevertheless, I think that it is still possible to distinguish between the representations in each case—i.e., I think that it is possible that Jane's representation is conceptual, while her dog's representation is not. Hence, again, the unity and continuum claims will prove to be mistaken. As we did before, I want to look at each representation in turn. And, as I've said, I think that it is the manner of the tokening that will make the biggest difference.

I take it as fairly obvious that when compared to a bee a dog's perceptual apparatus is relatively complex. The dog is capable of taking in and processing a much greater amount of information, and it is also capable of a fair degree of learning (at least when compared with an insect).[91] Thus it seems reasonable to suppose that Jane's dog is capable of much greater feats of representation than the bee from our earlier example. Nevertheless, let's examine the manner of the dog's tokening of the representation of a visually detected black ball. There is no doubt (at least as far as representationalists are concerned) that the dog's representation is part of a complex causal-functional network of representations. There is also no doubt that in virtue of this system of representations the dog is able to perceive and navigate its environment. The dog has what the bee has, only (we may suppose) to a greater degree—and presumably humans have what dogs have, only to a greater degree again. Indeed, we can tell roughly the same story about the

[91] Indeed, it is the perceptual and behavioral capacities of dogs that lead many people to suppose that these animals must have some level of conceptual abilities, even to the point of forming (at least rudimentary) beliefs and desires.

'triggering' of the bee's representation, the dog's, and Jane's. Each begins with the light waves reflecting off an object impacting the sensory receptors of the representer in question, and this in turn, we may suppose, causes the tokening of the representations that we are interested in.

Now, in the case of the bee we supposed that the representation was, while part of a causal-functional network (of a fairly simple variety, most likely), not much more than a causally determined response to the stimulus of the flower in the bee's visual field. What was most important, though, was that the representation was tokened in absence of any recognition, on the part of the bee, that it was appropriate that that representation be tokened in those circumstances.

So now let's imagine what might be going on in the dog when it represents the presence of a black ball in front of it. In principle I see no reason for the story here to look very much different from the story we told about the bee. Given the rather more contentious nature of this example, though, i.e., given that it will seem much less plausible to some people to suppose that a dog's perception of an object is non-conceptual than it did to suppose that a bee's perception of an object is non-conceptual, I think we are warranted in moving a bit more slowly through it.

First of all, I want us to note that the dog's representation is also a response to an environmental stimulus—but, then so too is Jane's representation: i.e., these are all cases of *perception*. So there is nothing (much) to distinguish the various representations here. We've already noted that Jane's and her dog's representations are responses to the impacting of their sensory receptors by the light waves reflected from the black ball itself. There can be little doubt, of course (as we've also already noted), that the dog's

causal-functional network is much more sophisticated than the bee's, but at root it seems we can imagine both systems as similar in kind, differing only in degree. And, again, I'm happy to admit that (on some level) Jane's representational network is simply more sophisticated yet. Thus, as far as the causal story goes, and as far as much of the functional story is concerned, we seem to be in the position at this point of playing right into the hands of the representationalist who endorses both the unity and continuum claims.

Recall also that in supposing that the bee couldn't recognize that its representation was warranted our argument turned *to some degree* on the fact that such an awareness seemed to require *representational capacities* that were entirely too sophisticated to ascribe to a simple organism like a bee (and note that I am here talking simply about the *mechanism* for representation—i.e., probably, the sophistication of the bee's brain—not about anything having to do with norms). So what about Jane's dog? Does it seem reasonable to suppose that it has the sophisticated representational capacities required for recognition of the appropriateness of its tokening of a representation of a black ball? To many people the features that seemed to disqualify the bee, *viz.*, being able to have a second-order representation *that* it was representing such-and-so; being able to represent the context of the occurrence of the representation separately from the actual representation itself; even being able to represent the correctness or incorrectness of the representation—all these capacities, while quite sophisticated, are not so obviously sophisticated that we couldn't imagine a dog possessing them (at least, to some degree—and, of course, if *degree* is the only difference between the dog's representation and

Jane's, then—again—we have lost the point to those who would endorse the unified notion of representation).

Suppose, though, that we *grant* the dog all these further and more sophisticated representational capacities. To some extent such a concession is really beside the point. For recall what really distinguished the bee's non-conceptual representation above from my conceptual representation: my representation was rule-governed and I tokened it in *recognition* of the rules that governed its tokening. The causal story, recall, was more-or-less irrelevant when it came to classifying my representation as conceptual. What was important was somehow being able to know that my representation is rule-governed, and that I am tokening it correctly. But this, I supposed, required a kind of training, and that in turn required trainers: others who were *already* in the space of reasons and could correct me when I misrepresented, until eventually I got the hang of it myself.

In order for Jane's dog's representation of the black ball to count as conceptual, then, it needs to be part of a *rule-governed* representational system, and the dog must token that representation in recognition of its rule-governed nature, and of the appropriateness of its tokening in the present circumstances. This, in turn, would require that the dog have been trained up by others, already in the space of reasons, to be self-correcting with respect to its representations. But here we run into a wall, which to my mind proves decisive. For how can the dog have been so trained? The representation in question (of a black ball in front of the dog) is an entirely *internal* event (internal, we may suppose, in the sense of occurring in the dog's brain, and hence not generally publicly observable). Yet, if the representation is not publicly observable, then no

training of the required sort can take place, for no one could have witnessed the dog's earlier tokenings in order to correct them when they were misrepresentations.

Dogs, of course, can be trained to *do* many things, but 'having a representation' isn't one of them unless there is something that the dog *does* that can be *equated* with its having a certain representation. This brings up an interesting point, for Jane may very well have trained her dog to respond to the command 'Get your black ball!' by going and finding the ball and bringing it to Jane. This behavior would seem to indicate the presence in the dog's representational system of the required representation, *viz.*, the 'black ball' representation, and supposing the dog were well trained we might even imagine that it has 'got the hang of' tokening 'black ball' in the presence of the object. But, first of all, behavior in *consequence* of the tokening of a representation isn't the tokening itself, so there will be a split here between the *behavior* that is corrected for and the *representation* which isn't. (To see this, simply notice that even when the dog fails to grab the 'right' object in response to the command, if it 'sees' the ball, it may still have the representation—the *correct* representation—even though its behavior isn't of the desired sort. But then when we correct the dog, we are correcting not a mis*representing* (for the dog *correctly* represented—in a non-conceptual way, I'd argue—the ball when it saw it), we are correcting a mis*behaving*. And unless the representing and the behaving are the same, the training of the dog isn't of the right sort.) Secondly, even if this first problem can be overcome (as it might—and will have to be in Jane's case, of course) the argument that Jane's dog is trained, and thus has the conceptual representation 'black ball' is only going to work given an atomistic semantics for representations. But as we

saw when we looked at Fodor's view in Chapters One and Two, I do not think that an atomistic semantics for representations is going to work.

Now, however, we must consider Jane's perception of the black ball. For isn't her perception in just the same situation as the dog's, *viz.*, internal in the sense of not publicly observable, and hence uncorrectable by others? I do not think that it is, but the *reason* for this brings us up against a point that we have yet to really flesh out. When I first introduced the normative nature of my belief that today is Thursday I noted that being rule-governed required a *practice* of giving and asking for reasons. I discussed the reason for thinking that such a practice was part of the picture, but I didn't go on to spell out the importance of such a practice. Yet, its central import to the very notion of a conceptual representation flowed beneath everything that I said afterward, for what else is the training of would-be conceptual representers by those already part of the space of reasons if not a social *practice*, a *public* practice of bringing others into that same space? We must now explore this issue a little further (though we will return to it again later).

Let's begin with the obvious objection: "By your own description," says our objector, "Jane's representation is an internal state triggered by the impacting on her sense receptors of light waves reflected from the ball. In what way is this a *public* event? How can this be any more part of a social *practice* than the dog's tokening?" My answer, of course, is to invoke the idea that forms the very core of the LTT. We've already agreed that the representations in question are in some sense symbolic, i.e., are constituted by symbols of some kind. So, what kind of symbols constitute Jane's representation? My answer is the only one that, so far as I can see, makes it possible for them to be part of a social practice: they are *linguistic* tokens. That is, they are symbols

of a public language. But our objector continues: "Suppose I grant that Jane's representation is constituted by symbols of natural language—however that would work. Nevertheless, the present token is still *internal*; Jane isn't talking, she's just silently perceiving an object." This is true. But notice: her *present* tokening is silent. We may suppose that Jane has already 'got the hang of' correcting her representational activity, and so her tokening *now* of a conceptual representation of a black ball needn't be public. All that is required is that some of her *previous* tokenings of the representation of a black ball were public, so that they'd be subject to correction.[92]

Here we come to the problem noted above in the dog's case. Unless there is a sense in which the tokening of a representation is equated with some publicly observable behavior the idea that conceptual representing requires a *public* practice of correction and conditioning will never get off the ground. Given my claim that Jane's representations are constituted by linguistic symbols, then, the solution is obvious (though not simple). Imagine Jane before her representation counted as conceptual. Imagine her tokening *out loud* the words 'black ball' in the presence of the object of our example. Now, in the sense that this is simply behavior in response to internal states of Jane's representational system we seem to have the same problem as we did with the dog: we can correct the behavior if it goes wrong, but the representation is a different matter altogether. There is a crucial difference here, though. In the dog's case, given that it doesn't speak English, it

[92] Even this is too strong, for it may be that Jane became a thinker, and got the hang of color and shape concepts, without ever seeing a black ball. Upon first seeing a black ball, then, she may silently token |BLACK BALL| even though she has never publicly tokened this representation. What she *has to* have done, however, is publicly tokened enough color and shape representations (e.g., 'red', 'yellow', 'cube', 'pyramid') to have entered the space of reasons with respect to such concepts, so that she can conceptually represent the current situation—and for her current token to really be a 'black ball' token, she probably will have to have tokened 'black' and 'ball' both publicly at some time in the past, though not necessarily together.

seems unusual (to say the least) to suppose that *its* representations are constituted by English symbols. Nor (and this point is even more important) is the dog's behaving an act of *representing*. But in Jane's case the very opposite is (I am claiming) true: Jane's behaving *is* an act of representing, and since she is being brought up to speak English, I am supposing that both the overt tokening (her *saying* 'black ball') and the covert tokening (the internal brain state, say) are tokens of *English symbols*.[93] Then, as Sellars has said many times over, there will be a very real sense in which when language-using creatures speak they are not simply exhibiting behavior caused by representations—*they are tokening those representations themselves*.

Suppose, though, that my opponents were to grant all this for argument's sake. Still there seems to be a problem, for even if Jane has built her current ability to token |BLACK BALL| in the present circumstances on a public practice of overtly tokening the English words 'black ball', nevertheless there had to have been *something* going on in Jane upon perceiving a black ball *before* she began to speak at all. That is (says the objector), certainly it's not the case that *every* time Jane has seen a black ball the representation that got triggered was a token of English symbols (whether covert or overt). In this case I think that the objector is exactly right: there *has been and probably still is* something going on in Jane upon seeing a black ball that, while necessary for Jane to conceptually represent the presence of a black ball, is *not* the tokening of English symbols—but I don't see this as a problem. What is the nature of this additional

[93] In imagining that Jane's covert (internal) tokenings are also of *English* symbols, I am simply picturing a situation in which, say, certain neural states in Jane encode English symbols. While intrinsically these states are simply physical states of Jane's brain, given that Jane is an English speaker and thinker, I am supposing that those brain states can be interpreted as, again, neural encodings of actual English symbols. They are, to put it bluntly, just another way of expressing English tokens—much as such tokens can appear as written symbols on a page, as spoken symbols when one talks, hand shapes and movements when one uses sign language, or even as coded tokens in clandestine communications.

something? Presumably it bears some resemblance to the *dog's* representation of the black ball—perhaps it's even an additional symbol-tokening that forms part of a causal-functional network and is (generally) caused in response to the impacting of the visual sensors by light waves reflected from a black ball. But even if it is an additional representation in its own right, it is *not* conceptual, for in order to be *that*, it would need to be part of a rule-governed system of symbols, and tokened in recognition of the norms that govern these symbols, which is only possible, again, when the tokens are *public*—hence, speakings (in a broad sense—what Sellars calls 'languagings' so as to include all forms of linguistic expression).

Let me put that last bit another way. I imagine that in adult human beings (who we are granting have conceptual representations), *perceiving* is *conceptually informed.* I take it that this is a relatively safe assumption, since we can all attest from our own first hand experience that thinking beings' perceptions are always conceptually informed.[94] I think it is clear that Jane will see the black ball *as* a black ball, and that requires that she have the concepts 'black' and 'ball'. But if having a concept is to be able to token a conceptual representation with the requisite content (and what else, on a representationalist view, would having a concept be?), then Jane must be able to token conceptual representations having the contents 'black' and 'ball'. Thus, Jane's *seeing* of the black ball is a conceptual representing *of* a black ball *as* a black ball. Her seeing occurs, therefore, like my belief in our first set of examples, within the space of

[94] The concepts in question may not always be the same across persons: what the astronomer sees as a brightly burning ball of gas the uninformed (say any of Plato's contemporaries) will see as a pinpoint of light in the sky. This isn't to say that their *perceptions* are different, only that the concepts that inform those perceptions can differ. But to anyone who doubts that all the perceptions of a thinking being are conceptually informed in one way or another I simply challenge you to provide a counterexample.

reasons—which is to say that it is part of a practice of asking for and giving reasons to justify one's representations, a practice that can only take place when participants in the practice recognize that their representations are rule-governed, as I argued above. This is just to say, however, that when *thinking* beings, say, *see* something their representations have something akin to that had by the representations of their unthinking brethren while also having something not possessed by the unthinking creatures: a normative feature. To put it another way, *conceptual* representation is built on top of a base level of non-conceptual representings.[95]

Let me summarize where we stand with the current example set. Both Jane and her dog have their retinas impacted by light waves reflecting off a black ball. In both cases this stimulation of sense receptors triggers a representation, say in the brain of the perceiver. Jane's representation is part of a causal-functional network, as is her dog's representation. Jane's dog's perception of a 'black ball' is simply constituted by this functionally defined representation, and it does not include an awareness of the black ball *as* a black ball; it is not part of a *norm*-governed representational system; and,

---

[95] Indeed, I don't really see how it could be otherwise. If the ability to have conceptual representations is developed as one learns to speak there has to be *something* going on in the language learner for the trainers to build upon, otherwise how could they get started? Other representationalists have supposed that this is reason to grant innate concepts and/or awareness of 'logical space' prior to language learning. But I take the lesson to be very different. Rather than innate concepts I require only non-conceptual representations. (Given that those representationalists who endorse the unity and continuum claims don't acknowledge a sharp distinction between conceptual and non-conceptual representations, this way out isn't really available to them.)

I'm not sure, however, that we need suppose there to be two *completely separate* representations in Jane's case—or in any case of conceptual representation. I think that in perceptual situations, such as we are imagining Jane to be in, there has to be *something* akin to a non-conceptual representation that gets tokened; but it could very well be that once one becomes a 'conceptual representer' the non-conceptual representation gets 'absorbed', say, into the conceptual representation, so that there is still only one representation when, e.g., Jane looks at the black ball—a representation that is *like*, in some ways, that had by Jane's dog, but is *conceptually* informed, as the dog's is not. (This suggestion matches up nicely with the earlier interpretation of 'recognizing the appropriateness' of one's representations as a matter of the recognition being 'built into' the representations themselves.) Furthermore, thinking beings have all sorts of representations that do not involve any sensations, or perceptions, at all—and in these cases there needn't be *anything* that we might suppose the thinking and non-thinking beings have in common.

consequently, it cannot be tokened in awareness of any rules governing its tokening. It is, in short, a non-conceptual representation. Jane's perception of the black ball, however, is *not* simply constituted by this functionally defined representation: her representation of the black ball includes an awareness of it *as* a black ball; her representation is part of a norm-governed system of representation, and is tokened in recognition of the rules that govern its tokening because her representation is constituted by linguistic symbols that, during her 'training' period, as she was moving from mere animal to thinking being, she was conditioned by her trainers to token under just these kinds of circumstances.

Assuming that the previous paragraph has summarized things nicely, I can add even more complication to the picture of thought that is emerging. As I mentioned earlier, there is no reason that my second example set had to involve a *dog* and a human being—it could just as easily involve *two human beings*. Of course, while that may initially sound shocking, we are now, I hope, in a position to see the banality of the point (on the assumption that the details of the various parts of the argument just sketched really do work out below, of course), for all I'm suggesting is that we compare a human being who has been properly trained and has finally become a conceptual representer—a human being who has reached the maturity of a thinking being—and a human being (most likely a pre-linguistic infant) who has not. If having conceptual representations is a matter of having been conditioned to token the right representations in the right circumstances, and to recognize (in the appropriate sense) that these representations *are* the right ones, because they're governed by rules and one is obeying the rules, then it should be *obvious* that no human being—no being whatsoever—could simply *start out* as

a conceptual representer. Thus, I could just have easily compared Jane and her newborn son as Jane and her dog.[96]

I set out at the beginning of this section to sketch an argument to the effect that there is a feature of thoughts, *viz.*, their norm-governed nature, that cannot be captured by most modern representationalist views. I think now that I have shown, roughly, why that is the case. For representationalists who endorse the unity and continuum claims, there can be no difference in principle between the representations in non-thinking organisms and those of thinking beings. So, of course, it cannot be the case that the representations of thinking beings are fundamentally distinct from other types of representations by being rule-governed, socially conditioned, and tokened only in recognition of the rules that govern their tokening. But this means that there cannot be any principled difference between the misrepresentations of thinking beings and those of non-thinking beings, or even machines. And that seems to me to just be obviously mistaken. When a thinking being makes a mistake it is a mistake for which that being can be held responsible. Yet, the only reason that we believe thinking beings, such as our fellow humans, to be responsible for their errors (whether in representing something as the case when in fact it is not, or in claiming that a representation is warranted when it is not) is that we believe their representations to be governed by norms—norms that we expect our fellow thinkers to recognize. We suppose, that is, that our fellow thinking beings are fundamentally *rational* creatures.

---

[96] There is, of course, one *very, very* important difference between Jane's newborn son and Jane's dog: the former, but not the latter, is *capable* of being trained into a full-blown thinking being. That is incredibly far from being a small difference—it makes all the difference in the world that the baby is, as it were, born *ready* to become a thinker, while the dog will *never* be one.

The burden of the rest of this chapter, then, is to make the case that thoughts are indeed governed by social norms regarding language use. And even if I cannot prove, here (in the scope of a single dissertation), that this approach works as a complete theory of mind, I hope at least to show that it is a *coherent* possibility; that is, I hope to be able to clear the logical space for this view as a first step toward defending it as the right way to go. I am not alone, however, in thinking that this approach is the correct one; I am not alone in believing that thinking is a norm-governed activity based upon the social practice of language use. These ideas have come to me from Sellars' work, and so it is to the Sellarsian framework of Psychological Nominalism that I next turn.

4.2 The Sellarsian Framework for the Integration of Language and Thought: Psychological Nominalism.

In this section I will undertake to explain and argue for the various features of Sellars' Psychological Nominalism that are more-or-less at the heart of the LTT. Of course, some of what Sellars says will need to be adapted a bit for my own purposes—but that will primarily be left to the next chapter. I do not, of course, intend here to undertake a thorough analysis of all the features of, and arguments for, Psychological Nominalism; to do so would require entirely too much space. But this section will be quite substantial anyway, as there remain many arguments to be explored—arguments that are central to the entire framework of the LTT, and serve to ground many of the claims made in other chapters.

A great deal of the large corpus of Sellars' work is relevant to the issues now at hand. In what follows I will divide my treatment of Sellars' view, and of his various works, into three parts: (1) an exploration of his 'verbal behaviorist' framework; (2)

issues regarding the learning of language; and (3) the final integration of language and thought.

4.21 Language and Rules: The Verbal Behaviorist Framework

Let's begin with an observation that Sellars makes at the beginning of "Some Reflections on Language Games":

> It seems plausible to say that a language is a system of expressions, the use of which is subject to certain rules. It would seem, thus, that learning to use a language is learning to obey the rules for the use of its expressions. However, taken as it stands, this thesis is subject to an obvious and devastating refutation. (1963, p.321).

The refutation in question takes the form of a *reductio*. The idea that learning to use a language involves learning to obey the rules of that language seems to require that one have access to a meta-language in which the rules in question can be formulated. But then learning the meta-language would require a meta-meta-language, and so on infinitely—the absurdity of which seems to falsify the original thesis that learning a language is learning to obey its rules[97] (Sellars, 1963, p.321).

If we are to hold onto the (intuitively quite plausible) claim that learning a language is learning to obey its rules, we will obviously need a theory of language and language learning that does not fall afoul of the above refutation. Of course, on what we've been throughout this work calling the traditional view of the relationship between thought and language, this issue seems not even to come up. But, as we shall see, according to Sellars' analysis the traditional view is actually imperiled by a very similar

---

[97] Since this refutation takes the form of a *reductio*, the absurdity may be taken to refute the premise that learning the rules of a language requires knowing a meta-language, or that the rules of a language are formulated in a meta-language, instead. We'll consider these options, briefly and somewhat tangentially, below.

reductio. More important for our purposes, however, is the fact that Sellars has a solution to offer that does not endorse the traditional view; and an examination of his solution will serve as a convenient way of introducing the framework of 'verbal behaviorism' (VB), the insights of which will be central to the development of the Linguistic Theory of Thought.

There are a number of concepts and distinctions that are ultimately going to be important in not only laying out the VB framework, but in dealing with the issues that arise in the subsequent sections of this chapter below. There is no obvious place to begin in introducing these important concepts and distinctions, so I will simply begin with Sellars' discussion of the above thesis about language and his way around the obvious refutation. The concepts and distinctions that I'm interested in will fall out in their own time, and we will have plenty of occasion to talk about them, as they will come up again and again throughout the rest of this chapter (as well as the rest of this dissertation).

To begin, then. Immediately following his statement of the *reductio* that seems to defeat the thesis that learning a language is learning to obey its rules, Sellars introduces the first distinction that will become ever more important to our project here. He writes:

> Now, at first sight there is a simple and straightforward way of preserving the essential claim of the thesis while freeing it from the refutation. It consists in substituting the phrase 'learning to *conform to* the rules...' for 'learning to obey the rules...' where 'conforming to a rule enjoining the doing of A in circumstances C' is to be equated simply with 'doing A when the circumstances are C'—regardless of how one comes to do it.... A person who has the habit of doing A in C would then be conforming to the above rule even though the idea that he was to do A in C had never occurred to him, and even though he had no language for referring to either A or C. (P.322).

The distinction between *conforming* to rules versus *obeying* them will become a dominate theme throughout our discussion of language, language learning, and the relationship between language and thought. For now we must simply note the way in which this distinction proposes to get us around the earlier *reductio*. The suggestion is that while learning to obey the rules of a language seems to require mastery of a metalanguage which in turn requires mastery of a meta-metalanguage and so on infinitely, conforming to the rules of a language (given the definition of 'conforming to a rule' in the quote above) does not (p.322). The proposed solution thus seems to quite simply and easily free us from the obvious refutation with which we began.

This solution, however, may seem (Sellars tells us) to have too high a cost. He writes:

> Is conforming to rules, in the sense defined, an adequate account of playing a game[98]? Surely the rules of a game are not so 'externally related' to the game that it is logically possible to play the game without 'having the rules in mind'! Or, again, sure one is not making a move in a game (however uncritically and unselfconsciously) unless on is making it *as a move in the game.* And does this not involve that the game be somehow 'present to mind' in each move? And what is the game but the rules? So must not the rules be present to mind when we play the game? These questions are both searching and inevitable, and yet an affirmative answer would seem to put us back where we started. (P.323).

The initial appeal of changing 'obeying' to 'conforming to' in the thesis about language learning came from the fact that we defined 'conforming to rules' in such a way that one could perform actions that accorded with the rules without having to be able to state those rules (or even be aware of those rules) first. This extricated us from the regress, of

---

[98] On the previous page Sellars writes: "[T]here are many modes of human activity for which there are rules (let us stretch the word 'game' to cover them all)" (p.322). Thus, language is a 'game' in this stretched sense. While I do not generally talk about language as a game, I will do so when doing so makes it easier to keep in line with Sellars' writings.

course, but the worry now is that merely 'conforming to' the rules of language (or any 'game') in the sense defined doesn't actually constitute using language (or playing the game) at all.

An analogy with another game, here, might be helpful. Imagine a small child moving a piece on a chess board. Suppose that the move actually conforms to the rules of chess—it is white's opening move, and the child advances a pawn two squares. This move conforms to the rules of chess. But we can imagine the child making this move without having any notion of the rules of chess, or even any notion that moving the pieces is part of a game. Yet in that case it seems we would not want to say that the child is 'playing chess'—for in order to play chess one must at least know that one is playing a game, and the child is merely moving the pieces for fun, say. We might even complicate the picture by supposing that the activity continues (with an adult, or perhaps another child, moving the black pieces—that's not really relevant at this stage[99]), and that each time the child moves a piece, the move conforms to the rules of the game. The example becomes less and less plausible as the moves pile on, of course, but in principle we might imagine it taking place. Still I do not think that we would say the child is playing chess, for it is still the case that the 'moves' the child is making have nothing to do with the game of chess. The child is simply playing around with funny-shaped objects. And yet, if we require that language learners be somehow 'aware' of the language game and its rules in order to learn language in the first place, we're back off on our regress (more on this in a moment).

---

[99] Though it will become relevant below, for (though I do not wish to get too far ahead of myself here) as we develop Sellars' account of language learning we will see that the presence of those who already speak the language (know how to play the game) is quite necessary.

Sellars' solution is to develop an understanding of 'conforming to rules,' 'obeying rules,' and other related concepts that avoids simply putting us right back at the start. The development of this understanding is my primary purpose in this section (though said development will carry over into the remaining sections to some degree as well). Thus, let me say up front, I believe that, properly understood, the above appeal to 'conforming to' versus 'obeying' the rules of language will, in fact, allow us to escape the *reductio*. To see how, though, requires some work. Let me continue to follow along with "Some Reflections on Language Games" for a little longer before I break off to discuss the most relevant points.

Perhaps somewhat ironically (though perhaps not, given his penchant for finding some grain of truth in every philosophical position) Sellars locates a guidepost to the proper (on his view) understanding of the language game in the view of 'Metaphysicus,' the proponent of the more traditional theories of mind. Sellars writes:

> As a matter of fact, [Metaphysicus] promises a way out of our difficulty which combines the claim that one is not playing a game—even a language game—unless he is *obeying* (not just *conforming to*) its rules, with the claim that one may obey a rule without being able to use the language—play the language game—in which its rules are formulated. To do this he distinguishes between the verbal formulation of a rule and the rule itself as the *meaning* of the verbal formula. He compares the relation of rules to rule sentences with that of propositions to factual sentences. Whether as Platonist he gives rules an 'objective' status, or as Conceptualist he makes their *esse* dependent upon *concipi*, he argues that they are entities of which the mind can take account before it is able to give them verbal clothing. (P.323).

And here we have the quite traditional view of the relationship between thought and language. One may 'grasp' a proposition before one is able to put it into words. So, too, Metaphysicus' solution goes, one may understand the rules of the language game without

being able to express them in words. Yet if the language learner can be aware of the rules of a language without needing a metalanguage in which to express those rules, then our regress seems to vanish. Thus Metaphysicus claims to have solved our problem.

Sellars is not convinced, though. In fact, he argues that this 'solution' is subject to the very same sort of *reductio* that started this whole mess. Briefly, Sellars' argument goes like this. We ask ourselves what is entailed by the non-linguistic 'awareness' of the rules of language that makes Metaphysicus' solution seem to work. Sellars writes:

> It is clear that if Metaphysicus is to succeed, becoming aware of something cannot be to make a move in a game, for then learning a game would involve playing a game, and we are off on our regress. Yet when we reflect on the notion of being aware of propositions, properties, relations, demands, etc., it strikes us at once that these awarenesses are exactly *positions* in the 'game' of *reasoning*. (P.324).

Granted, on Metaphysicus' view one needn't be able to express the rules of the language game in order to obey them; but one must still be aware of those rules, because we have already seen that merely conforming to them will not do. The game now is the game of reasoning, not the language game, but the worry is the same: if learning to play the game involves playing a meta-game, our regress reappears. To be honest here, I'm not really concerned with whether or not one takes Sellars' argument against Metaphysicus to be successful. I have many other reasons for rejecting Metaphysicus' solution, all stemming from the fact that I reject the theory of mind that underlies it. That is not relevant here. We have looked briefly at Metaphysicus' 'solution,' and Sellars' rejection of it, because Sellars claims that there is a kernel of truth to found there—or, at any rate, that in a limited respect Metaphysicus had the right idea. And that idea launches us in the right direction.

Sellars has us ask ourselves, "What was it about the proposal of Metaphysicus which seemed to promise a solution?" (p.324). The answer, according to Sellars, "is that Metaphysicus sought to offer us an account in which learning a game involves learning to do what one does *because doing these things is making moves in the game* ... where doing what one does *because of the moves* need not involve using language about the moves" (p.325). Where Metaphysicus went wrong, says Sellars, "was in holding that while doing what one does because of the moves need not involve using language about the moves it does involve *being aware* of the moves demanded and permitted by the game," which is what led to the regress (p.325).

And now we come to the crux of the issue. How is it possible to make moves *because* those moves are demanded by the game without it being the case that one is aware, in some way, that those moves are demanded by the game? Our problem was that 'merely' conforming to the rules of a game doesn't seem to be enough to learn the game—one's moves need to be moves *in the game*, in some way, if they are going to count as learning to play—but if we require on the other hand that the learner obey the rules of the game in learning to play, we end up with a vicious regress. Sellars' solution to this problem is to point out that we have tacitly accepted a false dichotomy between

> (a) *merely conforming to rules*; doing A in C, A′ in C′, etc., where these doings 'just happen' to contribute to the realization of a complex pattern;
>
> (b) *obeying rules*; doing A in C, A′ in C′, etc., with the intention of fulfilling the demands of an envisaged system of rules. (P.325).

Believing that these are the only two options requires us "to suppose that the only way in which a complex system of activity can be involved in the explanation of the occurrence of a particular act, is by the agent envisaging the system and intending its realization"

(p.325). But we have no reason to believe this, for "surely there can an unintended relation of an act to a system of acts, which is nevertheless a necessary relation—a relation of such a kind that it is appropriate to say that the act occurred because of the place of that kind of act in the system" (p.325). Here is an example of the middle way that Sellars is imagining:

> What would it mean to say of a bee returning from a clover field that its turnings and wigglings occur *because* they are part of a complex dance? Would this commit us to the idea that the bee *envisages* the dance and acts as it does by virtue of intending to realize the dance? If we reject this idea, must we refuse to say that the dance pattern as a whole is involved in the occurrence of each wiggle and turn? Clearly not. It is open to us to give an evolutionary account of the phenomena of the dance, and hence to interpret the statement that *this* wiggle occurred because of the complex dance to which it belongs ... in terms of the survival value to groups of bees of these forms of behaviour. (P.326).

Individual bee movements occur because of the dance pattern of which they are a part, but this is a pattern of behavior that has been selected for through evolution, not behavior that the bee itself need recognize as a system which it then intends its individual movements to realize.

The solution to our initial problem can now be seen. We do, in fact, appeal to a difference between 'conforming to' and 'obeying' the rules of language—but when we say 'conforming to' we do not mean 'merely' conforming to (where the fact that the behavior conforms to a rule is purely accidental, in one sense); rather we mean that the behavior occurs *because* of a pattern, a system of behaviors, but not in virtue of the agent *intending* for the behavior to follow the rules. Sellars here refers to this kind of rule-conforming behavior as 'pattern governed' behavior, as against 'rule obeying' behavior

(p.327). He then provides the following explanation of just what 'pattern governed' behavior is:

> To learn pattern governed behaviour is to become conditioned to arrange perceptible elements into patterns and to form these, in turn, into more complex patterns and sequences of patterns. Presumably, such learning is capable of explanation in S-R-reinforcement terms, the organism coming to respond to patterns as wholes through being (among other things) rewarded when it completes gappy instances of these patterns. Pattern governed behaviour of the kind we should call 'linguistic' involves 'positions' and 'moves' of the sort that *would be* specified by 'formation' and 'transformation' rules in its metagame if it *were* rule obeying behaviour. (p.327).

The theme of pattern governed behavior as a form of (often quite complex) behavioral conditioning[100] is going to come up repeatedly as we move forward. In fact, it would probably not be wrong to say that the distinction we have just developed between pattern governed and rule obeying (or rule governed) behavior forms the backbone of Sellars' philosophy of language and mind. Hopefully we can now see why: it offers us a way of preserving the initial, and quite plausible, suggestion that learning a language is learning the rules of that language without falling prey to the vicious regress that seems to threaten if we make that suggestion without this critical back-story.

---

[100] There is a danger that lurks when talking about behavioral conditioning. At least since the extinction of behaviorism as a philosophical theory of mind (if not during its short-lived heyday), there has been a tendency to think of behavioral conditioning as a relatively simple phenomenon—one that could not possibly account for any truly complex behaviors, and most certainly not linguistic behavior. When Sellars, and when I, then, speak of verbal behaviorism, and of pattern governed behavior as behaviorally conditioned, the danger is very real that we will be misunderstood. For my part, while I wholeheartedly reject behaviorism as a theory of mind, I think that the power of behavioral conditioning is often underappreciated by philosophers of mind. But then, there is behavioral conditioning and there is behavioral conditioning. Misleadingly simple examples of dogs salivating when bells are rung and pigeons pecking food-dispensing buttons tend to get in the way when one first mentions behavioral conditioning in conversation. (I say "misleadingly simple" because I seriously doubt the neurological processes involved in developing these conditioned responses to stimuli could reasonably be called "simple", at least for those of us who are not neuroscientists.) Quite obviously, the conditioning required for learning the language game is anything but simple.

Yet there is still more to say before we can feel that we have provided a solid foundation for verbal behaviorism. I will leave a detailed discussion of the issues that arise concerning language learning primarily for §4.22 below, but there is no way to completely extricate the concepts and distinctions that I wish to discuss further in this section from the general issue of how one learns language (one's initial language, anyway). To further explore the distinction between rule-conforming and rule-obeying linguistic behavior, as well as a couple of related concepts and distinctions, however, I wish to leave "Some Reflections on Language Games" behind for now—for I wish to speak outside the initial worry that has motivated the discussion in this section thus far. I want to say a little bit more about the difference between behavior, particularly linguistic behavior, that conforms to a pattern and is produced by stimulus-response conditioning, and linguistic behavior that is truly rule-governed. Then we will need to introduce and discuss another important distinction, one between two types of rules: what Sellars calls 'ought to be' and 'ought to do' rules of behavior.

Thus far we have three types of behavior on the table. First there is behavior that 'merely' conforms to rules. For example, the child who just happens to move chess pieces in accordance with the rules of the game, even though the child has no notion of those rules, nor even that the objects with which it is playing are pieces in a game. Next we have pattern governed behavior. This is behavior that also conforms to rules, but it does not 'just happen' to do so; rather, pattern governed behavior is rule-conforming because it is part of a system of behavior that has, in some way, been selected for. The rules of the game, that is, are responsible for the fact that the behavior conforms to them—such conformity is not mere happenstance. Nevertheless, behavior of this sort

falls short of the third sort of behavior: rule governed behavior. The difference, as we have seen, is that in cases of rule governed behavior, the conformity of the behavior in question with the rules of the game is intentional—someone 'envisions' the behavior as part of a game. With pattern governed behavior there is no such intentionality present—at least, not on the part of the behaving entity itself. Take for example the bee dances that Sellars mentioned. The individual movements of a given bee conform to the rules of the dance not by pure coincidence, but because conformity to those rules has been evolutionarily selected for. Yet this does not commit us to the idea that some intelligence envisioned the conformity of individual bee movements to the rules of a dance. And, of course, there is no suggestion that the bee itself is aware of the rules governing the dance, nor even that it is aware that it is partaking in the bee-dance 'game.' Language use, on the other hand, is supposed to be an example of truly rule governed behavior.

Now the basic thesis of verbal behaviorism is that learning one's first language is a matter of being conditioned to conform to the rules of the language by those who are already full-fledged speakers, until one comes to 'grasp' those rules oneself. VB replaces the innate grasping of contents, of the classical account of language learning, with behavioral conditioning. So we can immediately see how the distinctions that we now have in play are central to developing the VB framework, upon which Sellars builds his Psychological Nominalism. The initial goal of language teachers on the VB model is to create in the learner a form of pattern governed behavior—that is, of linguistic behavior that conforms to the rules of the language, not by mere happenstance, but because it has been conditioned to do so. But of course, while the language *learner's* behavior will be pattern governed—which means, given our definition above, and most importantly in this

case, that the learner does not intend her linguistic behavior to conform to the rules of the language, nor does she even recognize that her utterances are in fact 'moves' in the language game—the language *teacher's* linguistic behavior is rule governed, because her utterances are, if I can put it somewhat crudely and incautiously, self-consciously moves in the language game.[101] More broadly speaking, while, say, the system of bee dances never leaves the level of pattern governed behavior, language is by right a rule-governed system.

None of this would make the slightest bit of sense, though, without our initial distinction between *conforming to* and *obeying* rules. Language is a system governed by rules, but if one can only participate in that system by obeying those rules, then language use would seem never to get off the ground (not even, if Sellars' argument to this effect is sound, by appeal to some sort of pre-linguistic 'awareness' of the rules). If, on the other hand, it is possible to conform to the rules of language in such a way that we can say at once both that the rules are efficacious in bringing about said conformity, but that thus conforming does not require one to be able to express, nor even recognize, those rules, then we seem to have at least the beginnings of an account of language use and learning that can satisfy the initial thesis (quite plausible on its own) that learning to use a language is learning to follow its rules. Then given our definition of pattern governed behavior, we are already well on our way to a clear fleshing out of verbal behaviorism. Yet there is an important piece still missing.

---

[101] The explanation for how a mature language speaker's utterances can be considered rule governed is complicated, and I will have a lot more to say about it as we continue on. I therefore ask to be excused at this juncture for putting the point so crudely. And to forestall the most obvious worry about the way I have put the point here: I do not mean to suggest, in saying that a mature language speaker's utterances are 'self-consciously' moves in the language game, that the speaker thinks, with every utterance, "Now I shall make a move in the language game." Again, I will explain more carefully what this point comes to as we move on, so I ask the reader's indulgence of this seemingly misleading way of speaking here.

The missing piece is Sellars' distinction between 'ought to be' and 'ought to do' rules—particularly linguistic ought-to-be's and ought-to-do's, though the distinction isn't at root one that applies only to language. While we will be interested in this distinction primarily with respect to the rules of the language game, it may help make the idea clearer if we start with an example that does not involve language. Thus, consider the following story, adapted from an example that Sellars presents in "Language as Thought and as Communication" (1969, p.508): Imagine that we have a clock in our small town square. We want the clock to chime every hour on the hour, so that all the town can keep track of time. Thus we have the following *ought to be* rule: (A) "The town clock ought to chime every hour on the hour." Given our desires, this rule makes perfect sense. This rule, however, implies a further rule, a rule of action, which serves as an example of an *ought to do* rule: (B) "All things being equal, we ought to bring it about that the town clock does chime every hour on the hour."

Now, the first, and perhaps most important, thing to notice here is that while the ought to be rule (A) does not require of its subject (the clock) that the subject have any concepts (of clocks, chimes, towns, rules, or anything at all), the ought to do rule (B) does require of its subject (persons) that the subject have the appropriate concepts (see Sellars, 1969, p.508). Ought to be rules in general are statements about how things should be, and they do not generally require that the subjects to which they apply be capable of conceptualizing the state of affairs described in the rule, nor even that they be capable of having concepts at all (as illustrated by the inanimate clock). Ought to do rules, on the other hand, are rules of action and so apply to agents. Those agents, then, will typically need to be capable of at the very least having the concepts deployed in the

rule—and they may even need to have some general concept of a rule. In the clock case above, this is all fairly obvious. For (A) to apply to the town clock requires nothing conceptual on the part of the clock; but if (B) is to operate at all, those of us in a position to be subject to it are going to need to have a host of concepts. I would also suggest that (B) will require at least some general notion on our part of a rule, for even if we don't represent (B) to ourselves, we will need to represent something like (A) to ourselves, otherwise it will be a mystery why (B) should be the case. I do not wish to digress at this point, however, into a discussion of what is required to understand a rule. Rather, I wish simply to point out, as I have just done, this one important distinction between ought to be and ought to do rules. I think the difference and its significance is fairly straightforward in our clock case. How do ought to be and ought to do rules, and the distinction between them, apply to cases of language use, though? This is the important question if the distinction is going to be essential to the framework of VB, as I have claimed it to be.

Here is where the issue can start to get a bit murky, though, for in the case of language use, the subjects of both the ought-to-be's and ought-to-do's are going to be people[102]—hence, agents. So even if ought to be rules do not generally require that their subjects possess the concepts relevant to the rule, in the case of language use the subjects of the ought-to-be's generally will possess those concepts, or at the very least they will be capable of having the relevant concepts. While this clearly does not obliterate the difference between ought to be and ought to do rules of language, it does seem to complicate things a bit. Let's get an example on the table to give us something more

[102] I am not ruling out the possibility that the language game could be played by animals, extraterrestrials, or computers. I'm limiting the discussion here to human beings purely for simplicity's sake.

concrete to work with. Consider the following ought to be rule: (C) "People ought to respond to the presence of red objects with 'Red!'" The ought to do rule that seems to correspond with this ought-to-be is: (D) "All thing being equal, we ought to bring it about that people do respond to the presence of red objects with 'Red!'" (see Sellars, 1969, pp.511-512). Now, while the demands of (C) and (D) are clearly different (the former seems to require of us a certain kind of verbal behavior in response to a particular stimulus, the latter seems to require that we behave in unspecified ways towards others—though with a definite goal in mind), on the surface it would seem that both rules are rules of action, telling us what we should do in particular cases. Yet what is the difference between ought to be and ought to do rules if not that only the latter are rules of action? Ought-to-be's are supposed to be rules about how things should *be*, not what their subjects should *do*—after all, in the clock case it would be absurd to suppose that the rule (A) "The town clock ought to chime every hour on the hour" was a rule that demanded a certain kind of action of the clock. (This would be absurd because it is pointless to make demands of inanimate objects, for the simple reason that since they are not agents of any form, they are entirely insensitive to demands of any sort.) When the subjects of the ought to be rules are agents, though, as in the case of (C), the temptation is strong, I think, to suppose that the rules in question are somehow rules for action. At the very least it seems clear that linguistic ought to be rules like (C) do place demands upon their subjects, and that those subjects are the kinds of things which can, and are, sensitive to such demands.

Nevertheless, it is a mistake to take linguistic ought-to-be's as rules of action. Let me explain. It is true that the subjects (people) of linguistic ought to be rules can be

sensitive to the demands that the rules place on them. They *can* be sensitive to these demands, but they need not be—and in the case of language *learners* Sellars, and I, will claim that they do not, in fact, grasp the rule's demands (by which I mean they grasp neither the concepts deployed in the rule, nor the fact that they are subject to said rule at all). In order to make this case, we are going to need the concept of pattern governed behavior that we labored to make clear above. Before we get to that, however, let me belabor the current point a moment longer. Again, while I acknowledge that the subjects of linguistic ought to be rules are the sorts of entities that are in fact capable of conceptualizing both that they are subject to a given rule, and the content of that rule, still I claim it is the case that linguistic ought-to-be's are not rules of action. They tell us how things should be, not what we should do. So (C) "People ought to respond to the presence of red objects with 'Red!'" does not, in fact, tell its subjects what to do. It describes a state of affairs that ought to obtain (*ceteris paribus*, given the English language, etc.). The rule that demands action is the related ought to do rule (D) "All things being equal, we ought to bring it about that people do respond to the presence of red objects with 'Red!'", which demands of us that we take whatever actions are required to bring it about that people conform to (C), whether they recognize (C) as a rule or not. Perhaps the most important thing to keep in mind, then, is that while the subjects of (D) most definitely will need to possess quite a few concepts (more than one might even initially think, I'd argue, but that is a issue for later), the subjects of (C) needn't have any at all. And this point is not meant to be contentious nor problematic, but rather obvious, for it is simply a description of an at root fairly simple distinction: ought-to-be's tell us how things in the world should be, i.e., what states of affairs should obtain, while ought-

to-do's tell us what actions are to be taken by the agents of their domain, usually in bringing about a conformity of the states of affairs of the world with a related ought-to-be.

With this in mind, then, let us address the relations between the pattern-governed-versus-rule-governed distinction and the ought-to-be-versus-ought-to-do distinction, specifically with respect to language. As I mentioned, neither Sellars nor I require that language learners have any concepts at all, though they will be subject to countless linguistic ought-to-be's. But to make matters more complicated, we should note that full-fledged language *speakers* are also subject to linguistic ought-to-be's—and in those cases the subjects will be required to possess the relevant concepts, and must be sensitive to the rules in a way in which language *learners* are not. In order to sort out what may be beginning to look like something of mess, let me treat the two distinctions (pattern governed versus rule governed, ought to be versus ought to do) as two distinct axes that intersect in such a way as to give us three situations to consider: (1) pattern governed behavior ruled by ought-to-be's; (2) rule governed behavior ruled by ought-to-be's; (3) rule governed behavior ruled by ought-to-do's.[103]

Case (1) is probably the simplest of the three, though for many philosophers it is also the most contentious. What (1) encompasses are those noises produced by language learners[104] that conform to the ought to be rules of the language, even though the learners themselves are not aware of the rules (and hence are not 'obeying' them); nor, obviously,

[103] There is a fourth situation that would arise should we draw out our two separate axes as a grid: pattern governed behavior ruled by ought-to-do's. This is not a real possibility, however, as I hope will become clear once we have looked at (1)-(3). Though, of course, we can see the problem right from the start: this impossible fourth situation would be the case of behavior which conforms to a pattern without the agent intending that it do so while at the same time being governed by a rule whose defining feature is that the agent possess the relevant concepts and recognizes that she is carrying out a rule of action.

[104] For at least the rest of this section we will be dealing only with the language game.

do they intend for their behavior to conform to those rules. These language learners are not yet speakers for, though their verbal tokenings conform (in many cases quite accurately and consistently, we might even suppose) with the ought to be rules of the language (e.g., (C) above), they do so only as behavior which conforms to a pattern, and not as behavior which is rule governed.

Case (2) is likely the most complicated, conceptually and argumentatively, of the three. It is also perhaps the most important distinguishing feature of Sellars' VB. What (2) encompasses are those linguistic utterances produced by mature speakers of a language that constitute the bulk of our everyday volume of linguistic behavior (that is, most sentences spoken by most people most of the time). They are utterances that *obey* (the word is important here) the ought to be rules of the language. That is, these utterances are produced not according to rules of action, but rather in accordance with rules for how things should be, but they are truly rule governed pieces of behavior, performed with the intention that they should fulfill the demands of the linguistic system's rules. Put yet another way, mature speakers of a language are aware of the fact that particular linguistic tokens are warranted in particular situations, so their linguistic behavior is rule governed, not simply pattern governed, because these speakers intend (loosely—this needn't imply that they first think to themselves, "I will now produce a linguistic token that accords with *this* rule of my language") for their linguistic behavior to realize the rules of their linguistic system; yet the rules that they intend for their behavior to obey are not rules demanding actions of speakers—they are not ought to do rules of linguistic behavior—but rules stating what should be the case with regard to utterances of the language in the relevant situation (i.e., ought to be rules of language). I

would belabor this point further, but I fear that I would only be saying the same thing in yet another way.

Finally, in case (3) we have that linguistic behavior which is definitive of language *teachers* trying to bringing learners into the fold. Case (3) thus intersects with case (1). Where the learners in (1) are conforming to ought to be rules of language in a pattern governed way, the teachers in (3) are explicitly following the corresponding ought to do rules of language intended to produce, through conditioning, the pattern governed behavior of the language learners. The utterances of these teachers are rule governed linguistic tokens, as in (2), because the speakers are aware of the fact that, and intend that, their tokens realize the rules of their linguistic system. Yet unlike the tokens in (2), in this case the speakers are actually following rules of action: they are trying to bring it about that the learners conform to the ought to be rules that the teachers must therefore at least implicitly recognize as rules.

The goal of this section has been to give us a fairly clear understanding of the framework of Sellars' verbal behaviorism. I have not tried to give an exhaustive argument for this position, but rather to describe its most important features, for these features are going to continue to come up again and again throughout the rest of this project. Not only will the distinctions covered here come up over and over again, though, they also form the essential core of Psychological Nominalism, a view which is very close to my own Linguistic Theory of Thinking. Hence our need to be clear about them. Verbal behaviorism is, as I keep saying, a framework. What we must turn to now is the fleshing out of that frame into something truly substantive. We must delve more deeply into some of the issues already raised (for example, I will be discussing, in much greater

detail, the issues surrounding language learning in the next section), and we must then begin the task of tying together the various topics of the preceding chapters, as well as this one, to form a coherent whole. For now, however, I believe that we have a good foundation upon which to continue our work.

4.22 Learning Language

In the previous section we got a brief glimpse of Sellars' theory of language learning, of how those human beings[105] who have not yet learned any language at all are brought by their teachers into the linguistic fold. The overall purpose of this section is to develop the details of that theory, discussing the issues that arise as the picture comes into sharper focus. Let me note quickly up front, though, that when we talk about 'language learning' here we are *not* talking about the learning of a second language by someone who already speaks a language. We are talking about the transition from pre-linguistic human being to a person who truly speaks a language.[106] Since, ultimately, it will be one's ability to become a true speaker of a language that will enable one to become a thinker, the details of how language learning takes place and is possible at all are of great importance to both Sellars' Psychological Nominalism as well as my own Linguistic Theory of Thinking. In looking in detail at the process of language learning, and the

---

[105] This theory needn't be limited to human beings; it will apply to any entity capable of learning a language (possibly, e.g., apes, robots, aliens, etc.). But rather than constantly use bulky phrases like 'creature capable of language learning' I will simply talk about people.

[106] Two small notes are important here. First of all, I will often talk about someone learning to 'speak' a language, but I do not intend for this to mean that I am only talking about linguistic utterances produced with the larynx. That is, I am not excluding sign language, writing, or any other method of publicly tokening bits of language. I use the word 'speak' as a mere convenience to cover all modes of public linguistic tokening. Secondly, while I will be talking about the learning of one's *first* language, as I say here, I am not supposing that one cannot learn more than one language at a time. This may seem a minor point, but I feel I should at least mention that the process of language learning detailed herein applies, with probably only minor differences, to a child (say) who is trained up in a monolingual manner and one who is trained up as multilingual.

issues that arise from the kind of picture that Sellars' espouses, we will be building upon the VB framework established in the previous section.

So what was the basic picture of language learning that began to emerge in the previous section? The simple story goes something like this. A pre-linguistic human being (let's just say a child) goes through a period of behavioral conditioning in which its teachers, who are already mature speakers of the language, attempt to bring the child's linguistic behavior in line with the rules of the language. In the terms deployed earlier, the teachers obey ought to do rules of language to bring it about that the learner conforms to ought to be rules of language. At some point, and somehow (issues to be addressed below), the learner comes to 'grasp' the ought to be and ought to do rules itself, and its linguistic behavior becomes true speech. To introduce a new way of conceptualizing what's going on: the pre-linguistic child begins by making noises that are mere *parrotings* of speech, and ends by producing utterances that count as truly speaking the language. This picture of language learning is really part-and-parcel with verbal behaviorism.

There are some difficult issues that arise, though, when we begin to really take seriously this theory of language learning. Let's start with perhaps the most difficult question: Exactly how does one go from being a behaviorally conditioned being whose linguistic behavior is nothing more than mere parroting, to being a full-fledged speaker of the language, whose utterances are truly meaningful? Obviously there is a huge difference between someone who merely produces noises in echo of those noises produced around him, like a parrot, and someone whose utterances are meaningful tokens of a language. An examination of these differences incorporating elements of our

discussion of VB above will help us not only see more clearly how the transition from language learner to language speaker might possibly take place, it will also more clearly develop certain features of VB. That said, I feel obliged to note up front that I do not believe I yet have a perfectly satisfactory answer to the question of how the transition from language learner to speaker takes place. I have some suggestions for how to go about addressing this issue, and I think that I can point in what seems, to me at least, to be the right direction, but I will not here be able to lay this issue to rest. Nevertheless, let's get started.

First of all I think that we should recall a central part of our discussion of meaning in Chapter Three, because perhaps the most significant difference between the parrotings of the language learner and the speech of the full-fledged language user is that the latter's utterances are meaningful tokens of his language while the former's tokenings are merely noises, and only by courtesy considered words (in anticipation of the learner becoming a speaking member of the linguistic community at some later date). Hence, the question of what makes for a meaningful linguistic token is central to the issue at hand. Briefly, then, let me repeat a little of what I said in Chapter Three. What makes a linguistic token meaningful? We saw in Section 3.2 that, on Sellars' 'functional classification' view, meaning is a way of classifying a linguistic token in terms of the function that it plays in the language of which it is a part. That is, to give the meaning of a linguistic token is to specify the role that it plays in the language game. Sellars writes that this

> functional classification involves a special (illustrating) use of expressions with which the addressee is presumed to be familiar, i.e. which are, so to speak, in his background language. Some of the functions with respect to which utterances are classified are purely intra-linguistic (syntactical), and, in simple cases, are correlated with formation and transformation

> rules as described in classical logical syntax. Others concern language as a response to sensory stimulation by environmental objects—thus, candidly saying, or having the short term propensity to say, 'Here is a penny', or 'This table is red'. Still others concern the connection of practical thinking with behavior. (1974, p.421).

So far, however, all we've said is what, on Sellars' view, it is for a linguistic token to 'have a meaning'; we have not said how a speaker is supposed to go about producing meaningful tokens. In fact, with what we have said so far, the tokens of the language learner will be just as meaningful as the mature speaker's, insofar as they are both tokens in the language—but the claim we're trying to make sense of in this section is that the linguistic tokens of a pre-linguistic child are not meaningfully produced; they are mere parroting.

We are playing with a distinction here, of course, between talking about meaning in an abstract way (i.e., talking about the meanings of linguistic types and tokens independently of the context of their production), and talking about the meaningful production of individual linguistic tokens. Even if we grant, then, that abstractly considered, to talk of the meaning of linguistic tokens is to classify them functionally, we may still suppose that in order to produce those tokens in a meaningful way a speaker must, as it were, 'know the meaning' of what she says. Supposing this, however, would be a mistake on Sellars' view. Sellars claims, of course that "there is all the difference in the world between parroting words" and producing truly meaningful linguistic tokens, but, he writes, the difference

> is not that the latter involves a non-linguistic 'knowing the meaning' of what one utters. It is rather that the utterances one makes cohere with each other and with the context in which they occur in a way which is absent in mere parroting. Furthermore, the relevant sense of 'knowing the meaning of words' ... must be carefully distinguished from knowing the

> meaning of words in the sense of being able to talk about them as a lexicographer might—thus, defining them. Mastery of the language involves the latter as well as the former ability. Indeed they are *both* forms of *know how*, but at different levels—one at the 'object language' level, the other at the 'meta-language' level. (Pp.429-430).

The difference between mere parroting and truly speaking turns, according to Sellars, on the level of mastery that a language learner has gained over her own pattern-governed linguistic behavior (recall case (2) of our intersection of the two central distinctions of §4.21 above). When a child is first learning language she is merely conforming to the ought to be rules of linguistic behavior, but she is conforming to the ought to be rules without intending to do so, or even recognizing that her noises are part of the language game.

Once the child has got the hang of the language, though—perhaps once her patterned behavior reaches a certain level of complexity, say—her utterances become truly meaningful; she now not only conforms to the ought to be rules but does so with awareness (to some extent) that such rules are operative. She must also have some mastery of the ought to do rules of her language as well. Mastery of the ought to do rules is important not primarily so that the new full-fledged speaker can train *other* language learners, but so that she can now train (i.e., correct) *herself.*

This was about as far as we got in Chapter Three, though we lacked the fleshed out understanding of the distinctions of §4.21. With §4.21 under our belts now, though, we are in a position to say more about the issue here than we were able, or intended, to in Chapter Three. In the language of this chapter, what we are endeavoring to explain is the process by which one goes from conforming to ought to be rules of language without intending to do so, or recognizing that one is indeed engaged in a norm-governed activity,

and conforming to said rules with awareness of the activity of language use as subject to rules. We have suggested that the learner's behavior is simply conditioned, pattern-governed behavior, while the speaker's is intentional and rule-governed. But, again, how (and when) does one go from being a mere parrot to being *aware* (in some sense) of the norms of language use?

One possible direction someone could pursue is the suggestion that becoming a speaker involves crossing a threshold of some kind—say, reaching a specifiable level of complexity in one's linguistic behavior. This approach would envision becoming a speaker as something like turning on a light: there will be a moment when 'the switch is flipped' and the child's linguistic behavior goes instantly from mere pattern governed parroting to rule governed speech. Given the above framework of speaking as rule governed behavior performed with the intention of realizing the relevant system of rules (i.e., playing the language game), the suggestion here would be that there is a sharp line between pattern governed conformity to the ought to be rules of the language game and rule governed conformity to those rules—a line that one crosses, presumably, only when one meets certain criteria (which it would then be incumbent upon us to specify). The criteria might include one or more of such requirements as attaining a critical mass of time spent parroting the language in question, reaching a given level of complexity in one's pattern governed linguistic behavior, attaining a specified degree of reliability in producing the 'appropriate' linguistic tokens, or any number of other conditions. I imagine that there are numerous criteria, and combinations thereof, that we might come up with.

On the other hand, we might suppose that the process is more gradual than this. Perhaps there is a gray area between a being whose linguistic behavior is entirely pattern governed, and a mature speaker whose linguistic behavior is entirely rule governed. This is not to say that there won't be criteria (perhaps even the same criteria mentioned above) that must be met in order to transition from learner to speaker, but we might imagine a certain amount of flexibility in the criteria, so that, e.g., one's linguistic behavior becomes less pattern- and more rule-governed as one spends more time producing linguistic tokens, as one's tokens becomes more reliable, etc., until the behavior is, say, partially rule governed, though still partially mere parroting. To give ourselves an analogy, we might think of a person learning to play chess without being told the rules beforehand. He begins by mimicking the moves of the other player, being corrected when his moves violate the rules, but slowly he begins to actually 'grasp' some of the rules. Perhaps, e.g., he figures out that pawns can only move forward and can only attack diagonally. Yet much of his behavior is still just mimicry; he hasn't worked out all the rules yet. Of course, the chess learner is consciously trying to discern the rules of the game in a way that the language learner is not, so the analogy isn't perfect. Nevertheless, it gives us something of an idea of how a gradual transition might work.

There is an issue, though, that we haven't brought up explicitly yet, but which is revealed by the analogy just made with chess. This is the issue of the role of thought in the process of language learning. The person learning to play chess is a thinking being, and so figuring out the rules of the game, even if one does so by first engaging in merely

pattern governed behavior[107], involves a lot of reasoning. In the case of the language learner, though, do we want to say the same thing? Should we suppose that the language learner is aware of the learning process, and is actively engaged in deducing the rules of the language game? Obviously to suppose this would be to abandon the verbal behaviorist framework entirely, for that framework was recommended in part so that we could avoid attributing to language learners a pre-linguistic awareness of logical space (recall Sellars' rejection of Metaphysicus' view). We haven't really discussed this part of the VB framework, though, so perhaps we should just note here that if we were to suppose that the language learner was engaging in a reasoning process, the goal of which was to determine the rules of the language game, we could not longer characterize that behavior as pattern governed, since by definition pattern governed behavior conforms to ought to be rules without the behaving being intending that the behavior so conform, or even recognizing that there is a game afoot.

Yet the issue of how thinking figures into language learning is an important one, and we cannot simply brush it aside (as I think is fairly clear once we start to ask—as we have in this section—how the transition from language learner to full-fledged speaker actually takes place). I do not wish to address fully, in this section, the question of the role of thought in language learning, however, for I think that we need to add some further elements to our discussion before we'll be able to develop a satisfactory answer. I will add those elements in the next, and final, section of this chapter. We can, though, lay a bit more of the foundation for that final stage. So let me say a few more things about the issue of language learning here. Some of what I will have to say will flirt with the

[107] We should, of course, note here that the behavior in the chess case isn't *quite* pattern governed behavior in the sense defined earlier, since the chess learner at least recognizes that he is playing a game.

issue of thought's relationship to language learning, some of it will simply be a further discussion of the differences between the threshold and graduated views of language learning mentioned earlier.

In anticipation of §4.23 below, let me discard here a couple of dead-ends when it comes to discussing the role of thought in language learning. First of all, I want to be entirely clear that I do not take consciousness to be the issue. Consciousness, of some sort, will have a role to play, but I actually think that 'consciousness' doesn't denote a single phenomenon. I'll have more to say about that issue in Chapter Six, but for our present purposes the red herring I'm trying to avoid here is the supposition that what separates the language learner from the true speaker is that the speaker, but not the learner, is *conscious*. The suggestion would be that the behaviorally conditioned, pattern governed language learner is merely some sort of automata, but that when consciousness awakens in the being in question it becomes capable of truly speaking. I'm not supposing that anyone has ever suggested that this is how language learning (especially on a VB model) actually takes place; I am simply trying to forestall what might seem like an obvious objection: the absurdity (and I do take it to be absurd) of suggesting that language learners are unconscious, and only become speakers when they suddenly (or gradually) gain consciousness. There is *something* to this suggestion, a grain of truth that I will develop in Chapter Six, but in the way that the suggestion is stated here it is, in my view, entirely mistaken. Language learners are, on my view, quite clearly conscious. In fact, language learning couldn't take place if the learners weren't conscious. I must advise the reader to be cautious here, though, since when I say that language learners are conscious I do *not* mean that they have thoughts. I take being conscious to be quite

different from having thoughts. At any rate, I want to be clear up front that 'being conscious' is not, on my view, what separates the language learner from the full-fledged speaker, so the sudden (or gradual) appearance of consciousness will not explain the transition that we're interested in.

Secondly, as I noted briefly, almost parenthetically, above, the transition from learner to speaker cannot turn on some sort of non-linguistic awareness of the rules of the language. We could argue that such supposed awareness would itself be problematic (this is the thrust of Sellars' argument, discussed in §4.21 above, that the view of Metaphysicus is subject to the same sort of *reductio* as afflicts the naïve thesis of language learning with which Sellars begins "Some Reflections on Language Games"). That, however, is unnecessary—and even beside the point. The problem for our present purposes in supposing that the transition from pattern governed language learner to rule governed language speaker involves coming by a non-linguistic awareness of the rules of the language game is that such a supposition is incompatible with verbal behaviorism. As I mentioned earlier, since pattern governed linguistic behavior is defined as a conformity to the rules of the language game without a recognition that one's behavior does conform to those rules, or even that one's behavior is part of a game at all, the dropping in of some non-linguistic awareness of the rules to explain language learning defeats the purpose of proposing a VB framework in the first place. After all, what purpose would behaviorally conditioning language learners to conform, in a pattern governed way, to the rules of language serve if learning to actually speak required that one somehow come by a non-linguistic awareness of the language game and its rules? If we're going to embrace a picture of language learning that places the learner in the

logical space of reasons prior to becoming a speaker, then we might as well suppose the learner to start out in that space from the very beginning and do away with the apparently pointless period of behavioral conditioning all together. Again, I'll have more to say along these lines in the next section, but I want to make clear right up front that I do not think an appeal along these lines is a real possibility for explaining language learning within the VB framework.

Now, returning to the two suggestions for how language learning might take place within our present frame of discussion, whether by the crossing of a distinct threshold, or by gradual progress through a gray area, the only question I wish to pursue further here is which of these suggestions I take to be the more promising. And I will say that I find the graduated view of language learning to be more appealing. I do not, however, have any compelling arguments for preferring the graduated to the threshold view. Nor do I even believe that the threshold view simply couldn't ever work. For all I know, once we develop a clearer empirical understanding of the process of language learning, we might very well find that there is, in fact, a distinct point at which one crosses from being a pattern governed parrot to being a rule governed speaker. My intuition, though, is that this won't be the case. Why do I have this intuition? There are a number of reasons, some more philosophically interesting than others. Probably the strongest reason I have for thinking that the graduated view of language learning is the right way to go is that it seems to fit more smoothly with the behaviorist elements of the framework we're working in. Consider: the core of VB is the idea that one begins to 'use' language in a mimicking way, and that one's trainers reinforce proper deployment of linguistic tokens. This is a process that progresses by degrees. The infant babbles non-sensically; then she

begins to produce what, by courtesy, we call words. At first her production of these words fails to match up with the context in any real way; over time, though, as we selectively reinforce the production of certain tokens in certain situations, and other tokens in other situations, her words come to more reliably match up with the context of their tokening. Eventually the child seems to be 'getting the hang of it.' Now, if this process advances by degrees in this way, it seems reasonable to me to suppose that the transition from the behavior of the learner to that of the speaker also proceeds by degree. Again, though, this isn't an argument for the graduated view so much as an explanation of an intuition.

As I said near the beginning of this section, I do not feel that I am in a position yet to propose a definitive answer to the question of how the transition from learner to speaker takes place. What I have tried to do here is lay out the issue as clearly as possible, discuss some of the worries that arise when looking at language learning on a VB account, and say something about the direction in which I think the conversation will need to continue (both in the remainder of this project, to the extent that I've here indicated is possible, as well as in research that goes beyond what I hope to accomplish at this point). We've also seen that the discussion of language learning within the VB framework pushes us into a consideration of the relationship between thought and language—which is, of course, the primary topic of this dissertation. In the next section, then, I will take what we have so far of Sellars' VB framework, and what we were able to say here about language learning, and fill in what we need of the remainder of the details of Sellars' Psychological Nominalism. There we will see how Sellars envisions the

relationship between thought and language—a picture that I will then adapt in Chapter Five to form the heart of my own Linguistic Theory of Thought.

4.23 Language and Thought: The Myth of Jones and Our Rylean Ancestors.

That the issue of language learning, on the VB model, is inextricably tied up with our theory of mind seems to me beyond doubt. Indeed, even before we looked specifically at the issue of language learning, the necessity for introducing a theory of mind was at the very least implicit in our discussion of verbal behaviorism. Recall that what we claimed separated pattern governed from rule governed behavior was an *intention* on the part of the agent to realize the system of rules that constitute the game (whatever it might be). That is, what separates pattern governed from rule governed behavior is that the latter, but not the former, requires thought. When we then come to the question of how one makes the transition from language learner to language speaker, this distinction becomes central to the whole issue.

Now the view that Sellars calls Psychological Nominalism (PN) amounts to the claim that there is no thought, no conceptual awareness, no 'awareness of logical space,' either prior to or independent of the acquisition of a language (1997, p.66). Whether this claim adds anything to what we've been calling verbal behaviorism or not is open to question, once we see the implications of adopting the VB framework.[108] Regardless, the full-blown realization of Psychological Nominalism is of primary interest to us now. The

---

[108] Indeed, in "Meaning as Functional Classification" Sellars defines Verbal Behaviorism (his capitalization) in more-or-less the same way as he defines Psychological Nominalism in EPM. I have given VB a more narrow construal up to this point in order to keep the discussion focused and simpler, but I have no vested interest in arguing for a distinction between VB and PN if Sellars indeed considers them basically the same thesis. I mention this here to acknowledge that the use I have been making of the phrase 'verbal behaviorism' might not be the use that Sellars makes of it, but I do not think anything important hangs on this point.

purpose of this section is to bring together the various pieces of our discussion thus far and lay out Sellars' picture of the relationship between language and thought.

To do this I will be revisiting some of the topics and issues introduced earlier, but I will also be introducing a number of new topics for consideration. I am going to center a large part of my discussion here around Sellars' Myth of Jones from "Empiricism and the Philosophy of Mind." My ultimate purpose is to present Sellars' picture of a new paradigm in psychology and philosophy of mind. Let us begin, though, by briefly revisiting the discussion of language learning to see if, and if so, how, the introduction of the primary claim of Psychological Nominalism alters the picture.

As already noted, it may be that the primary thesis of Psychological Nominalism only makes explicit what was already implicit in our discussion of verbal behaviorism. If that is the case, then we shouldn't expect for the PN picture to really change our discussion of language learning. We can, however, revisit the denial, made briefly above, of the idea that learning a language involves a non-linguistic awareness of the rules of the language. I said when we first looked at this suggestion that to suppose that language learning involved such an awareness would be to abandon the verbal behaviorist framework. To make this point I noted that if there were, in the language learner, a non-linguistic awareness of the rules of the language game, or of the fact that she was trying to learn the language game, her behavior would no longer be pattern governed in the sense defined by VB. Pattern governed behavior on the VB model is behavior that conforms to ought to be rules without any intention on the part of organism that the behavior realize the system of rules in question. But if the transition from language learner to full-fledged speaker were to require that the child develop a non-linguistic

grasp of the rules of the language game, then insofar as her behavior was pattern governed these conditioned responses would serve no purpose in allowing the child to actually become a speaker, while to the extent that the language learner had a non-linguistic grasp of the rules of the language game, or of the fact that she was playing the language game, her behavior wouldn't be pattern governed. All of this follows from our definitions of pattern governed and rule governed behavior above.

Once we realize this, however, it should become clear that a primary implication of verbal behaviorism, if not one of its ultimate purposes, is the denial that there is any non-linguistic awareness (i.e., non-linguistic conceptual grasping) of the system of rules that make up a language, and hence that learning a language is, while a matter of learning to conform to its rules, not a matter of first 'grasping' those rules non-linguistically. Yet if we're going to deny non-linguistic awareness as a route to learning the rules of one's first language, we are quite likely going to want to also deny that there is any such thing as non-linguistic awareness *of anything at all*. Why? Because to grant a pre-linguistic child, say, an awareness of the space of reason, but deny that she goes about learning language by deducing the rules of the language—i.e., by grasping, non-linguistically, the rules of the language—is to endorse a highly implausible state of affairs. To be in the space of reason is to intend (in some sense) that one's behavior conform to the rules that govern that space, and I have a hard time seeing what sense it would make to suppose that the language learner has an awareness of the rules of the game of reasoning but lacks any notion that her linguistic behavior is also subject to rules. What we're imagining is a child who has conceptual mental states and a conceptually rich awareness of her surroundings but who produces linguistic tokens mindlessly, as it were, in a merely

parroting manner—and that strikes me as a highly unlikely, if not down right silly, thing to suppose is the case.

At any rate, what we are now grafting onto the VB framework (or perhaps just making explicit as a consequence of that framework) is the rather radical suggestion that awareness of logical space, i.e., the ability to have conceptual mental states in the first place, rests upon the acquisition of a language. This is the defining thesis of Psychological Nominalism. So the full Sellarsian story of the language game, and of language learning, incorporates a fairly radical shift away from the traditional view of the relationship between language and thought: it ties thinking to language use in a way that most theories of mind past and present would find objectionable, if not downright absurd.

Yet if I have so far given the impression that Psychological Nominalism is a fairly straightforward theory I can almost hear the incredulity of my opponents as they muster all manner of taxing questions. Thus I must hasten to acknowledge that Psychological Nominalism is anything but simple and straightforward. Indeed, the subtleties and apparent contradictions in the view form a veritable mine-field of issues, so that I could not hope to disarm them all here. I do want to address some of these issues, though, in the space that I have left in this chapter—and I will address a few more in Chapter Five. To set up our discussion I want to introduce Sellars' Myth of Jones, from EPM. I believe that this simple story will provide us with a vivid thought experiment with which to approach a couple of the more potentially troublesome aspects of Psychological Nominalism. Let us indulge for a moment, then, in a quick description of the details of this Myth.

The Myth of Jones is told in sections XII-XVI of "Empiricism and the Philosophy of Mind," and is a story of how some fictional humans (our 'Rylean ancestors') might have developed the notion of certain private inner episodes called 'thoughts.' Sellars begins the story by having us imagine a time in pre-history in which there are a people who speak a limited 'Rylean language.' This language contains only the vocabulary to talk about publicly observable phenomena in space and time. Even though the vocabulary of this language is thus quite limited, Sellars quickly adds that it also has great descriptive power—for it makes use of various logical operations like conjunction, disjunction, quantification, and counterfactual conditionals. Now the first question that Sellars wants us to consider is this: What would have to be added to this Rylean language in order that our fictional ancestors "might come to recognize each other and themselves as animals that *think*...?" (1997, p.92). According to Sellars, we need two things.

The first addition that one would have to make to this Rylean language is to enrich it with "the fundamental resources of semantical discourse" (p.92). That is, we would have to allow the Ryleans to say of each other's "verbal productions that they *mean* thus and so, that they say *that* such and such, that they are true, false, etc." (p.92). Having added semantical discourse to the Rylean language, we have given it a dimension that Sellars thinks makes much more plausible the idea that our fictional ancestors could eventually come to talk about 'thoughts' just as we do. For "characteristic of thoughts is their *intentionality*, *reference*, or *aboutness*, and it is clear that semantical talk about the meaning or reference of verbal expressions has the same structure as mentalistic discourse concerning what thoughts are about" (p.93).

The second element that Sellars says must be added to a Rylean language to allow talk of 'thoughts' to develop is the notion of *theoretical discourse*. He writes:

> Informally, to construct a theory is, in its most developed or sophisticated form, to postulate a domain of entities which behave in certain ways set down by the fundamental principles of the theory, and to correlate—perhaps, in a certain sense to identify—complexes of these theoretical entities with certain non-theoretical objects or situations; that is to say, with objects or situations which are either matters of observable fact or, in principle at least, describable in observational terms. (1997, pp.94-95).

The most important thing about theoretical discourse (for the purposes of the Myth) is that it allows talk of unobservable 'entities' into the language of our fictional ancestors, by allowing them to posit these entities on the *model* of observable phenomena.

So we now have a Rylean language enhanced by both semantical discourse (allowing talk of meaning and reference) and theoretical discourse (allowing talk of theoretical entities correlated with observable fact). With the language of our Ryleans thus enhanced all Sellars' story needs now is for one of our fictional Rylean ancestors to make the necessary theoretic leap, and begin the talk of 'thoughts.' This is where Jones comes in.

Sellars wants us to suppose that, in his fictional Rylean society, a genius—we'll call him Jones—now appears. Jones, we might suppose,

> in the attempt to account for the fact that his fellow men behave intelligently not only when their conduct is threaded on a string of overt verbal episodes—that is to say, as *we* would put it, when they "think out loud"—but also when no detectable verbal output is present ... develops a *theory* according to which overt utterances are but the culmination of a process which begins with certain inner episodes. *And let us suppose that his model for these episodes* which initiate the events which culminate in overt verbal behavior *is that of overt verbal behavior itself. In other words, using the language of the model, the theory is to the effect that overt verbal behavior is the culmination of a process which begins with "inner speech."* (1997, pp.102-103).

If we then suppose that Jones comes to call the theoretical 'inner episodes' introduced by his theory 'thoughts' we have a framework of *thoughts* as inner episodes which can accommodate much of the classical notion of mind. But, it is of the utmost importance to realize that the 'unobserved,' 'non-empirical,' 'inner' episodes of the Jonesean theory "are 'nonempirical' in the simple sense that they are *theoretical*—not definable in observational terms" (p.104). And thus, for Sellars' purposes, it is important to notice that, although we can now say that all meaningful speech is the overt culmination of 'inner episodes' called 'thoughts,' this

> is perfectly compatible with the idea that the ability to have thoughts is acquired in the process of acquiring overt speech and that only after overt speech is well established, can "inner speech" occur without its overt culmination. (1997, p.105).

This, then, is the Myth of Jones. Again, my primary interest in the Myth is as a thought experiment meant to illustrate how it might work to conceive of thoughts as theoretical entities—in particular, how we could conceptualize thought episodes as having many of the features ascribed to them on more traditional theories of mind (as 'private, inner episodes' and as preceding meaningful overt linguistic tokens), while at the same time embracing the primary thesis of Psychological Nominalism that says there is no awareness of logical space prior to or independent of the acquisition of a language. (As we'll acknowledge in just a moment, the Myth of Jones is really neutral on the question of whether the central thesis of PN is true or not, but that means that it is compatible with PN, and that's all I want.)

The Myth of Jones is, I think, a nice, clear story that illustrates rather vividly how one might use the notion of thoughts as theoretical entities to hold onto certain features of

the Cartesian conception of thoughts while jettisoning Cartesian metaphysics. (I take it that this is the primary purpose of the Myth as far as Sellars is concerned.) Once we spend just a little time reflecting upon the Myth of Jones for our own purposes, though (as a story that helps to illustrate the new paradigm of the relationship between thought and language that arises from the VB framework), we find that there are some rather serious questions that begin to arise. I wish to deal with one such question—an apparent difficulty that, if left unaddressed, would threaten to undermine the entire foundation of the LTT (as well as Sellars' own Psychological Nominalism, for that matter).

The simplest way to introduce the worry I have in mind is probably like this. According to the Myth of Jones, while thoughts are posited as theoretical entities to explain behavior not accompanied by overt linguistic tokens when such tokens would be otherwise expected, the thoughts themselves, even in speakers lacking the concept of thought, have always been there (i.e., the Ryleans were thinking before Jones came up with his theory; they just didn't *know* that they were thinking). Thus, in the Myth, it is acknowledged that these posited entities called 'thoughts' are episodes, tokens of which occur prior to any given production of meaningful, overt linguistic tokens. The problem is that, even if the Myth casts aside much of the unwanted Cartesian metaphysics of thoughts, the claim that meaningful linguistic tokens are the overt culmination of private, inner episodes is really the heart of the traditional view of the relationship between thought and language, so it might appear that the picture in the Myth of Jones actually conflicts in some way with the central thesis of Psychological Nominalism. The appearance of conflict here, however, as we'll see below, is misleading and rests on a confusion that can soon be dispelled. Of course, it may also seem mistaken in the first

place to suggest that the Myth of Jones *directly* conflicts with the thesis of Psychological Nominalism, since it could rightly be pointed out that Sellars' anti-Cartesian purpose in telling the Myth is somewhat independent of his endorsement of Psychological Nominalism. Nevertheless, the two aren't completely separate. The Myth may be just a thought experiment, but I think that Sellars does, in fact, endorse the idea that meaningful linguistic tokens are always backed by thought episodes.

There is a minor point that bears mentioning here before we turn to the major issue raised by Sellars' seemingly contradictory endorsement of both Psychological Nominalism and the picture in the Myth of thought episodes preceding meaningful speech. This minor issue has to do with the suggestion, oft made by fans of the traditional account of the relationship between thought and language, that words *express* thoughts. In one sense this claim is harmless; in another it is incompatible with Psychological Nominalism. The harmless sense of the claim that words express thoughts is present in the Myth of Jones, and is nothing more than the claim that meaningful linguistic tokens are the overt culmination of inner episodes known as 'thoughts' (though I realize we have yet to say exactly *why* this ultimately turns out to be harmless). The sense in which the claim that words express thoughts is incompatible with Psychological Nominalism is the sense in which it would be proper to say that someone *meant* by his words to express such-and-such a thought. As Sellars writes in "Language as Thought and as Communication":

> The familiar saw that words have meaning only because people mean things by them is harmless if it tells us that words have no meaning in abstraction from their involvement in the verbal behavior of language users. It is downright mistaken if it tells us that for an expression to have a certain sense or reference is for it to be *used* by people *to convey* the

> corresponding thought.[109] Rather, we should say, it is because the expression has a certain meaning that it can be effectively used to convey the corresponding thought. (1969, p.523).

The problem isn't with the idea that people sometimes do explicitly choose certain words to express their thoughts—there is no doubt that this is sometimes the case. The problem is in supposing that this is *always* the case with meaningful, overt linguistic tokenings. For on Sellars' view, much meaningful linguistic behavior is properly characterized not as the purposeful expression of thought but as thought itself (what he calls 'thinking-out-loud'). Again, though, this isn't to deny that meaningful, overt linguistic tokenings are the 'culmination of a process that begins with inner speech' as Jones' theory would have it.

The more pressing problem isn't the distinction between thinking-out-loud (about which, more below) and purposefully expressing one's thoughts, it's the question of how to reconcile the idea that meaningful, overt linguistic tokens are always preceded by thought episodes with the idea that being able to have thought episodes at all requires acquisition of a language.

Let me digress for a bit to explore what seems to be a similar problem. (This digression should put us in a better position to find the solution to the preceding problem.) Sellars has said in at least a couple of places[110] that while language is 'prior in the order of conceiving' to thought, thought is 'prior in the order of being' to language. That is, that while the concepts and categories of thought (e.g., meaning) are in the first instance *linguistic* concepts and categories, thoughts are actually ontologically prior to overt linguistic tokens. To help explain this idea, Sellars has used an analogy between

---

[109] Recall our rejection of the 'communication-intention' theory of meaning in Chapter Three.
[110] In *Science and Metaphysics*, 1967, p.164; and "Reply to Marras," 1973, p.488.

micro-physical particles (henceforth simply 'particles') and macro-physical objects (henceforth simply 'objects'). Presumably the analogy should go something like this: just as there can be no objects without particles, so too there can be no meaningful linguistic tokens without thoughts; on the other hand, just as there *can* be particles without objects, so too there can be thoughts without language. But, of course, construing the analogy this way conflicts with Psychological Nominalism. So let's look more closely at exactly how the analogy with particles and objects is supposed to shed light on the relationship between thought and language.

Let's start with the particles and objects. Given our current understanding of molecules, atoms, etc., it seems to be the case that without particles, there would literally be no objects. Since every object is, in a loose sense, a 'collection' of particles, where there are no particles, there are no objects. But, of course, the converse doesn't hold: without objects there can still be particles. So in that sense, the objects are ontologically dependent upon the particles.

Can we say the same thing when it comes to thought and language? Can we say that without any thoughts there wouldn't be any language? The answer is, unfortunately, both yes and no. There *is* a sense in which Sellars would be willing to say that without thought there is no language. In a letter to David Rosenthal, Sellars writes: "From the truth of Jones' theory one can ... [conclude] that if *on a particular occasion* there are no thoughts, then on that occasion there is no meaningful (nonparroting) speech" (Sellars, 1972, p.501). So, in one way, without thought there is no language. The analogy seems to break down when we go the other direction, though, since we can easily imagine language existing without thought. A computer which prints out the Complete Works of

Shakespeare produces a lot of linguistic tokens, but the computer is not thinking. On the other hand, it is quite easy to imagine cases in which linguistic tokens are produced in the complete absence of thought. A computer which prints out the Complete Works of Shakespeare produces a lot of linguistic tokens, but the computer is not thinking.

The question that we need to ask here is what, exactly, the analogy with micro-physical particles is supposed to help explain? As I understand it, the analogy is meant to explain two things: (1) the fact (on Sellars' view) that all meaningfully produced (i.e., nonparroting) linguistic behavior is, as Sellars puts it in the Myth of Jones, "*the culmination of a process that begins with 'inner speech*'" (1997, p.103); (2) the fact (again, on Sellars' view) that semantical terms are first applied to overt linguistic behavior and only later used to describe 'thoughts.' The analogy works like this. Just as an object depends (for its very existence at each moment) upon the particles that constitute it, so too the meaningfulness of a linguistic episode depends (for its very meaningfulness[111]) upon the thoughts that give rise to it. Similarly, just as the concepts by which we describe the particles of physical theory were first used to describe macro-physical objects, so too the concepts by which we describe thoughts were first used to describe overt linguistic behavior. Thoughts are thus 'prior in the order of being' while language is 'prior in the order of conceiving.'

So far all the analogy with particles and objects does is put in general terms a particular feature of the Myth of Jones: the fact that the pre-Jones Ryleans of the Myth think, but do not *know* that they think. What Sellars wants us to see as possible (at least in principle) is the idea that a people could use the language of semantical discourse

---

[111] 'Meaningful' in the sense that it is produced in a nonparroting way.

without having any notion of *thoughts*. But the supposition (indeed the *crucial* supposition for Jones' theory to be at all correct) is that such people are still *having* thought episodes. They just do not have the conceptual scheme yet available to be aware of those episodes.

Calling thoughts 'prior in the order of being' but language 'prior in the order of conceiving' is merely to generalize this feature of the Myth of Jones. All of the meaningful linguistic tokens of the Ryleans are, according to Jones' theory, the overt culmination of processes that begin with 'inner speech,' but the pre-Jones Ryleans have not yet conceived of anything like an 'inner episode.' It is only after Jones introduces his theoretical explanation of his fellow Ryleans' intelligent behavior that they can come to recognize what was there all along (according to the theory): their *thoughts*.

Though Sellars' Psychological Nominalism commits him to the idea that one only gains the ability to have thoughts after one acquires a language, what we must notice here is that there is nothing in the idea that thought is ontologically prior to language, as we have explored it thus far, that either supports or defeats this idea. All that has been established up to this point is that meaningful linguistic tokens cannot be produced in the absence of thought episodes (an idea that we saw in the Myth of Jones).

Now the worry about the ontological priority of thought conflicting with Psychological Nominalism arose from the suggestion that the ontological priority of thought implies, as the analogy with micro-physical particles seemed to suggest, that there can be thought episodes in a world devoid of language. This supposed implication is obviously incompatible with the claim that the ability to think is acquired only as one

learns a language, and we have not yet dispelled this worry, for we have not yet dealt with this level of the analogy.

What I will claim here, then, is that it is a mistake to suppose that the analogy between thoughts and micro-physical particles (in terms of their ontological priority) implies that thoughts could exist in a world devoid of language (or, more to the point, in a being that had not yet acquired a language). The reason that this is a mistake is because it construes the analogy too closely, ignoring an important point of *dis*analogy. It makes sense to suppose that there could be a world of particles that never collected to form an object, because we think of micro-physical particles as, in some sense, having their properties intrinsically. That is, we suppose that the properties of physical theory that we ascribed first to macro-physical objects, and only later to the particles of molecular, atomic, and sub-atomic theory (e.g., mass) are actually properties of those particles themselves. This is *not* the case when it comes to thoughts.

The concepts and categories of meaning and intentionality, which it is the point of the Myth of Jones to illustrate could have been used by a people who had no concept of anything like what we call thoughts, came (in the Myth) to be applied to the theoretical entities called 'thoughts' on the model of their application to overt linguistic behavior. But it is not a part of the Myth that these concepts be taken to provide an intrinsic, descriptive characterization of thought episodes. Rather, Jones' theory provides a metaphysically neutral, *functional* characterization of the 'inner episodes' that he comes to call 'thoughts.' That it is a mistake to take the analogy with micro-physical particles so far as to suppose that the meaningfulness and intentionality of thoughts are descriptive

features of such episodes is made clear by Sellars in, e.g., *Naturalism and Ontology* when he writes:

> The emptiness of the classical account of thought episodes can be explained by the fact that it uses as its model for the description of the *intrinsic* nature of mental acts (i.e. what they 'consist of') aspects of linguistic activity which are largely functional in character. (1979, p.71).

This, of course, should remind us of the 'functional classification' theory of meaning that we developed in Chapter Three. As I've said, all the things that we've been discussing up to this point ought now to be coming together. In this case, the level of analogy between thoughts and micro-physical particles that would suggest that just as particles could exist without objects so too thoughts could exist without language is clearly one that Sellars rejects. What Sellars' talk of the priority of thought comes to is merely an endorsement of the claim made by Jones in developing his theory, that all meaningfully produced (i.e., nonparroting) linguistic behavior is "*the culmination of a process that begins with 'inner speech'*" (1997, p.103). This claim, however (as revealed by our examination of the analogy with micro-physical particles), does not conflict in the least with the claim that the ability to think is only acquired as one learns a language—i.e., it does not conflict with Sellars' Psychological Nominalism. This ends our digression.

Let's return, then, to the primary issue for which we need to find a proper understanding. How do we reconcile the idea from the Myth of Jones that meaningful linguistic tokenings are always preceded by thought episodes with the idea that the ability to *have* thought episodes in the first place requires that one have acquired a language? For starters, our digression has now shown us just how we should understand the claim, in the Myth of Jones, that meaningful utterances are the overt 'expression' of inner

episodes. That claim is to be taken as stating merely that *meaningfully produced* linguistic tokens cannot be mere parroting: they must be preceded by thoughts. This, however, should not come as a surprise to us, for this is just what we would have expected given our discussion of the VB framework above. We noted many times there that the difference between the language learner's pattern governed linguistic behavior and the mature speaker's rule governed linguistic behavior is that the former's behavior merely conforms to the rules of the language without the learner intending to realize that system of rules, while the latter's behavior conforms to the rules of language in full recognition (in some sense) of the fact that his behavior is realizing the system of rules that make up his language. The learner is merely parroting those speakers around him, producing sounds that are only considered words in anticipation that the learner will eventually become a legitimate speaker. For someone's linguistic behavior to count as meaningfully produced (i.e., not just parroting), he must recognize that his behavior is part of a rule governed system—that is, part of a language. But such recognition obviously requires thought.

The fact that the meaningful production of linguistic tokens requires some amount of thought on the speaker's part would previously have seemed a problem (because it would appear to undermine the VB framework). But what we can now hopefully see is that this isn't a problem at all, because, as the Myth of Jones illustrates, we can imagine someone *having* thoughts without having the concept of, or any sort of idea of, private, 'inner' episodes. That is, one can be thinking without knowing that one is thinking. Indeed, I'd be willing to go so far as to suggest that all people transitioning from non-thinking language learners to thinking speakers go through a period in which they are

having thoughts but do not have any concept of, or recognition of, the fact that they are thinking (i.e., they have no second-order thoughts to the effect that they just thought thus-and-so).[112] Regardless of whether this last bit is true or not, though, the only important point we need to recognize here is that we have now fleshed out the VB framework in such a way as to both connect thinking to language use, and to allow for thoughts to occur in the absence of any concept of thought episodes.

There is one last piece of the puzzle that I need to mention before I can sketch out the full picture of thought and language that forms the core of Sellars' Psychological Nominalism. I mentioned this piece in passing earlier, but it is important enough to warrant a bit more attention. What I have in mind is the fact that, according to Sellars, not only is it not the case that all meaningful linguistic behavior is the purposeful expression of thought (as opposed to simply being the overt manifestation of thought), it is the case that, in a sense, much linguistic behavior simply *is* thought. To put the point more carefully, according to Sellars' picture, a great deal of meaningful linguistic behavior is, as it were, the overt manifestation of thought episodes and is not produced for communicative purposes. Rather, the behavior is really just part of the thinking itself. Linguistic behavior of this type is what Sellars calls 'thinking-out-loud.' Such behavior is not produced by thinking, in any ordinary sense, it *is* thinking. This isn't to say, of course, that there isn't something going on in the person that is properly identified with thought and does not involve the larynx—the occurrence of certain brain states, say. But

---

[112] These would perhaps be like Sellars' children in "Language as Thought and as Communication" who only think-out-loud, not yet having learning to 'keep their thoughts to themselves.' That is, such children are now speakers of the language, not mere parrots, and so their linguistic tokens are produced in all the rights sorts of ways, including being preceded by thought episodes, but the children literally can't think in silence—they are constant chatterboxes. (See Sellars, 1969, pp.521-522.)

it is to say that, as far as the meanings of the thoughts go, there is no separating the linguistic behavior from the thought. There are some complications to this picture, of course, and my rather rough and crude way of presenting the issue here leaves much to be desired, but I will be taking up this topic in much greater detail in the next chapter, so I do not wish to spend any more time on it here. I think that we have at least said enough now to give us a fair grasp of the full picture of Psychological Nominalism.

Let me try, then, to present that picture. First of all remember that the primary connotation of 'psychological nominalism' is the claim that there is no thought (i.e., no conceptual awareness of anything at all) without the prior acquisition of a language. We begin at the beginning, then. A child enters the world without thought (though, as I briefly mentioned earlier, with consciousness of some sort—see Chapter Six for further discussion of the topic of consciousness). That child is surrounded by thinking human beings who speak a language (at least one). As the child begins to grow, and is exposed to countless instances of linguistic stimuli, it begins to mimic the linguistic behavior of the speakers around it. In turn, those speakers begin to selectively reinforce certain linguistic behaviors of the child. The child's behavior begins to conform to the rules that govern the language of its trainers. Still the child is not thinking, but its linguistic behavior is becoming pattern governed—the behavior conforms to the ought to be rules of the language, and those rules are involved in bringing about the behavior (by virtue, in this case, of the fact that the child's teachers are conditioning it such that its behavior does so conform).

Here things get a little tricky, and as I've said before I cannot give a thorough explanation of how the transition takes place, but at some point the child begins to move

from parroting, pattern governed linguistic behavior to a legitimate rule governed speaking of the language. We might imagine that the child's utterances have come to, as Sellars says, "cohere with each other and with the context in which they occur in a way which is absent in mere parroting" (1974, p.429). Also (and I do think this is important), the child may begin to correct itself, no longer needing its trainers to reinforce the right linguistic behavior. A self-correcting system isn't necessarily thinking, but in the case of language learning this seems a crucial step—an essential part of 'getting the hang of it' if you will. As the child's linguistic behavior ceases to be mere parroting, it becomes legitimate speaking, with all that this entails on the Psychological Nominalist's view: its speakings are thinkings-out-loud; the linguistic behavior is rule governed with the new speaker intending[113], in some sense, to realize the system of rules (that it had been conditioned to conform to, of course); its linguistic tokens are backed by the occurrence of thought episodes; etc.

Let me now conclude this chapter with a couple of points that will lead us directly into the task of the next chapter. First of all, we saw in the first part of this chapter (Section 4.1) that thinking is a normative affair. We now have a theory that can accommodate this feature of thoughts, for obviously thinking, according to Psychological Nominalism, is norm-governed, since it arises from the rule-governed practice of

---

[113] This point can seem a little troublesome, since one wonders how the newly developing speaker can come from the start to *intend* anything, given that this requires that she be capable of thinking. I imagine, though, that the sense of intention necessary to allow for the child's linguistic behavior to count as rule governed on the VB model could be cashed out as some combination of factors including the aforementioned self-correcting of one's behavior, a certain degree of reliability in the conformity of the child's conditioned linguistic behavior, etc., etc. As is probably obvious, the factors that are going to allow for the 'intending' necessary for full rule-governed speech to develop are just the sorts of criteria that will likely come to bear in spelling out how the transition from language learner to speaker takes place at all. Thus, what I have to say about the development of the requisite 'intentions' must remain as necessarily vague as what (little) I have had to say about the nuts and bolts of how language learning could possibly proceed.

language use. Thus, I contend, a theory of thought like that espoused by Sellars is in a position to do what the representationalist views of Chapter Two and Section 4.1 above cannot: give us a representationalist theory of thoughts that respects the essential normativity of the activity of thinking. We should also notice that, on the picture just laid out, thinking is going to be a social affair, since thinking, on this view, requires the acquisition of a language, and the acquisition of a language requires training by those who already speak the language. That is to say, no one is going to be able to become a thinker without participating in a public practice of overtly tokening the symbols of a language. These two features, respect for the essential normativity of thought and a recognition of the social nature of thinking, are the cornerstones of the view that I am here, in this dissertation, attempting to lay out the foundation for. In the next chapter I update what I take to be the essential insights of the Sellarsian framework for incorporation into a modern representationalist theory of mind. We have all the pieces now; all that remains is to put them in place. We turn now to the final development of the theory that we have been aiming for all along, the Linguistic Theory of Thought.

5. The Linguistic Theory of Thought

The time has come to tie together all the pieces we've looked at so far. Before we begin doing that, however, I wish to briefly sketch out what the finished product will look like. That is, before we get into the details of the LTT I want us to have a general picture of the view that we're developing.

The LTT is a version of contemporary representationalism (see Chapter One). More specifically, it is a theory that takes the mental representations that constitute thoughts to be not only symbolic, but linguistic (see Chapter Two). Even more specifically, and somewhat radically, the linguistic symbols that constitute thoughts according to the LTT are symbols of natural languages. In part because of this, and in part because of how the LTT envisions language learning, the Linguistic Theory of Thought is capable of both acknowledging and explaining the essential norm-governed nature of thought. As we saw in the previous chapter, this is one of the primary things that sets the LTT apart from nearly all other contemporary representationalist views.

The normativity of thought, according to the LTT, is a consequence of the fact that thoughts are constituted by the symbols of natural languages which are in turn governed by rules. Learning language, on this view, requires learning the rules of the language (see Chapter Four). Thus, learning to think involves coming to conform to the norms of the language—and not merely conforming, but coming to *recognize*, in some sense, those norms. Of course, the meanings of the linguistic symbols that constitute both one's language and one's thoughts are a matter of the functional role played, in the language, by those symbols (see Chapter Three).

Now, since learning language—and, hence, learning to think—requires learning the rules of the language, if we apply the lessons of Chapter Four to the emerging picture we will realize that becoming a thinker is inherently a social process (and this conclusion is only reinforced by our functional role theory of meaning). There is no thought, no awareness of logical space, without plenty of practice in the public use of language; and even once thought is established, the norms that govern it come from the society of language users.

In the end, the LTT is rather simple in its broad strokes, though far from simple in its details. Thoughts, according to the LTT, are the tokening of symbols of the language(s) that one speaks—and by "speaks" I mean to invoke the distinction of Chapter Four between language learners and full-fledged speakers. These symbols are meaningful apart from any individual person, as meaning is a matter of the functional role of types of expressions within a public practice of language use. This in turn allows us to reject the traditional view of the relationship between thought and language, whereby words are simply the public expression of what would otherwise remain private inner episodes of thought. Instead, we see that the ability to think at all is a linguistic affair, and that no one can have any thoughts at all without being a trained and practiced member of a linguistic community.

We've seen the arguments for most of these details already. Below I will more carefully tie them together, adding what arguments remain to turn the Sellarsian picture into a representationalist theory for the 21st Century.

5.1 The Languages of Thought

The idea that thinking takes place via the tokening of symbols with language-like structure is not, of course, unique to the LTT. Plenty of others have embraced this idea, from Sellars, to Fodor, to psychologist Steven Pinker. In fact, it would probably not be wrong to say that such a view is the dominant theory among representationalists today. Nor does a respect for the essential normativity of thought, though it does (as I argued above) set the LTT apart from the vast majority of modern representationalist views, make the LTT unique, as a recognition of the essential norm-governed nature of thought is central to Sellars' philosophy of mind, and to the views of those who have followed him, or shared his insight. Yet almost every view of which I am aware that embraces the idea that thinking is the tokening of language-like symbols is alike is supposing that, at least on some level, these symbols are part of a 'language of thought' or 'Mentalese'—i.e., that the symbols that make up thought are linguistic in form and function, but are part of a special linguistic system distinct from any public language.[114] What truly seems to set the LTT apart from all those views with which it shares varying degrees of similarity, then, is the fact that, according to the LTT the symbols that constitute thought are not just

[114] There are the obvious examples of Fodor or Pinker, each of whom embraces the idea that all thinking is done in Mentalese. Then there are people who only go part way: most notably Peter Carruthers (see following footnote). On the other hand, philosopher David Cole seems to endorse a partial natural language thesis, though in a way quite different from Carruthers or myself, without appealing to Mentalese at all (see Cole, "Hearing Yourself Think", 1997; and "I Don't Think So", 1998). Cole always qualifies his statements about thinking in natural language by saying that 'much' or 'most' thinking takes place in natural language (what medium the rest of thought takes place in Cole doesn't say), but at the very least he does offer arguments against Mentalese—or, at any rate, arguments that purport to show that the Mentalese approach isn't as obviously correct as it is often, in the cognitive sciences, accepted to be.

language-like, they are actually symbols of our public languages. There is not just one 'language of thought'; there are many languages of thought.[115]

5.11 Thinking in Natural Language

While the suggestion that thinking takes place in the symbols of the natural language(s) that one speaks is, to my knowledge, rather rare among philosophers of mind (which is not to say that it hasn't been considered, of course—just not endorsed; but we'll come to that), my reason for suggesting it is fairly straightforward. I believe that it is in many ways the simplest view—or, at least, the simplest view that will get the job done. The main issue here is a debate between those who would posit a universal language of thought (Mentalese), like Fodor, and those (seemingly quite few) like myself who eschew Mentalese altogether. This debate takes place, of course, within that subset of representationalists who take mental representations to be linguistic in nature. (I call this group a subset of representationalists because there is nothing in RTM generally that requires mental representations to have a language like structure, let alone constitute an actual language. Yet the linguistic conception of mental representation is really the dominant view within cognitive science, for many reasons. Recall our discussion of Lycan in §1.13 above.) As I said, though, the issue here is really just the question of

---

[115] It's probably worth noting that in a footnote in "Notes on Intentionality" Sellars writes, "There is indeed, every reason to suppose that Japanese inner speech differs systematically from English inner speech in a way which reflects the differences between these two languages" (1972, p.331). While this does not suggest that the natural languages just are the languages of thought, it does suggest that Sellars embraces the idea that Mentalese can vary widely between different thinkers in a way that reflects the differences in their spoken languages. This should be contrasted with people like Fodor or Pinker who seem to suppose, or even require, that the language of thought is universal in all its characteristics. On the other hand, as we'll see in the next chapter, Peter Carruthers actually endorses a view in which at least some thinking does in fact take place in natural language. But Carruthers also relies heavily on a notion of a universal Mentalese, and I have serious reservations about whether the role he assigns to thoughts in natural language is even coherent. As I said, though, we'll deal with that in the next chapter.

whether or not we need Mentalese. I will be addressing this question directly in the following subsection below (§5.12); there I will discuss some of the reasons that other theorists have rejected a natural language theory of thought, and I will show that when we take all the pieces we've looked at so far (a functional role theory of meaning, the Sellarsian framework for Psychological Nominalism, etc.) and combine them in the right sort of way we can respond to these worries. The upshot is this: I believe that if we can explain thought without adding, on top of all the natural languages, a mental language, then that is the route that we should take. We apply Ockham's razor and keep our ontology as small as possible. Before we get to that, however, I would like to say just a little bit about what a theory that takes the language of thought to be the language(s) one speaks might look like. Doing so will better position us to assess the arguments against Mentalese that will follow.

To begin, let's quickly recap the key parts of our theory so far. As a version of representationalism, the LTT is (as we've seen) one among many theories that suppose that to have a thought is to token a mental representation of some kind. We saw in Chapter Two that this approach divides into questions along two distinct, though related, axes, which we called form and content. When it comes to the latter, the question of a psychosemantics, the LTT embraces a Sellarsian functional role theory of meaning. With regard to the former, the question of the kind of representations that constitute thoughts, the LTT sides with those philosophers who embrace the suggestion that thought requires some sort of language-like symbols (again, recall Lycan's arguments as we discussed them in §1.13 above). We are now adding a final feature along both axes.

Obviously, in claiming that thoughts take place in natural language, i.e., are mental tokenings of natural language symbols, we are giving a final, determinate character to the type of representations our theory supposes constitute thoughts. While other theories posit a semi-mysterious Mentalese, the character of which remains somewhat less than determined, the LTT gives us a clear, full picture of the symbols of thought—the representations that constitute thoughts are no more mysterious than the representations used in writing or speaking.[116]

Along the other axis, if we combine what we have already said so far about linguistic meaning and language learning (see Chapters Three and Four) with the suggestion that natural language symbols constitute the form of thoughts, we find ourselves with a very clear picture of the psychosemantics of the LTT. Since I am claiming that thoughts are constituted by tokenings of the symbols of the language(s) that one speaks, and have already argued that the content of natural language symbols is determined by the role those symbols play in the language of which they are a part, it follows quite simply that the contents of thoughts are also determined by the role played, in the linguistic community, by the symbols that constitute them.[117] That is, whatever content the overt tokens of the symbols of a language have in virtue of their functional role, that content is also the meaning of the covert tokenings of those symbols when one has a thought.

---

[116] There are, of course, important differences between mental tokens of a natural language and written or spoken tokens of the same. I will address those differences in subsequent sections below.

[117] This move is possible, of course, only because our theory of linguistic meaning, as developed in Chapter Three, does not presuppose—does not rest in any way upon—the prior existence of meaningful thoughts. That's not to say that there aren't potential difficulties still; I will be dealing with some of these potential worries in Section 5.2 below.

The picture that emerges from these points is fairly straightforward, I think (at least in its broad strokes). Again, to put things bluntly, to have a thought, according to the LTT, is to mentally token symbols of the language(s) one speaks. An English speaker thinks in English; a French speaker thinks in French. According to the LTT these aren't metaphors or figures of speech, they are literal truths. As an English speaker my thoughts *just are* tokens of English words and sentences. We should note, however, that we are not suggesting that thoughts are merely silent speakings—talking without moving one's mouth, sort of like muttering under one's breath. They are distinct tokenings, in whatever medium is relevant, of linguistic symbols that can be tokened in other mediums as well (e.g., as sound waves when speaking, as marks on a page when writing, etc.).[118]

While I take this picture to be ultimately pretty straightforward, there is room for some confusion in the details. The latter half of the preceding paragraph actually points us in the direction of the largest trap of confusions into which we, if we are not careful, can easily fall. For many, the most natural way to attempt to grasp the type of picture that I'm suggesting is, perhaps, to imagine thinking as a sort of internal monologue. Maybe one imagines thoughts on this view as akin to reading words written in the mind; or maybe one imagines thoughts as the speaking of an internal voice that only the thinker

[118] The medium that is relevant for 'mental tokenings' depends on the being under consideration. Neuroscience gives us, I think, good reason to suppose that the medium within which human thoughts are tokened is brain matter—that the linguistic symbols that make up thoughts are encoded as electrical and chemical occurrences in the brain. There is no reason to suppose, however, that other media couldn't serve just as well for other beings. For example, perhaps silicon chips and computer hardware will one day serve as the medium for the thoughts of sentient robots. In the end, the medium is more-or-less irrelevant, though. What's important is that the symbols be of the right sort, i.e., be symbols of a public language, and that they be tokened in the right way, e.g., in a non-parroting manner. The medium is doubtless important for any number of reasons from a scientific standpoint, but theoretically the LTT requires no assumptions about 'hardware.'

can 'hear.' Both of these suggestions, and anything like them, however, lead to problems and are not, in my view, fruitful ways to understand the idea I'm proposing.

In many important respects (e.g., how meaning gets determined) thinking in natural language really is like speaking a natural language. But imagining thinking as an internal monologue that one must follow along with (and could, perhaps easily, vocalize should one choose to) is a mistake. Thinking takes place in natural language, but there are important dissimilarities between thinking something and saying it, or writing it, etc. In many ways the elements of dissimilarity between, say thinking and speaking will mirror dissimilarities between, say writing and speaking. After all, tokening an English sentence by writing it on a piece of paper and tokening that sentence by producing sound waves with one's mouth differ in quite significant ways. The medium in which the tokens occur is very different, the forms they take (marks on a page for visual consumption versus vibrations in the air for auditory consumption) are completely distinct, the speed with which they can be produced may vary drastically—all these and more will distinguish writing from speaking. And similar considerations will distinguish thinking from writing, speaking, signing, etc.

While I think that these dissimilarities are important (as are the similarities), perhaps the most important point to make here is that the above suggested ways that one might imagine thinking in natural language are alike in making it sound as though thinking is something that the intentional agent itself must somehow 'witness.' The mistake here, though, is obvious: the intentionality of the agent is a consequence of the fact that the agent thinks, not something independent of that fact. So the image of one 'reading' a stream of natural language symbols written in the mind proves to be a

dangerously mistaken way of trying to understand thinking in natural language primarily because it invites this mistake of at least implicitly positing an intentional agent independent of the stream of thoughts—an agent that must 'read' the symbols that constitute those thoughts. The stream of natural language symbols that constitute the tokenings of various thoughts are not to be 'read' by the intentional agent, they are the very things that make the agent an intentional one to begin with. (This is not to say, of course, that one cannot be self-consciously aware of one's thoughts. I think it's clear that we can be. But this is not the same thing as simply having those thoughts to begin with.)

This potential point of confusion arises quite naturally because we are used to thinking of public tokenings of linguistic symbols, in speech, or writing, as things that we, as thinking beings, do 'witness.' We hear others (and ourselves) speak; we read what others (and ourselves) have written; and in such cases there is a separation between the linguistic tokens and us as intentional agents hearing or reading them. When we think, however, self-conscious awareness of those thoughts aside, we are not playing the role of audience or interpreter, witnessing the tokenings of linguistic symbols that constitute our thinkings. Those tokenings are the foundation of our intentional, conceptual selves to begin with.

So, while from one point of view thinking is to be analogous to speaking or writing, from another point of view thinking will be very different from speaking or writing. (Notice that in this latter case the speakings or writings are not thinkings-out-loud, for even though there are still differences between silent episodes of thinking and the publicly observable episodes of thinking-out-loud, the differences are not nearly as pronounced as the differences between thinking, on the one hand, and speaking for public

consumption, on the other.) Again, then, just so long as we avoid the mistake of imagining thinking in natural language to be like reading or hearing an internal monologue, the picture presented by the LTT is fairly straightforward: thoughts are the mental tokenings of representations constituted by the symbols of the natural language(s) one speaks.

This means that the mental representations of the LTT have a great deal in common with the mental representations of the LTT's representationalist peers. When Fodor, or Lycan, or Pinker suggests that thinking occurs in some sort of 'mental language' I don't think they're all that far off base. Thoughts are mental tokenings of linguistic symbols, and as such are distinct occurrences from the overt speakings, or writings, or whatever that are public tokenings of linguistic symbols. Where the LTT differs from the views proposed by these, and other, theorists is in supposing that the distinct episodes of the tokening of linguistic symbols that constitute thoughts are still tokenings of natural language symbols rather than symbols of a special 'Mentalese.' Of course, the psychosemantics of the LTT is quite distinct from the psychosemantics of, say, Fodor's Language of Thought hypothesis—but as we've seen, differences in psychosemantics have largely been the only real distinguishers of one representationalist theory from another anyway. The holistic, functional-role semantics of the LTT works with the other parts of that theory just as the atomistic, causal-covariance semantics of Fodor's LOT works with the rest of his theory.

This point is important mainly because I want us to recognize that the LTT, while departing radically from other representationalist views in some respects (see, e.g., Section 4.1), is in many ways just building upon the insights of its representationalist

predecessors. As I've said all along, my purpose in this project is to carve out a space for the LTT among related views. Within a certain range of such views, however, the LTT is distinct in supposing that we needn't posit a special language of thought if the languages already available to us, the languages that we speak, will do. Thus the major point of contention between the Linguistic Theory of Thought and other 'language of thought'-type views is the question of whether or not natural languages really can serve as the languages that constitute thoughts. I now turn to some of the reasons that others have supposed that they cannot so serve, and my replies to these worries.

5.12 Do We Need Mentalese?

There are a great number of reasons that someone might suppose that if thought is the tokening of linguistic symbols, then those symbols must be symbols of a special language of thought rather than symbols of natural language.[119] These reasons vary in force and in scope: some are intended to speak strongly (taken with other considerations as well) in favor of Mentalese, while others are simply supposed common sense reasons for rejecting the natural language hypothesis; some purport to show that natural languages are entirely unsuited to be the languages of thought, while others aim only to show that not all thoughts could take place in natural language. The list of possible objections that might be raised (and, again, I recommend Pinker's book, or Cole's response to it, as places to find a nice list of these various objections) runs the gamut from the mundane, and probably rather trivial (e.g., the common phenomenon when speaking of sometimes having to try to 'find the right word,' or the fact that bits of

[119] Chapter Three of Steven Pinker's *The Language Instinct* contains the largest single collection of arguments, hints, and allegations to this effect that I am aware of.

natural language can sometimes be ambiguous—but thoughts never seem to be), to the more theoretically sophisticated (including, e.g., the worry that if thoughts took place in natural language translation between languages would be impossible; the worry that language learning would be impossible on a natural language theory of thought; and the objection that animals which lack any language whatsoever can nevertheless think). I haven't the space here, nor is there really the need, to go through all the possible objections individually. Ultimately the worries about taking natural languages as the languages of thought, and the replies to these worries, generally center on just a few issues, all of which are related to the core elements of the LTT. If we look at just a few examples, then, I think it'll be clear how one could go about responding to the other worries.[120]

Let me begin with one of the seemingly more trivial objections, and then we'll work our way up to a couple of the most serious objections. Though the objection itself may seem trivial, I begin with it because I think that it will help us see a common feature (which I will then discuss) of many of the worries one might have about the natural language hypothesis, so this particular objection will lead us into a more general, and generally fruitful, discussion. The objection I have in mind goes something like this: the sentences of any natural language can be ambiguous, but one's thoughts are normally not so. Thus, for example, if I utter the English sentence "I'll be waiting by the bank" my utterance may be ambiguous, given that the word 'bank' has at least two distinct meanings—a financial institution, or the side of a river. Surely, though (goes the

[120] Indeed, after we've looked at the issues central to the handful of examples I'll be discussing here we could, had we the space and inclination, go through Pinker's list ticking off responses to each point one by one—with a certain mind-numbing repetition, I might add.

objection), the *thought* expressed by that sentence isn't ambiguous. I obviously know, when I utter that sentence, whether I am talking about a financial institution or a riverbank. But then the thought must be something other than the tokening of that English sentence, for the sentence, but not the thought, can be ambiguous.[121]

Now, insofar as this point is correct, I contend, it is irrelevant; and insofar as it is relevant, it's simply mistaken. Let me explain. It is true, of course, that, taken in abstraction (or taken from the point of view of someone, say, hearing it), the English sentence "I'll be waiting by the bank" is ambiguous. This can hardly be relevant to the issue of whether or not a natural language, like English, can be a language of thought, however, since thoughts do not occur in abstraction, nor (as we've just seen) as tokens to be interpreted by the thinker as audience. To make this worry relevant, then, we need to shift away from consideration of the English sentence in abstraction to consideration of that sentence in context.

Once we do that, though, we see that it is simply mistaken to assert that there is any ambiguity in the sentence. When I say "I'll be waiting by the bank" I know whether by 'bank' I mean 'financial institution' or 'side of a river.'[122] Yet, though this is true, haven't we then just reiterated the worry that we began with? The objection just is that I *do* know what I mean—that my thought is not ambiguous. But the sentence can be. What we really need to know is how, if thoughts are simply to be tokens of natural language sentences, they can still manage to be unambiguous for the thinker. My point,

---

[121] This objection is actually given a fair bit of serious discussion by Pinker, but I think that some form of this objection could also be teased out of the views of many of the modern representationalists that we've been talking about throughout this project.

[122] Or, if I don't know what I mean my thought is just as ambiguous as the utterance—but then, the whole force of the ambiguity example is supposed to come from the fact that the utterance, but *not* the thought, can be ambiguous, so this won't help the friend of Mentalese.

though, is that this is no more troublesome than accounting for the fact that speech is generally unambiguous for the speaker. The fan of Mentalese, of course, will argue that speech is unambiguous for the speaker because it is the expression of unambiguous tokens of Mentalese.[123] This sort of solution is actually open to me as well, in the case of speech (since I grant a distinction between covert and overt linguistic tokenings), but that's rather beside the point. The real solution on my view is to recognize that the linguistic token (be it thought or overt speech) occurs in a particular context, that is, as a token with a certain functional role. And this contextualizing of the tokening typically eliminates ambiguity.

My thought, "I'll be waiting by the bank," occurs not in abstract isolation, but within a functional network that includes connections to the world and to other bits of language. This is the essence of a functional role semantics. Without these functional connections the representations would be meaningless. But once we include these functional connections in the picture, ambiguity more-or-less disappears. When I think, "I'll be waiting by the bank," the supposedly ambiguous word 'bank' isn't ambiguous at all because its tokening is functionally connected to all sorts of other things (e.g., things in the world, such as a financial institution, most likely a specific building, as well as other pieces of English, like the symbols 'financial institution'), that completely determine its meaning. Of course, someone who hears me say "I'll be waiting by the bank" out of context may not know what sort of bank I'm referring to, because he will

[123] This will be the case, at least, for all those fans of Mentalese who do not believe that Mentalese tokens can ever be ambiguous—and I imagine that includes most of them.

not know what functional connections are operative. But I cannot produce that token outside of a context that determines its content.[124]

The real issue here (as with related worries about vagueness, saying what one means, translating between languages, etc.), of course, is the issue of how words become meaningful in the first place. Those who are moved to embrace Mentalese by examples like the ambiguity example above will, as I've noted, be tempted to respond to my argument by saying something like, "Of course the sentence 'I'll be waiting by the bank' isn't ambiguous to the person who utters it—but that's because *she* knows what she's thinking, and hence knows what she means by 'bank'." Yet it should be obvious that such a response presupposes that overt linguistic tokens of a natural language 'get their meaning' at least in part from the thoughts that they 'express.' The only reason that the ambiguity example seems to work as an objection to the natural language view is because we are led to implicitly accept a separation between an utterance in natural language and the thought that utterance is supposed to communicate to others. I'd be concerned with the question begging nature of this objection if it weren't for the fact that I find the whole presupposition behind it mistaken. Once we identify and reject this particular presupposition, the ambiguity example loses its force.

Now, have we successfully rejected this presupposition? The answer to that depends upon what one thinks of our discussion of meaning in Chapter Three. The point of endorsing a functional role theory of meaning, so far as the current issue is concerned,

---

[124] Some time after having developed this argument myself, I came across a very similar argument by David Cole (see "Hearing Yourself Think", 1997). Cole's argument also appeals to context to defeat the ambiguity argument of the fan of Mentalese. He even uses 'bank' as his example of an ambiguous word (though that my not be terribly surprising—it's an oft-used example in philosophy). Cole's overall view of thought is somewhat different from mine, however, as I mentioned in an earlier footnote.

is so that we can talk about words being meaningful without invoking the independent meaningfulness of antecedent thoughts. I find the arguments for such a theory of meaning compelling, but suffice it to say that so long as it's possible to have a theory of the meaning of overt linguistic utterances that does not rely upon the prior meaningfulness of thoughts then objections like the ambiguity objection aren't going to provide any reason to add Mentalese to our ontology. According to the LTT, the thought and the utterance are both tokens of the same English sentence, and the meaning of each sentence token (the overt utterance, and the covert thought) is determined by its functional role within the language game. That role is defined by the position that it occupies in the network—by its relations to other linguistic tokens and to the environment—and by this measure the two different readings of the ambiguous English sentence are actually readings of two different sentences, since those sentences occupy different positions in the language game, while the thought and related overt utterance are the same sentence, since they occupy the same position in the game.[125]

I think that it is important to note at this point that I am not suggesting that ambiguity somehow disappears on my view, or is only superficial. Ambiguity remains, on my view, just as problematic as it has always been, and in just the same sorts of cases, too. My point is simply that the fact that words and sentences used in *communication* can be ambiguous gives us no reason to suppose that natural languages cannot be the

---

[125] Hence: (1) "I'll be waiting by the bank" where 'bank' means 'financial institution' and "I'll be waiting by the bank" where 'bank' means 'side of a river' are, on this analysis, tokens of different sentences (though constructed of the same symbols—e.g., in this case, marks on a page); (2) my thought "I'll be waiting by the bank" and my overt utterance of "I'll be waiting by the bank" are tokens of the *same* sentence (in that they play the same role—I'm not positing sentence types as abstract entities here). (Compare: (1a) a wooden pawn on A4 and a wooden pawn on B2 are different pawns, though constructed of the same material; (2a) a physical realization of a pawn on A4 and a computer generated pawn on A4 in a recreation of the same board position are instances of the *same* pawn (in that they both occupy the same position in the game)—this is how, e.g., IBM's Deep Blue was able to play against Kasparov.)

languages of thought. Ambiguity is, and always has been, a problem of communication, not a problem with the words themselves. So there is no reason to suppose that words that can be ambiguous in interpersonal communication will remain so when they are taken as the vehicles of thought.

Now, as I've already said, I think that the issue of how words become meaningful is at the center of most of the small, individual worries (most of the worries on Pinker's list, for example) that motivate theorists to endorse Mentalese over a natural language theory of thought. There are, however, at least two other significant worries that do not fall into quite the same boat. One is related, in that it has to do with meaning, but is significant enough on its own to deserve separate discussion (not to mention the fact that the above response won't work): this is the question of how meaningful language use could ever get off the ground on a functional role theory of meaning (or any theory like it, really, that didn't at least partially ground the meaningfulness of overt linguistic tokens in the prior (inherent?) meaningfulness of thoughts). The second issue is wholly unrelated to worries about meaning, but is perhaps one of the strongest (if not *the* strongest) motivators for rejecting natural languages as the languages of thought: this is the fact that, if natural language is required for thought, then all beings (e.g., non-linguistic animals, pre-linguistic humans) that lack natural language simply cannot think.

Unlike with the example of the ambiguity objection, while these two further examples of objections to a natural language approach to thought do raise issues central to the LTT, I cannot give a thorough response to either of them (due in part to time and space constraints, but also in part to the fact that I have not yet been able to work out thorough answers). Nevertheless, let me say a little bit about each worry. Hopefully I

will be able to say enough to at least show that the natural language approach isn't inherently absurd.

The question of how meaningful language use could have ever got started in the first place if the meanings of linguistic tokens are not at least partially determined by the prior meaningfulness of thought episodes obviously brings us back into contact with the worry in sections 4.22 and 4.23 about how one can move from language learner to language speaker. Of course, the issue here magnifies that worry quite a bit, since at least when looking at the transition from language learner to speaker we can take the existence of a meaningful practice of language use for granted. In the current case, though, not only can we not take such a practice for granted, it is the very possibility of such a practice ever coming to exist that we are questioning. Imagine, then, a period in pre-history. We have a group of animals, pre-human in the sense that they do not yet possess anything like a language (the specifics here aren't the point—all we want to do is imagine creatures that are physically capable of learning language, but haven't yet developed language; they could be Neanderthals, or like gorillas in the wild—whatever works). We might imagine that these pre-humans begin to make noises in response to their environment. Perhaps they even mimic each other's noises some of the time. The question is: At what point do their noises become meaningful? At what point do their noises begin to constitute a language?

Now if meaning is a matter of the functional role of linguistic tokens the key to the noises of these pre-humans coming to have meaning is going to be those noises cohering with their environment and with each other, and being systematic enough that they form a functional system within which pieces and moves can take determinate

shape. Yet we have to be careful in saying this, for we do not want to open the door to just any sort of systematic behavior constituting a language. The 'dance' of bees, for example, is not to count as a language. What can we do? Well, bringing up the bee dance might remind us of what separated the bees' behavior from the linguistic behavior of human beings in our earlier discussion: the bees' behavior is merely pattern-governed, while the linguistic behavior of human beings is rule-governed. Yet this distinction doesn't seem to make our problem go away; if anything, the noise-producing behavior of our pre-humans is merely pattern-governed behavior just like that of the bees' dance. Of course, if we remember our discussion in Chapter Four we will recall that essential to the noises of a language learner becoming rule-governed speech was recognition by the producer of the would-be linguistic tokens that such tokens were part of a rule-governed system. This issue of 'recognition,' we may recall, proved a bit problematic in the end, though we were at least able to suggest that it might have something to do with the reliability of the production of the linguistic tokens, with the ability of the system to self-correct, and perhaps some other features.

So, when does the pattern-governed noise-producing behavior of our pre-humans start to become the rule-governed speech of the earliest human beings? I have no clear answer to that question. I have suggestions to make—again, the ability of the system to self-correct seems important; the reliable correlation of the production of certain noises with certain states of affairs in the environment, and the production of certain other noises seems important—but I have no clear answers. I hope, however, that the lack of clear answers to this question does not seem to defeat the suggestion that thinking takes place in natural language. I don't find the question of how language could have become

meaningful without depending in some way upon the prior meaningfulness of thought to be any more vexing than the question of how (on other views) meaningful thought could ever have arisen in the first place. The answer, whether you're an advocate of Mentalese, or attempting to develop a view like the LTT, lies in the intricacies of the evolution of species from single celled organisms to thinking human beings. That's not an argument in favor of a natural language theory, to be sure, but I do think this point at least blunts the worry that the proponent of Mentalese might try to level at the LTT.

Finally, though, we have the worry that is at once both the most straightforward and also the one most likely to seem decisive to the vast majority of people, at least on a first pass. As I've mentioned, if thinking is the tokening of mental representations constituted by symbols of a natural language that the thinker speaks, one obvious consequence is that any being that lacks a language simply does not, and cannot think. Yet that strikes many people as so obviously absurd as to immediately falsify the claim that thinking takes place in natural language. To many people, nothing could seem more contrary to common sense than to suggest that animals and infants do not have thoughts. How else are we to explain the complex and seemingly intelligent behavior of so many non-linguistic and pre-linguistic beings if we are not to ascribe to them at least rudimentary beliefs, intentions, and desires? If my guard dog thinks that there is an intruder, it will bark; if my newborn child wants to be fed, it will cry. Monkeys can choose to deceive, babies can keep track of numbers. And this isn't even to mention the fact that it seems to many cognitive scientists today rather perverse not to suppose that there is a continuity in cognitive abilities from the simplest insects to the most talented human thinkers.

What can I say in response to this? Just as the worry is the most straightforward, so too my reply is the least complicated. To put it bluntly, this is a bullet that I'm perfectly willing to bite. That doing so is not fatal is by no means obvious; but neither is it obvious that it is fatal. What does the issue turn on? Well, in a sense it turns on the success or failure of this project as a whole. Thus this is not a worry that we can even hope to defuse at this stage. I think that the issue also turns quite a bit on just how we imagine the alternative to the idea that animals and infants think—that is, in suggesting that they do not think, are we saying they are unconscious? Are we saying they're mere automata? Are we suggesting that their behavior is nothing but simple stimulus-response patterns? I think the answers to these questions are 'No' in every case, but then we should suggest alternatives. Yet the suggestion of alternatives is no simple matter, because the issues are so complex. I will have a bit to say about consciousness, sensation, and the interaction of both thinking and non-thinking organisms with their environment in the next, and final, chapter of this project. And I do hope that what I have to say there will help to defuse the force of this worry somewhat. Nevertheless, even more so than with the previous worry, I simply am not in a position, nor do I have the space here, to respond to this worry thoroughly. Again, it is an undeniable consequence of the LTT that non-linguistic and pre-linguistic beings do not have thoughts. But that is a consequence that I am prepared to accept.

In the following sections I will bring us back down from these lofty heights of supposedly theory-shattering objections, and vast, undetermined (in the present project) replies. With the idea that thinking takes place in natural language now on the table and, as much as is possible under the present circumstances, defended as a feasible suggestion,

I turn to more concrete issues of the relationship between thought and language within the framework of the LTT.

5.2 Learning to Think

If to think is to mentally token symbols of the natural language(s) that one speaks, then just as one must learn to speak a language, one must learn to *think*. We're not, of course, talking about learning to think in the way that, e.g., one's education is supposed to teach one to think—instilling knowledge of this or that, honing one's reasoning skills, etc. Rather, we're talking about becoming able to have conceptual mental states in the first place. Most philosophical theories of mind do not find themselves in the position of having to provide a robust account of learning to think in this sense, since most theories of mind suppose that at least the ability to have basic concepts (say corresponding to one's perceptions of the environment) is more-or-less innate, pre-programmed, hard-wired, or something along those lines. Yet, again, if having conceptual mental states is the tokening of symbols of a natural language, then one will only come to have such states as one acquires a language. And it seems reasonable to suppose that the acquisition of language and the acquisition of thought will occur hand-in-hand with each other. What I want to do in this section, then, is quickly revisit the issue of language learning, and then talk about how the idea that one thinks in natural language shapes and changes how we conceptualize the relationship between speaking and thinking, and between learning to speak and learning to think.

5.21 Language Learning Revisited

Of central importance at this juncture is Sellars' notion of 'thinking-out-loud.' I want us to understand 'thinkings-out-loud' (and I think this is how Sellars wants us to understand them) as overt linguistic tokenings that are produced not for the purpose of communication but merely as (sometimes uncontrollable) overt instances of thinking itself. That is, the utterances are to be identified with the thinkings, rather than seen as consequences of thought episodes. This is not to say that there are not covert, 'inner' episodes that accompany thinkings-out-loud, only that it would be a mistake to identify these 'inner' episodes as the 'real' thinkings, and the utterances as mere consequences of thought. Let's put that issue aside for the moment, though. What is important at this point is just to be clear that there is a kind of public linguistic tokening that is to be taken not as the consequence of thinking, but as thinking itself.

This idea, which I have taken from Sellars, is actually entirely at home within the framework of the LTT. We saw above that according to my theory, to have a thought is to token, in the appropriate way, the symbols of a natural language. But of course speaking is perhaps the most familiar way of tokening such symbols. The only question would be whether or not linguistic tokens produced by speaking could count as having been tokened in a way that is appropriate for thought episodes. With one important qualification, I see no reason to suppose that they could not. The qualification is the obvious one: the tokens must be *meaningfully* produced if they are to count as thoughts, i.e., as thinkings-out-loud. This, however, brings us back to the issue of language learning, for as we saw in Chapter Four, the central issue that faces a VB model of

language learning is the question of how one goes from producing linguistic tokens in a purely parroting manner to producing those tokens meaningfully.

Now, again, while I admit that I am not yet in a position to say exactly how language learning proceeds, I do think that I can paint a broad picture; and given the final element of the LTT (that thinking takes place in natural language) I think we can add one more brush stroke. Previously what we imagined was a child who began by mimicking the linguistic tokens of his teachers and ended by having a full command of the language himself. While we were fuzzy on the details, the basic picture involved a period of behavioral conditioning wherein the teachers followed the ought to do rules of language to bring it about that the learner's behavior conformed to the ought to be rules of the language. Once the learner's conformity to those rules became ingrained in the right sorts of ways the learner became a speaker. (I also favored the gradual picture of this transition, the reader may recall, but that's really neither here nor there for our present purposes.)

So what changes when we add the idea that thinking itself is the meaningful tokening (covert or overt) of one's natural language? Clearly, the main thing that changes is that when the language learner is being conditioned to produce linguistic tokens in accordance with the rules of the relevant language he is not being trained merely to speak, but to think as well. There is a simultaneous acquisition, that is, of both the ability to participate in the language game *and* the (for lack of a better phrase) 'thinking game.'

Of course, this is also more-or-less the case on Sellars' view; but while the relationship between thinkings-out-loud and Mentalese isn't entirely clear, the

relationship between thinkings-out-loud and the covert thought tokens of the LTT is entirely clear: they are simply different modes of tokening the very same symbols. In fact, we could fairly safely presume that the language learner is simultaneously producing overt linguistic tokens and their covert counterparts throughout the learning process. This assumption may even seem necessary, for only if it is true will the conditioning of the production of the appropriate overt tokens also be a conditioning of the production of the appropriate covert tokens. Recall our discussion in Section 4.1 of the problem faced by the conditioning of Jane's dog: in order for the correcting of its overt behavior to be a correcting of its 'inner' representations of the world, the overt behavior and the inner representation had to be necessarily linked, so that the conditioning of the former could also count as a conditioning of the latter. In the case of Jane herself, then, we saw that the only way to overcome the problem was to suppose that her overt behavior and her inner representations *were* linked in some way. I suggested there what I have made more of a case for since: the simplest way to suppose that Jane's overt behavior and inner representations are necessarily linked is to suppose that they are *both* instances of the same phenomenon—*viz.*, the tokening of particular linguistic symbols.

As the child takes its first steps toward acquiring a language, then, it is also practicing with the essential building blocks of thought. The tokening of symbols of a natural language, both overtly and covertly, is a process that gives the potential speaker and thinker contact with the raw materials out of which is constructed a conceptually informed view of the world. To truly develop this idea, however, we need to say more about just how the relationship between the overt and covert tokenings of a language is to be understood.

5.22 Talk and Thought

To begin with we should probably say a little bit about just what it means to 'covertly token' some piece of a natural language. I've been making casual use of that idea, and in one sense it's just another way of talking about the mental tokening of symbols that is at the heart of the representationalist picture—with the added idea that the symbols in question are symbols of a natural language. To the extent that we're just talking about the mental symbol tokenings of the representationalist picture the issue is probably not all that important. Theoretically these 'covert tokenings' could be anything from symbols realized in Cartesian 'mind-stuff' to various patterns of brain activity. Since contemporary representationalists are almost always physicalists, though, most of the time the mental symbol tokens will probably be viewed as some sort of neural activity (at least in the creatures we are familiar with).

Of course, we also have to remember that, if contemporary representationalists are physicalists, they are functionalists first (representationalism is a form of functionalism). So even though most contemporary representationalists would probably say that the mental symbol tokenings that constitute thoughts are going to turn out on some level to be patterns of neural activity, there is nothing special about neurophysiology *per se*. Just so long as we can realize the same functional network in other materials, brain matter is ultimately irrelevant. For the representationalist all this means is that if we can get the appropriate representations to occur in the right sorts of ways (and the 'right sort of way' for any given theory will be largely determined by its

psychosemantics) anything from robots to Martians can have the same thoughts that we have. The material, or medium, in which thoughts occur is ultimately irrelevant. [126]

Again, then, the question is this: What is the proper understanding of the relationship between the covert and overt tokenings of language (disregarding, as much as possible, the media within which those tokens occur)? If overt tokenings in the form of thinkings-out-loud are properly to be construed as thoughts themselves, and not merely the expressions of thoughts (as the traditional view would have it), what is the role of the covert tokenings? Are they part of the thought, or related thoughts, or conceptually similar precursors to the thought? Do we even need them at all? Moreover, since we are supposing that the ability to speak a language, and hence the ability to have thoughts in the first place, is gained through the behavioral conditioning of would-be language users by those already in the space of reasons, how are we to understand the role of this public training process in developing the right sorts of covert linguistic activity? In short, how does one actually learn to think?

What we're interested in here is the way that 'inner' episodes of the tokening of linguistic symbols figure into the overall theoretical framework of the LTT. Let me approach this issue with a quick reminder run-up. Like all representationalist theories, the LTT proposes that we view thoughts as mental tokenings of representations. According to the LTT, the representations in question are symbols of the natural

---

[126] Philosophically, at least. I imagine that there are many empirical considerations that would constrain the types of material that could realize mental states. Certain of these constraints may even be suggested to us by the psychosemantics that we choose. For example, if we are to endorse a functional role semantics with a wide content element it will be imperative that any thinking being be such that it is capable of having its mental representations connect up with the world in specific ways. But this will almost certainly place speed of processing constraints on the materials within which mental states could be realized. Such empirical constraints will be important for cognitive science as a whole, but they can be safely sidelined in our present discussion.

language(s) that one speaks. The content of those representations is determined by the role that the symbols play in the language. So thinking, according to the LTT, is mentally tokening meaningful symbols of one's language. But what does it mean to 'mentally' token such symbols? As opposed to what?

Well, as opposed to tokening them out loud—that's what it usually means, I think. For the most part it's not actually all that clear what representationalists mean when they talk about 'mental' tokenings of representations. The usual formulation is that mental representations are tokened 'in the mind or brain'—but that is hardly a useful explanation of what is meant, especially when it is the mind that we are trying to explain. Moreover, since the LTT attempts to take Sellars as seriously as it takes contemporary representationalism, and on Sellars' view there is a kind of public tokening that is to be considered not simply the expression of a thought, but thought itself (*viz.*, thinkings-out-loud), the suggestion that thoughts occur 'in' the mind or brain is even more problematic.

At least part of the issue here, of course, is the fact that most modern representationalists are traditionalists about the relationship between language and thought; i.e., they adhere to the view that overt linguistic utterances are merely the expressions of thought. Thus their picture involves, almost necessarily, a distinction between the 'inner' episodes ('in' the mind or brain) that are one's thoughts and the 'outer' expressions of those thoughts, in speech, or writing, or whatever. To the extent that Sellars wants to preserve this sort of distinction (and it's clear that he does want to preserve it for the most part), his story is going to end up being much more complicated. After all, Sellars will be faced not only with the task of saying what it means for a thought episode to be 'inner' (the same task faced by contemporary representationalists

and addressed—if one can call it that—by their rather weak appeal to the thoughts occurring 'in' the mind or brain), but also with the task of explaining how these 'inner' episodes relate to instances of *public* behavior that are nevertheless considered to be episodes of thinking. That's no easy task, to be sure.

One might begin to wonder, at this point, why there have to be inner episodes at all. That is, why couldn't it just be the case that all thinking was the overt tokening of linguistic symbols (with the qualification that such tokenings be performed 'in the right sort of way,' of course)? A large part of the answer to that question, it seems to me, is that representationalists are keen to avoid the pitfalls of behaviorism. If we were to start identifying thinking with producing overt linguistic behavior, we might very well feel as though we were sliding down the path toward a behaviorist dead-end. Thinking, goes the idea, is more than just behavior—even linguistic behavior. And this is all the more evident in those cases in which one thinks but doesn't say anything. (To explain 'silent thinking' on an account that dismissed the idea of 'inner' episodes we would, it seems, have to resort to appeals to dispositions to produce overt linguistic behavior—and once again behaviorism would rear its ugly head.) The desire to avoid behaviorism is strong, and it is a motivation that I share.

Thus, it seems to me, we have at least two requirements that we can put on any given instance of the tokening of symbols of a natural language if such a tokening is going to count as a thought. First, the tokening must count as making a move in a language game. This will invoke all that we have said so far about what makes a given tokening a move in a language game, including that the token have a functionally defined meaning, and that it be produced in 'recognition' by the thinker that it is a rule-governed

symbol of the language, etc., etc. This requirement ensures that the tokening will meet the normativity condition set out in Section 4.1 earlier. It also ties the token itself to the appropriate language, thus satisfying my desire to make one's natural language the language of thought. Secondly, though, the tokening must be more than just an instance of overt verbal behavior. That is, to avoid the behaviorist trap, there must be more to thinking than overt verbal behavior and dispositions to behave. The most obvious suggestion, here, is to suppose that there are states of the being in question that actually are the thoughts that the overt verbal behavior can then express. These are the 'inner' episodes of the representationalist picture. For that matter, these are going to be the functional roles of the basic functionalist picture. And that, I think, provides us with the key to unlocking a clearer idea about just what these 'inner' episodes of thought really are.

They are the functionally characterized states of a thinking being that serve as the primary bearers of meaning. They are, on Sellars' view, and according to the LTT, the rule-governed linguistic tokenings produced not primarily as instances of public behavior, but rather as episodes of conceptual representing. That such episodes may be publicly observable is neither here nor there. It is their function as the primary bearers of meaning, as episodes of conceptual representing, and as the causes of meaningful overt linguistic behavior, that makes them 'mental' tokenings of linguistic symbols—in whatever medium they might occur. Thus, 'inner' and 'covert' in our discussion of linguistic tokenings should not be taken as implying a certain location; rather they should be taken as distinguishing such tokenings from those that rely upon such 'inner' tokenings both causally and in terms of where they get their intentionality.

With this idea now on the table, let's return to the question of the relationship between such covert (or mental) and overt linguistic tokenings. To begin, consider the following questions: Are all instances of rule-governed linguistic behavior instances of thinking? Isn't there sometimes a difference, say, between thinking something and telling someone about it? I might think that I would like to see a movie tonight; upon arriving home I might then tell my wife that I would like to see a movie tonight. Thinking-out-loud is one thing, but much of the time people really do use language to communicate their thoughts to others. Sometimes an utterance may even be serving both purposes at once, i.e., it may be both a thinking-out-loud *and* intended to communicate that thought to others. Yet, when I use speech to communicate a thought to others I must have had that thought in the first place.

In these cases it seems likely that we are going to look to the 'inner' token as the true thought episode. What we have just seen, however, is that referring to the thought as an 'inner' episode can generate needless confusion. There is nothing special about the thought being 'inner' or covert; what matters is simply that the thought is the primary bearer of meaning and the cause of the utterance.

As an empirical matter, I think it is almost certainly the case that all overt linguistic behavior, whether it is the meaningfully produced tokenings of the mature speaker or the mere parroting of the learner, is preceded by neurophysiological events in the brain of the person whose behavior it is. But the fact that such events occur in the brain is, we have now seen, of little consequence. Given what we have just said, sometimes it will be appropriate to identify the neural event with the 'inner' thought episode (the mental tokening of a representation) that is causally responsible for the overt

utterance; other times it will be more appropriate to say that the utterance is itself an episode of thinking (even if it, too, is preceded by a neural event); and in the language learner, of course, neither the covert neural event nor the overt utterance will count as a thought. Everything depends upon what token, if any, is primarily possessed of intentional content.

Finally, then, we can return to the interesting question of how one learns to think. To the extent that we have an understanding of how one learns to speak a language (and I have acknowledged that we do not yet fully understand this process, even from a philosophical standpoint), we can also begin to understand how one can learn to think. We train would-be speakers into mature language users by conditioning their overt linguistic behavior to conform to the rules of our language. According to the LTT, the covert tokenings that accompany these overt utterances are just another form of the very same linguistic symbols that figure into the overt behavior. Given this, it seems to me fairly unproblematic to suppose that the conditioning of the overt behavior is at the same time a conditioning of the inner tokenings. And we saw in Section 4.1, with the difference between Jane and her dog, that this is just the sort of situation required to produce conceptual representers: unless the conditioning of the overt behavior is simultaneously the conditioning of the inner representations, no linguistic form of representationalism that respects the normativity of thoughts is going to get off the ground.

We can thus paint the following picture of how a pre-linguistic child is brought into the space of reason. Start at the point where the child has begun the transition from language learner to language speaker. We are imagining that the child has begun,

perhaps only very tenuously and intermittently, to 'get the hang of it.' Yet since the transition from learner to speaker has begun, at least some of the child's utterances are meaningfully produced. This could not be the case, however, unless those utterances were accompanied by thought episodes. Now, if the conditioning of the overt linguistic behavior is also a simultaneous conditioning of the covert linguistic tokenings, it seems reasonable to suppose that to the extent that the child has begun to get the hang of producing overt linguistic tokens she has also begun to get the hang of producing the corresponding inner tokens (though this does not require the supposition that she is 'aware' of those inner tokenings—the reason for this will be made clear in a moment). Take a given utterance meaningfully produced by this child, then. We might suppose that the overt linguistic behavior is causally produced by an inner linguistic tokening. But it may seem more accurate to think of the overt utterance and the covert tokening as occurring more-or-less simultaneously. The significance of this point is quite simple: the child that we are imagining is only just beginning to be a thinker. We can therefore suppose that all of her meaningful overt linguistic behavior is of the thinking-out-loud kind. That each instance of such behavior is accompanied by a corresponding covert tokening will be theoretically relevant in the long run, but is not overly important at this stage. To put the matter simply, I am imagining that the child, while beginning to get the hang of producing meaningful linguistic tokens, has not yet learned to produce the covert tokens without their overt manifestations. In essence, she literally cannot think without speaking.

Over time, though, this child will become more and more a master of the language, and eventually she will begin to be trained to 'keep her thoughts to herself.' As

this happens there will be a shift, in the sense that the child will begin to have thoughts without producing any overt linguistic behavior at all. Her 'inner' thoughts will, more and more, remain covert (and so, 'inner' in that more traditional sense). The silent nature of thoughts that one keeps to oneself, however, is no more important to their role as thoughts than is the fact that the child began to learn the language through the medium of speech rather than hand-signing.[127] If we then go back into this story and ask, "When did the child learn to think?," the answer is that she was becoming a thinker at the very same time that she was becoming a speaker. Her ability to meaningfully produce linguistic tokens was simultaneously the ability to have thoughts, for thoughts according to the LTT are constituted by tokens of the language that one speaks.

It is this last point that most strongly suggests the topic that I will take up in the next section: that thought, as envisioned by the Linguistic Theory of Thought, is fundamentally a *social* affair. Not only is the ability to think developed as, and only when, one learns to speak a language, but the thoughts themselves, as tokens of the symbols of the natural language(s) one speaks, are ultimately rooted in social practice. The very rules that govern the use of those symbols, and hence allow them to be meaningfully produced, are grounded in social practice. What I will do now is explore this idea more carefully and fully, for I do think it is a fairly radical suggestion.

5.3 Thought as Social Practice

The traditional view of the relationship between thought and language has an obvious feature that we have, nevertheless, said very little about directly. I have in mind

[127] Incidentally, I think we can further suppose that the child will actually still think-out-loud some of the time (as indeed, perhaps, do we all; sometimes I really don't know what I think until I say it).

the fact that, on the traditional view, there is a sharp divide between the private nature of thoughts and the public nature of language use. We could perhaps even say that on the traditional view the privacy of thoughts and the publicity of language use are essential features of each. Thoughts are supposedly inner episodes, occurring in the mind, and accessible only to the one whose thoughts they are. Language, on the other hand, is a social convention, a public practice that allows us, otherwise solitary thinkers, to share with each other what we are thinking. This has been, and still is, the most common conception of both the phenomenon of thinking and using language. Indeed, many, if not most, contemporary cognitive scientists seem to endorse the view that language function is just one module of the mind, in no way essential to the mind's other functions. This isn't to say that the ability to use language isn't recognized as an amazing feature of human behavior, but most theorists assign no special role to language use in explaining human cognition. (Whether these same theorists take thoughts to be essentially private is another matter. No doubt some do; but on this point I think the contemporary playing field is more divided. Certainly there are many representationalists, in large part influenced by advances in neuroscience, who entertain the possibility that we might one day be able to 'read' people's thoughts by scanning their brains.)

There are, of course, related issues about thought and language that we could raise. For example: (a) Do thinkers have immediate access to every thought that they have, or are some thoughts 'subconscious' and inaccessible? (b) Is a thinker's access to her own thoughts infallible? (c) Is language purely a matter of convention, or are there parts of it not subject to the whims of society? These are all interesting questions, to be

sure, but I will not address them here. I mention them simply so that we can avoid the temptation to digress further into these related topics.

My concern here is the more obvious and more fundamental one. Simply put, Is it true that thoughts are essentially private, and language use essentially public? On the traditional view of the relationship between thought and language, language is merely a tool for the public expression of otherwise private thoughts. As I have made abundantly clear many times over, this is a view that I reject. Yet doing so does not, necessarily, entail a rejection of either the claim that thoughts are by their very nature private episodes, nor the claim that language use is a matter of public practice. With all that has been said so far, it is abundantly obvious that I take language use to be a matter of public practice. In saying this, however, as we have seen, I am not endorsing the view that language is merely a tool for making thoughts public. On the other side, though I have been making the case for a strong connection between thought and public practice, and am concerned in this section to argue that thought is fundamentally a social affair, there is also a clear sense in which I would agree that thoughts are, or at least can be, private.

On the more traditional view what the privacy of thoughts seems to come down to is something like epistemic access. I, and I alone, can know what I'm thinking. I enjoy a certain privileged access to my own thoughts, as you do to yours. Independent of questions of how I gain this privileged access to my own thoughts, whether such access is infallible, and whether I have this access to all, or only some, of my thoughts, the basic idea of privacy is again just that while I can know what I'm thinking, there is no way for anyone else to directly observe my thoughts. Of course, that's where language is supposed to come in. If no one can directly observe my thoughts, then if I want those

thoughts to be known by others I need some sort of way to make them known. The most effective method of making my thoughts known to others is to simply tell them what I'm thinking. There are other methods, of course (e.g., engaging in this or that (non-linguistic) behavior; facial expressions; things like that), but language is by far the most powerful and efficient method for making my thoughts known. Yet since, on the traditional view, language is not taken to play any role in making thoughts possible, it is an entirely public matter—i.e., it is a social convention created (though this needn't, and probably shouldn't, be taken as implying intent on the part of early, pre-linguistic thinkers) to allow us to communicate our private thoughts.

So before we continue I should probably make a couple of brief comments on where I stand with respect to the issues of privacy and public practice as they figure in the above picture. I reject the traditionalist picture of the relationship between thought and language, to be sure. Nevertheless, there is a sense, as I mentioned earlier, in which I would agree that thoughts have a certain amount of privacy to them. While I obviously do not agree with the idea that thoughts are independent of language use, I am willing to grant that they may not be publicly observable—and, so, it may be that I have a certain privileged access to my thoughts after all. As we will see below, this isn't going to be much of a concession to the traditionalist, though, once I flesh out the rest of the story. Furthermore, I should add that I do not see privacy as in any way essential to thoughts. I don't want to rule out the possibility, that is, that it might actually one day be possible for other people to directly observe my thoughts. (There has been a suggestion floating around in philosophy of mind for some time now that neuroscience may open up the possibility of 'brain reading'—being able to read someone's thoughts by observing their

neural activity—and I see no reason to reject this possibility out of hand.) So, though I am happy to grant that one's thoughts (at least when one is keeping silent, i.e., not thinking-out-loud) may have a certain sort of privacy, I do not see this privacy as an essential element of thoughts. Of course, on the other issue, it should be clear, as I've said, that I take language use to be a thoroughly public matter.

In fact, given my endorsement of the verbal behaviorist framework, and of a functional role theory of meaning, it should be clear that I take language use to be a matter of public practice right down to its very core. The traditionalist may see language as a societal convention for communicating thoughts, but he also believes that the intentional content of linguistic tokens derives from the prior meaningfulness of the thoughts those linguistic tokens express. I, on the other hand, see the intentionality of language as arising from the public practice itself. So, according to the LTT, language is a social phenomenon both in the fact that linguistic tokens are first and foremost public events, and in the sense that the intentionality of these events is determined by social practice.[128]

Much of this has been made clear before, mainly in Chapters Three and Four. Building upon those chapters and the preceding sections of this chapter, I now want to make a case for the idea that thought itself, the very core of our cognitive lives, is a social phenomenon. Let me be clear about what I mean by this, to head off an easy misunderstanding. I do not mean, in calling thought a social phenomenon, that even

---

[128] This is true, of course, even in those cases in which we might want to say that the semantic content of an overt linguistic token is derived from the content of the covert linguistic token (the thought) that caused it, since the intentionality of the thought itself is determined by social practice (a point that I will develop more thoroughly in just a moment).

covert thought episodes are publicly observable.[129] I don't want to rule that possibility out, as I said above, but that is not primarily what I mean. Rather, what I mean is that the phenomenon of thought is one that arises only within linguistic communities, and one that is completely governed by social practice. I believe that taking the public practice of language use as foundational to thought carries with it a very strong sense in which having thoughts is fundamentally a social affair.

Let me clarify this point a little. There are at least two things that I mean when I call thought a social phenomenon. The first is the more obvious, and should by now be quite clear: since thoughts are constituted (on my view) by symbols of natural language, in order to have thoughts at all one must be a practiced member of a linguistic community. That, in a (perhaps somewhat weak) sense, makes thought a social phenomenon—it only arises in language-using creatures, and only in virtue of their public linguistic practice. There is much about the LTT that surely strikes many philosophers as at least potentially problematic, and there are undoubtedly very complex issues that will need to be dealt with if the LTT is to be developed beyond the basic framework I'm creating in this project. However, once the central thesis of the LTT is accepted, I think that the further point that, on such a view, thought is a social phenomenon in this first sense follows rather straightforwardly. In fact, it may seem like little more than a restatement of the central thesis of the LTT.

There is, though, a second, stronger, and perhaps more interesting sense in which thought is, on my view, a social phenomenon. Put bluntly: If one takes the LTT seriously, then what one thinks—what one *can* think—is socially determined. That is,

[129] Overt thought episodes—that is, episodes of thinking-out-loud—are, of course, always publicly observable.

the very contents of our thoughts are determined by the practices of the linguistic community to which we belong. While I think that this will strike many people as a particularly radical thesis,[130] it really follows fairly clearly from what we've already said. Consider: thoughts are constituted by symbols of one's natural language; those symbols get their meaning from the role that they play in the public practice of language use; hence the meaning of one's thoughts is a matter of the role played by the linguistic symbols that constitute those thoughts in the public practice of language use. That is, since linguistic meaning is determined by social practice, and thoughts are just tokens of those same linguistic symbols, then the meanings of one's thoughts are determined by social practice.

The conceptualized world that one lives in is, by these lights, a social construct. I realize that this will offend some philosophers' realist sensibilities, since I am, to a certain extent, making truth and reality relative to one's conceptual scheme. I am not, however, questioning the brute existence of the world; I am merely suggesting that, to borrow a phrase from Quine, how one 'slices' that world is determined by the linguistic community to which one belongs. Consider the following quote from Quine's "Ontological Relativity":

> [B]egin by picturing us at home in our language, with all its predicates and auxiliary devices.... In these terms we can say in so many words that this is a formula and that a number, this a rabbit and that a rabbit part, this and that the same rabbit, and this and that different parts. *In just those words.* This network of terms and predicates and auxiliary devices is, in relativity

[130] I do not claim that this thesis is new, however. My theory of thought, which leads me to make this claim, has new elements, I think, or at least a new synthesis to it. But the idea that what one thinks is determined by one's language has been suggested before; for example, in the Sapir-Whorf hypothesis—what linguists and psychologists now commonly call the thesis of 'linguistic relativity' (see, e.g., Lera Boroditsky (2003), Steven Pinker (1994, p.46)). I also think this idea is fundamental to Quine's naturalized epistemology (from which, in turn, the suggestion of some feminist epistemology that knowledge is social or communal arises).

> jargon, our frame of reference, or coordinate system. Relative to *it* we can and do talk meaningfully and distinctively of rabbits and parts, numbers and formulas. (1969, p.48).

If we replace all references to speaking in the above quote with references to thinking instead, we will have a fairly clear example of the consequences of treating thought as a social phenomenon. Our thoughts have determinate content, but that content is determined by social practice—that practice provides, in Quine's terms, the coordinate system within which thought is possible and meaningful. The traditional view has treated thoughts as more-or-less independent of societal factors. Thinking has always been regarded as an individual, solitary activity—something that one can engage in regardless of whether one is a member of society or a solitary human (or other animal) alone in the wilderness.

My suggestion, though—and, since it is a clear consequence of the LTT, it is rather more than a mere suggestion; it is part and parcel with the theory of thought that I am laying the foundation for—my suggestion is that thinking is *not* an individual, nor solitary, activity. Thinking is an inherently social activity; conceptualizing the world around one requires membership in a linguistic community, a community that supplies the representational system that allows one to *think* that this is a rabbit, this and that rabbit parts, etc., in the first place.[131] Again, this is not to say that thought episodes must be publicly observable. Rather I am saying that whether or not such episodes are public, their occurrence, and their content, are determined by a particular social practice, *viz.*, the practice of using language.

---

[131] Incidentally, I have long held that Quine endorses, or would endorse, this characterization of thoughts. Quine probably would not like my commitment to realism about the propositional attitudes, though, so it's right that I acknowledge here that Quine and I would have our differences. Nevertheless, I have always felt more in league with Quine than not.

My final task in this chapter is a kind of 'summing up' if you will. We have now seen all the elements that make up the framework of the LTT, and have seen how they fit together with one another to form a coherent whole. I do not, of course, claim that this picture is complete—far from it. There are many issues yet unexplored, and many of the issues that we have talked about here could use further discussion. My goal in this project is not, however, to offer a definitive and complete argument for the LTT. As I said at the outset, my goal is simply to describe the basic framework of the theory and show that there is a place for it amongst the myriad contemporary theories of mind. Along the way I have offered a few compelling (to my mind) arguments in favor of pursuing the LTT to a fuller degree. Below, then, I will take stock of what we have done so far before turning to consideration (in Chapter Six) of a few of the more significant objections to the LTT.

5.4 Sellarsian Representationalism for the 21st Century

One might reasonably wonder at this point, given my stated goal of finding a place for the LTT amongst its representationalist peers, whether I have really made a case for seeing the LTT as a form of contemporary representationalism, or instead advocated a radical discarding of all such views in favor of something quite different. One might wonder, that is, whether the LTT really shares all that much in common with the views of Fodor, Lycan, Millikan, Cummins, Block, et al. I have said that I take myself to be updating contemporary representationalism with the insights of a Sellarsian approach to language and thought. Yet I have also said that I find all the usual suspects, the various contemporary versions of RTM, severely lacking in certain regards; and there can also be

no denying that the LTT is in many ways a radical departure from the bulk of contemporary representationalism.

Perhaps, then, I should say something about the relationship of the LTT to what I do, in fact, consider its representationalist peers. To begin with, let me list some of the things that I believe the LTT shares with other representationalist views. First there are the obvious similarities. Like the theories of Fodor and Dretske, Cummins and Block, etc., the LTT treats thoughts as the mental tokenings of symbolic representations. Like those theories one of the most distinguishing features of the LTT is its approach to the development of a psychosemantics (recall that it is different psychosemantic theories that primarily distinguish one representationalist view from another); in fact, the psychosemantics of the LTT is at root a functional role theory, placing it fairly neatly alongside views like those of Block or Cummins. As a representationalist theory, the LTT is basically a version of functionalism (that large class of theories of mind that has been dominant for decades now, ever since the fall of the (short-lived) identity theory). Also, I am, as are all contemporary representationalists, a physicalist, which is primarily to say that we representationalists are not dualists of any stripe.[132]

Less obviously, though no less importantly, the LTT embraces, as does its representationalist peers, a certain degree of naturalism about minds. I agree with Fodor that beliefs and intentions are not going to be found among the fundamental properties of the universe. Though, of course, the issue of naturalism is also the locus of one of the

---

[132] I have avoided any discussion of physicalism *per se* primarily because I have been speaking for the most part to other representationalists—and, hence, in most if not all cases, other physicalists. There are a great number of issues that arise when one begins to probe just what a given philosopher's physicalism comes to: issues such as the qualia problem, supervenience, token identity, and more. These issues need to be discussed at some point, if the LTT is to become part of a larger theory of mind, but I believe that they can safely be ignored here since my primary audience is already committed to physicalism anyway.

largest differences between the LTT and other representationalist views (as I argued in Section 4.1). While most representationalists have taken adherence to naturalism to require us to treat those mental representations that count as thoughts as in no way fundamentally different from other forms of representation (the unity and continuum claims), I have argued that thoughts must, in fact, be treated as an entirely distinct form of representation. They are distinct in that they are norm-governed.

What I am trying to do here is find a middle ground between a Platonistic conception of the norms that govern thinking as 'free-floating,' 'universal' rules, and a hard-core naturalism that would reject any appeal to normativity (as an essential feature of thoughts) as being unscientific and non-natural. I believe that the normativity that arises out of the social practice of language use, and that helps make thoughts the unique kind of representational states that they are, is compatible with a commitment to naturalism, and with a denial that the propositional attitudes are irreducible, fundamental features of the universe.

Having said that, however, I must reiterate that while the normativity that I take to be essential to thought does not, in my view, conflict with my naturalist commitments, it is, nevertheless, of great importance. I suppose that I should also note that since the normativity of thoughts according to the LTT arises, as we have now seen, from social practice, my naturalism obviously embraces such practices. This is probably a weaker sense of normativity than Sellars would like, and it may be a weaker form of naturalism than some representationalists would like, but though I occupy this middle ground I see no reason to suppose that this ground is unstable.

At any rate, the naturalism of the LTT is a significant factor in tying it to other representationalist views, primarily for the following reason. I have argued that there must be a sharp divide between those representations that count as conceptual (thoughts) and those that do not. I have also fleshed out how I take conceptual representing to work. But what about the other side? What about all those representations, in non-linguistic animals and pre-linguistic humans, that do not count as conceptual? How do we explain them? Perhaps the largest single reason I have for viewing the LTT as a development of, a continuation of, other representationalist theories is that I think such theories have largely been on the *right* track when it comes to explaining the mental states of non-linguistic animals.[133] I do not, therefore, wish to jettison such views; on the contrary, it is obvious to me that one or more of the representationalist precursors to the LTT, or some development thereof, will be necessary for a complete theory of mind (remember that the LTT is a theory of *thought*, and not all mental states are thoughts, i.e., conceptual representings).

Lest it sound as though I am relegating other representationalist views to some sort of minor, background role, let me hasten to add that though I take the LTT to be the proper way to approach a theory of thought, there is nothing 'minor' about a theory of non-conceptual mental representation. Indeed, the LTT is a part of the larger picture of RTM, and, as we will see to some extent in the next chapter, there will be significant

[133] For the record, I find myself often inclining toward a teleological theory, like Millikan's biosemantic approach, when I think about non-conceptual mental representations. Though since I embrace a functional role theory for conceptual representations, one might expect that I would do the same in the non-conceptual cases. I admit that there is some attraction to functional role theories in the non-conceptual case as well. Yet I can't help feeling that at least some parts of the teleological picture are important—particularly the focus on *consumers* rather than producers of representations. Perhaps some combination of teleology and functional role would prove most efficacious. Since the issue of how to deal with non-conceptual mental representations lies largely outside the confines of this project, though, I do not feel the need to delve into the topic in any serious manner here.

interplay between those representational abilities covered by the LTT and those representational abilities covered by some version of one or more of the theories from which the LTT takes its initial cue.

I do not say all of this in order to attempt to placate those representationalists with whom I am disagreeing (as though it would), nor in some misguided attempt to sound magnanimous. I say this because I want to be clear that I really do see the LTT as arising out of the field of contemporary representationalism. The insights of contemporary representationalists, with their emphasis on the symbolic nature of mental representations and with their adherence to naturalism, are important (in my view) for the development of a proper and successful theory of mind. Yet I also believe that the insights concerning language and thought provided by Sellars have been largely overlooked by modern representationalists. I have therefore tried both to show that these insights, e.g., concerning the normative nature of thinking, are important, and how they can be integrated into the representationalist picture (primarily by turning to a Sellarsian psychosemantics and, for my own part, turning away from Mentalese and the like to embrace a natural language theory of thinking).

This, ultimately, is what this project is all about. Though (outside of Chapter One) I have been debating exclusively with other representationalists, my aim all along has been, not to undermine representationalism, but to endorse it and take it in what I think is the right direction. I have been building on the foundation of modern representationalism using the insights of Sellars. Or you could say that I have been trying to take Sellars' work and update it with the insights of modern representationalism. Either way of putting the point is equally true.

So we see that if one takes what I have done so far in the way in which it was meant to be taken, it really does amount (assuming one sees the project as successful) to finding a place for the LTT amongst its representationalist peers after all. What I have argued is that if representationalism is to succeed as a theory of conceptual representation, i.e., as a theory of thinking, in addition to a theory of non-conceptual mental states, then we must recognize that thoughts are fundamentally *different* from other forms of mental representation. But we can acknowledge and accommodate this fundamental difference without abandoning the basic framework of, and arguments for, the representational theory of mind. This, in essence, is what I claim to have shown in this project. Many questions and problems remain, and all that we have so far is the barest of frameworks for a successful theory of thinking, but that is a start, and in my view, a move in the right direction.

In the next, and final, chapter I will tie up a couple of loose ends, delving into more detail about the differences between the LTT and a few views that might seem most like it in certain respects. I will also raise and reply to a couple of the more immediate objections that one might have to the theory that I have now proposed.

6. On Consciousness, Sensations, and Competing Views – A Brief Survey

This Chapter is a bit of a hodge-podge of issues and material. Now that we have the LTT laid out as clearly and thoroughly as we can in the space we have here, I want to consider a handful of issues that might arise when one tries to evaluate the possibility of success for that view. To this end, I engage in a number different tasks in this Chapter. First of all, if I have cleared the space for the LTT among most of its representationalists peers, as I have intended to do, there are still some other views that warrant special attention. In sections 6.11 and 6.12 below I will address these views. First there is a view put forward by Peter Carruthers that might seem, at first glance, to be very much like the LTT—so much so, in fact, that unless one examines it more closely one could question whether I have really added anything to the field by elaborating my own view here. As we will see, however, Carruthers' view is quite different from the LTT. On the other hand, there are the views of other followers of Sellars, and it would behoove me to say something about how they would view my project.

The next section will deal with a variety of issues surrounding consciousness and sensations. Since consciousness is a subject that many people today write entire books on, I do not claim that my discussion here is anything but a small start. I focus my discussion, however, by looking at an essay by Ned Block. The main points of that section are to provide some reply to those who find the consequence of the LTT that non-linguistic animals and infants do not think unappealing, or even absurd, and to show how, though I have spent most of this project talking about the differences between the thinkers and the non-thinkers, there are also many similarities and continuities between them.

There are, of course, many potential objections to the LTT that I am not going to be able to cover in this Chapter, nor in this project. That should not surprise us, however, as my development of the LTT is an attempt to develop something of a new paradigm in philosophy of mind: I am in many ways proposing an entire theory of mind, and I cannot hope to defend such an attempt against all foes in the space I have here. As I have said since the beginning, my goal is the more modest one of providing the overall skeleton of the LTT, and carving out its place among surrounding views. And to that end, we now consider some more of those views themselves.

6.1 Competing Views

One hopes, in philosophy, to contribute something new, to move the conversation forward in some way, be it a small step or giant leap. To many, the LTT may seem like an attempt at a giant leap, for it is a radical departure from much of the history of philosophy of mind. Or, perhaps, it will seem like an attempt to make a leap already attempted by others (and an attempt that ended in failure at that). To such people, I have already said all that I can say here. There are others, however, to whom the LTT will seem like a more modest attempt to push forward an idea the fundamentals of which they have already embraced. I have in mind not only the representationalists that I have been talking with throughout this project (Fodor, Lycan, Millikan, Cummins, etc.), but others as well. And if I have said enough to distinguish the LTT from the views of Fodor, etc., I have not yet addressed those (aside, of course, from Sellars) whose views are much more like my own.

In this section, then, I wish to talk about some of these philosophers, and how the LTT differs from their views. Though my main intention here is simply to mark out the distinctions between the LTT and these other views, in line with this project's main goal of clearing the space for the LTT, I do hope to provide some arguments meant to persuade people to see the LTT as a more promising line of development in the conversation about the nature of thought.

I will first address in some detail a view that, on its surface, can look very much as though it just is what I'm calling the LTT. This is the natural language theory of thought proposed by Peter Carruthers in his book *Language, Thought, and Consciousness*. I will show that though there are superficial resemblances between Carruthers' view and my own, his theory is ultimately a confusing attempt to graft what I take to be the right way to approach thoughts onto a clearly Fodorian Mentalese picture. That is, Carruthers seems to me to be basically an adherent of Fodor's Language of Thought hypothesis, and his inclusion of an appeal to natural languages as the languages of thought in some cases is basically only a half-step in the right direction. (As we will see, I think Carruthers really is moving in the right direction, and he has addressed many issues (and critics) of the natural language view of thought so well that I would gladly cite his arguments for my own purposes. Again, I just think that Carruthers hasn't gone far enough.)

Following that I will turn to a brief consideration of how things look from a Sellarsian standpoint, i.e., from the point of view of someone who is engaged with the philosophy of Sellars, but who does not necessarily share my interest in adapting Sellars' view to a modern, physicalist, representationalist picture. An important part of my

project has been, and is, to find a sort of middle ground with respect to questions of naturalism and normativity. The relation between these two ideas is also an important part of Sellars' own work, and of subsequent scholarship regarding Sellars. In 6.12 I will use an essay by James O'Shea to guide a discussion of this essential issue. Most Sellarsians today seem to have taken one of two paths: either they embrace Sellars' naturalism while rejecting his insistence upon the importance of the normative, or they heartily agree with Sellars about the importance of the normative, but they remain skeptical that this can be reconciled with naturalism. If O'Shea is right, Sellars himself would reject both of these approaches to his work, insisting that each part has an essential role, and that they are compatible with one another. Regardless, I will use this opportunity to argue, somewhat indirectly, with my fellow Sellarsians at both extremes, for the LTT also treads the middle ground between naturalism and normativity.

6.11 Carruthers' Natural Language Thesis

Stripped to its bare essentials, the thesis of Peter Carruthers' book *Language, Thought, and Consciousness* is that all (at least human) conscious thought takes place in natural language. The key word here, for Carruthers, is 'conscious.' That is, while Carruthers puts forth arguments for a natural language theory of thought that, at first glance appears to be very much like the LTT, he qualifies his theory with the restriction that it only apply to *conscious* thoughts.[134] A huge portion of his book, then, is spent

[134] I offer here a quick note of clarification. Though Carruthers is never very explicit about this point, it is quite clear from what he says throughout his book that what he means by 'conscious thoughts' are thoughts the thinker is aware of having—call them 'self-conscious thoughts.' Only if one reads Carruthers this way can one make sense of his notion of 'non-conscious thoughts'—for such thoughts are clearly not thoughts in unconscious beings (which would make no sense), but rather thoughts that one is not consciously aware of having. These are the unreflective thoughts that Carruthers ascribes to animals, infants, and (presumably) the 'sub-consciousness' of adult human beings.

talking about consciousness. I will have a little to say about the topic of consciousness in Section 6.2 below, though I can hardly claim to have more than simply scratched the surface of the issues surrounding philosophical debates about consciousness, and in no way do I attempt to argue explicitly, or even implicitly, with Carruthers on this subject. To a certain extent, though, what I (or Carruthers) have to say about consciousness is irrelevant to the project of distinguishing the LTT from Carruthers' view; it is also irrelevant to my main criticisms of Carruthers' thesis. This is so because my debate with Carruthers has nothing to do with the details of what consciousness is or how it works, but rather with his restriction of the natural language thesis to only a subset of thought episodes, rather to thought generally.

In many ways, Carruthers and I have the same sort of approach to both the philosophy of mind and (especially) language. With regard to the latter, Carruthers argues, as I do, that language is not merely a tool for expressing thoughts (the view that I have been referring to as the traditional view, and that Carruthers calls 'the communicative conception of language' (1996, p.1)). Rather, he sees language as fundamental to thought (a view that he labels 'the cognitive conception of language' (p.2)).[135] But, as mentioned, he sees language as fundamental not to all thought, but only to conscious thought.[136] Cutting to the chase, on Carruthers' view animals, infants,

---

[135] Interestingly enough, Carruthers is also sympathetic (as so few people these days seem to be) to the Sapir-Whorf hypothesis of linguistic relativity (the thesis that Steven Pinker so vehemently attacks in his book *The Language Instinct*). Properly understood, I too could be considered sympathetic to the Sapir-Whorf hypothesis, though what one takes that hypothesis to be seems to vary from theorist to theorist (e.g., Pinker takes it to be an extreme thesis with what could only be called racist overtones; psychologist Lara Boroditsky takes the thesis to be much more moderate, though limited in effect; Carruthers seems to treat it as just another way of stating the natural language thesis; and I see it as an early recognition of one consequence of the LTT's picture of the relationship between thought and language).

[136] It is also of some consequence to notice what Carruthers says on p.215: "[M]y claim is certainly not that language is prior to thought! (That would be absurd. To be a language-user you have to be a thinker too.)"

children raised by wolves, etc., do have thoughts—it's just that their thoughts are not conscious thoughts. Carruthers is prepared to say that these non-conscious thoughts take place in Mentalese, on the Fodor model, though he also claims to be ultimately 'agnostic' about this (see p.66).

Thus it may at first appear that there is little more to say about the difference between the LTT and Carruthers' natural language thesis. I ascribe thoughts only to speakers, and argue that they are tokens of natural language sentences; Carruthers ascribes thoughts to non-speakers as well, but distinguishes between the conscious thoughts of speakers, which are tokens of natural language sentences, and non-conscious thoughts in whatever beings have them (and this would include any non-conscious thoughts in creatures that do possess language), which may be tokens of Mentalese sentences. I am not satisfied to leave the discussion at that, however, for two reasons. First of all, I think that there is something to be gained by looking more closely at the details of this difference between Carruthers' view and mine. Secondly, I want to say at least a few things about why I think one should prefer the LTT to the type of view that Carruthers espouses (since I do, after all, take Carruthers to be charting roughly the same sort of course that I am charting, even if he ultimately, as it seems to me, veers back into 'safer'—more well sailed—waters by the end).

Let's start, then, by looking more closely at the nature of the difference between Carruthers' view and the LTT. There are some aspects to this difference that I am going

---

As we have seen, 'priority' talk when discussing the relationship between language and thought can be complicated, and I don't think that Carruthers' denial, here, of the priority of language is quite as sharply opposed to the attitude I take as it first appears. But his denial of the basic idea is so forceful (notice the exclamation point and the use of 'absurd') that I take it there are some pretty basic ideas about which Carruthers and I would find ourselves in disagreement.

to avoid really saying anything about: *viz.*, anything that would require us to go into a detail about various theories of consciousness. This shouldn't hamper us too much, though, since as I've said, most of my important disagreements with Carruthers are relatively independent of what one has to say about consciousness. The most important difference, I think, between Carruthers' view and my own is the centrality of the role that each assigns to natural language. We've seen that the LTT, taking its cue from Sellars, sees natural language as necessary for the development of thought—one cannot be a thinker without being a participant in a public linguistic practice. Obviously, on Carruthers' view this is not the case. He does (see, e.g., Section 4.6 of *Language, Thought, and Consciousness*) want to claim that there are some thoughts that cannot be had without possession of a natural language, but he has also said that the idea that animals and infants don't think is so 'implausible' that he won't even consider it (p.65). In and of itself this is relatively unremarkable: most of the people that I have been discussing and arguing against throughout this project believe that non-linguistic beings can and do have thoughts. This part of Carruthers' view becomes important, though, when we realize what it means for language's role in the development of thought.

We should perhaps remind ourselves at this point that the claim that the development of thought in any being whatsoever relies on participation in a language community is not one that is tangential to the LTT; rather, that claim lies at the very heart of the theory. Thoughts, on my view, are constituted by natural language symbols which are meaningful in virtue of the role that they play in the linguistic community of which they are a part. As we have seen, then, that means that participation in a linguistic community is a prerequisite for acquiring the ability to think. Not only that, but it is the

public practice of language use that gives thoughts their most essential feature: their norm-governed nature. The rules that govern thought arise from the practice of using language.

Whatever Carruthers might say about the use of natural language in having a certain subclass of thoughts, if he allows that thinking can take place without natural language—that is, if he allows that non-linguistic beings can have thoughts just like linguistic beings—then he has at best relegated language use to a tangential role in cognition. I do not mean, in calling the role 'tangential,' that it is unimportant, of course; the role that natural language plays in the having of conscious thoughts on Carruthers' view is far from unimportant. Yet language is not central to thought on Carruthers' theory. And there are significant consequences when one puts language to the side like this.

First, we are likely going to end up with a theory that fails to acknowledge the normativity of thought. This is not, of course, necessary: Carruthers could develop an account of the norm-governed nature of thoughts that had nothing to do with the practice of using natural language. In point of fact, however, Carruthers, like Fodor, Dretske, Millikan, etc., before him appears to be a hard-core naturalist and thus seems to be embracing the unity and continuum claims (as discussed above in Section 4.1). Granted, he is not embracing those claims quite as whole-heartedly as these other theorists, for he *is* trying to make room for a special class of representations which are constituted by natural language symbols. Nevertheless, I have not found him anywhere attempting to acknowledge the normative nature of thoughts, and so even if his account doesn't

preclude developing an account of thought's rule-governed essence, Carruthers never actually offers one.

This leads us to the second consequence of denying language a central role in thought. Even if Carruthers is going to insist that there are differences between conscious thoughts, which take place in natural language, and non-conscious thoughts, which (perhaps) take place in Mentalese, he is not going to be able to deny that many of these 'different' kinds of thoughts will be at root the same. Consider the following example. I will have the conscious thought, "There's a food bowl"; my dog will have the non-conscious thought, "There's a food bowl." Being conscious, my thought (on Carruthers' view) takes place in natural language; being non-conscious, my dog's thought takes place in Mentalese (let's suppose, putting Carruthers' agnosticism aside for the time being). Yet, it would seem that my dog and I are really having the same thought, with the same content. It is therefore unclear just what purpose natural language serves when it comes to thinking. Even if we grant that there are certain thoughts (about, I don't know, the mystery of the value of pi) that can only be had by language-using creatures, most of the thoughts that we have every day are going to be of the "There's a food bowl" sort, i.e., thoughts that can occur in natural language or Mentalese with no difference in content, functional role, or anything else of substance. This, as I see it, greatly trivializes the natural language thesis. If true, that thesis becomes more an empirical curiosity than an insight into the nature of thought.

Let me elaborate this point a bit, for this is the feature of Carruthers' view that I find most objectionable. Carruthers goes through a great deal of trouble to argue against critics, in the cognitive sciences, of the natural language thesis. Yet many of these critics

(e.g., psychologist Steven Pinker, or Jerry Fodor) are actually *fans* of a language of thought hypothesis.[137] Being at least something of a fan of Mentalese himself, Carruthers must therefore find some element of the natural language thesis very compelling—otherwise why go through the trouble of writing an entire book defending that thesis against critics whose view he basically shares? One could argue (and I think there is quite a bit of merit to such a position) that Carruthers' real, albeit less explicitly stated, goal is to say something about consciousness. Plenty of people can endorse the same Mentalese picture while disagreeing with one another over the nature of consciousness. Be that as it may, however, Carruthers does endorse a natural language hypothesis, and he makes that endorsement the central theme, the very purpose, of his book. So it is all the more troubling when it begins to appear as though Carruthers' natural language thesis is so very impoverished.

Again, that Carruthers' view is impoverished seems rather obvious when one considers the role that he assigns to natural language in thought. It is only *conscious* thoughts that occur as tokens of natural language sentences. Non-conscious thoughts, in both non-linguistic and linguistic beings, are tokens of Mentalese. There may be great value to being able to have conscious thoughts (e.g., being able to have thoughts about one's thoughts might seem to require this ability), but there is nothing (on this view) essential to thoughts *per se* in either their being conscious or involving natural language. Being able to have concepts, and thus conceptually rich representations of the world,

---

[137] Though I disagree with him, I have always found it admirable that Fodor has continued to stick to his LOT hypothesis throughout the years, especially since he seems so very well informed about the latest developments in psychology, neuroscience, and computer science (which is to say, his view has managed to incorporate—perhaps even shape—much of this empirical research, and I take that as a plus when dealing with a philosophical theory of mind). I was thus somewhat perversely delighted when Fodor more-or-less asserted the complete denial of my own view within the first five minutes of his 2005 APA Presidential Address.

does not, on Carruthers' view, require access to language. Nor, of course, is being part of a linguistic community necessary for one to have conceptual mental states—and, hence, beliefs, intentions, knowledge, etc. By themselves, of course, these points are not criticisms of Carruthers' view (unless one already endorses something like the LTT)—but, then, they aren't meant to be. Rather, I am simply pointing out that insofar as Carruthers endorses a natural language theory of thought the view that he puts forward is severely limited in scope, applying as it does not to thoughts at their most fundamental, but only to a subset of thoughts.[138]

The upshot, obviously, is that someone like Carruthers is going to need essentially two theories of thought: one for the non-conscious thoughts and another for the conscious thoughts. And so far as I can tell, Carruthers' only real reason for going this route is his desire to avoid what he takes to be the absurd position that animals and infants don't have thoughts like we do. That may seem like a very good reason to many people, but as I have already discussed, accepting the claim that non-linguistic beings do not have thoughts is a bullet that I am willing to chomp down on whole-heartedly.

For our present purposes, then, I think that I have said all I need to about Carruthers' view. While it may appear on the surface to be akin to the LTT, it takes only a brief examination to realize that, despite various similarities, my view is not much closer to Carruthers' than it is to Fodor's. We turn now to what is both an issue of Sellars

---

[138] Of course, I am also bothered, just in general, by any view that seems to unnecessarily (in my view) complicate our theory of thought—in this case by supposing that one and the same thought can occur either in Mentalese or in natural language, thus requiring *two* full-blown types of representational system to account for one and the same phenomenon. (Though I don't think that we can take it entirely at face value—because I don't think that he really believes this without qualification—Carruthers says on p.163 that "all types of mental state admit of both conscious and non-conscious varieties...." And so, on his view, we're going to have to have both Mentalese and natural language accounts of each and every type of mental state—a situation that, I think, we ought to avoid if possible.)

scholarship and an essential part of my own position: the proper way to reconcile naturalism with an endorsement of a non-reductive normativity in the realm of thought.

6.12 Fellow Sellarsians

We saw earlier (Chapter Five) that the Linguistic Theory of Thought is in many ways trying to steer a middle course between an extreme naturalism (that would reject any kind of non-reductive account of normativity) and an equally extreme endorsement of the irreducibility of the normative. While it may be the case that this middle course I am trying to navigate is one that Sellars would reject, it is not obvious that he would do so. His followers, though, seem largely divided by this issue into what James O'Shea calls Sellars' 'right wing' and 'left wing' admirers (2006, draft version)—with the right wing (among whom O'Shea includes such luminaries as Dennett, Lycan, Millikan, and Jay Rosenberg) embracing Sellars' scientific naturalism while rejecting, or remaining skeptical of, his arguments for a "normatively structured 'logical space of reasons'" as irreducible yet compatible with naturalism (O'Shea, 2006); and the left wing (including McDowell and Brandom) taking the reverse position, embracing Sellars' normativity while resisting his scientism. O'Shea himself argues that Sellars' position is truly in the middle here—"naturalism with a normative turn" as he calls it. Thus, in O'Shea I seem to have found something of a compatriot.

O'Shea's main purpose in his "On the Structure of Sellars' Naturalism with a Normative Turn" (2006) is to argue for an interpretation of Sellars' philosophy that reconciles as compatible the conceptual irreducibility of the space of reasons with the causal reducibility of this space from the perspective of the descriptive sciences. The

hardest part of this task, though clearly also the most important part, is preserving the full force of both sides, i.e., taking seriously both the *irreducibility* (in one sense) of the space of reasons and the *reducibility* (in another sense) of that space. This task will fail if we end up with either 'normativity so-called' or a scientism that is incomplete (because it leaves out norm-governed conceptual mental states).

Since my view involves, as I've said, an attempt to stake out a middle ground that recognizes normativity as essential to conceptual representing while at the same time endorsing a full, un-impoverished naturalistic account of human beings in the world, O'Shea's interpretation of Sellars is of particular interest to me. This section of my project is not, however, just an evaluation of one contemporary Sellarsian's interpretation of a particular part of Sellars' philosophy. On the other hand, this section is also not an attempt to briefly survey the vast literature of Sellars' followers, not even on a limited number of topics. Rather, what I want to do here is, in essence, to cast my lot in with those Sellarsians, like O'Shea, who try to steer a middle course on the fundamental issue of normativity versus naturalism, taking Sellars as our guide. It should go without saying that this course is crucial to the success of my own view.

O'Shea takes as the centerpiece of his argument Sellars' essay "Philosophy and the Scientific Image of Man" (PSIM). And of central import from that essay is Sellars' attempt to create a 'stereoscopic' view of human beings in the world that fuses what he calls the 'manifest' and 'scientific' images of our species. It is an oversimplification, though not, I think, an outright mistake to think of the manifest image as the realm of the normative, and the scientific image as the realm of the naturalistic. Hence the stereoscopic view, the blending into one cohesive picture both the manifest and scientific

images, just is a striking of the middle ground sought by O'Shea (and myself) in addressing the naturalism versus normativity question in Sellars' philosophy.

Near the beginning of PSIM, Sellars has a passage that manages to touch on just about every difficult issue we've been dealing with throughout this project. He talks about the idea

> that anything which can properly be called conceptual thinking can occur only within a framework of conceptual thinking in terms of which it can be criticized, supported, refuted, in short evaluated. To be able to think is to be able measure one's thoughts by standards of correctness, of relevance, of evidence. In this sense a diversified conceptual framework is a whole which, however sketchy, is prior to its parts, and cannot be construed as a coming together of parts which are already conceptual in character. The conclusion is difficult to avoid that the transition from pre-conceptual patterns of behaviour to conceptual thinking was a holistic one, a jump to a level of awareness which is irreducibly new, a jump which was the coming into being of man.
>
> There is a profound truth in this conception of a radical difference in level between man and his precursors. The attempt to understand this difference turns out to be part and parcel of the attempt to encompass in one view the two images of man-in-the-world which I have set out describe. For, as we shall see, this difference in levels appears as an irreducible discontinuity in the *manifest* image, but as, in a sense requiring careful analysis, a reducible difference in the *scientific* image. (1963, p.6).

If we pay attention, we see Sellars here raising the issues of the normativity of thinking, of how to account for the transition from non-thinking language learner to thinking speaker, and, of course, of the attempt to find a middle ground that embraces both the reducibility and irreducibility of the space of reasons. With regard to the latter point, we even see Sellars raising the issue of the simultaneous continuity and discontinuity between human beings as thinkers and all the non-thinking animals (a subject that I will take up more explicitly in the next section).

Though the journey to a full understanding of how Sellars creates his sought after stereoscopic vision is arduous, the upshot of O'Shea's reading of Sellars goes something like this. When we look at things like the social practice of language use (and that, of course, is one of the most important things to look at for our purposes here) we find that they have a normative dimension, an element of rule-following, that arises from within that practice and is not logically reducible to uniformities of behavior, or causal relations among objects in a scientifically respectable ontology. The rules that govern linguistic practice (and hence thinking) are real, and cannot be analyzed in purely naturalistic terms. At the same time, such practices are causally reducible to uniformities of behavior and causal relations among objects in a naturalistic ontology. We can, that is, describe *without remainder* a social practice like language use in purely causal, scientific terms—even if, in so doing, we are not offering a logical analysis of such behavior.

Later in his paper, O'Shea quotes the following from Sellars' "Truth and 'Correspondence'":

> I am not claiming [in saying earlier that "Espousal of principles is reflected in uniformities of performance"] that to *follow* a principle, i.e. act on principle, is identical with exhibiting a uniformity of performance that accords with the principle.... I am merely saying that the espousal of a principle or standard, *whatever else it involves*, is characterized by a uniformity of performance. And let it be emphasized that this uniformity, though not the principle of which it is the manifestations, is describable in matter-of-factual terms. (Sellars, 1963, p.216).

O'Shea then comments that while "the normative principle itself is, on the one hand, conceptually irreducible to any ideal scientific explanation of it in causal-naturalistic terms," nevertheless "the patterns or 'uniformities of performance' themselves are in principle describable in purely naturalistic terms, and they are thus explainable *as* the

particularly shaped patterns that they are" (O'Shea, 2006). This point put in terms of our subject matter in this project, then, is that while the norms that govern linguistic practices cannot be reduced to explanations in purely naturalistic terms, the patterns of behavior that manifest those norms can be described in purely naturalistic terms. Hence, the space of reasons is logically irreducible, but causally reducible, to naturalistic explanation.

There is clearly more that could be said about this, particularly in the details, but I will leave it as is for now, for we have at least given ourselves a clear start in discussing this issue of trying to follow Sellars across the middle ground between the extreme naturalism of his 'right-wing' followers, and the extreme 'non-naturalism' of his 'left-wing' followers. I turn now to our final topic: the relations between conceptual representations, non-conceptual representations, sensations, and consciousness—which leads to a consideration of the *continuities* between thinking beings and other conscious creatures (while most of this project has been concerned with pointing out the sharp distinctions between the two).

6.2 Thought, Sensations, and Consciousness

Having now said something about some of those philosophers whose views can end up looking very much like my own on the surface, or who share my Sellarsian proclivities, it remains to say a little something about some much broader issues. What follows here is a necessarily brief discussion of the way that I imagine the LTT to fit into a broader theory of mind. I address this question by looking at the interrelations between thoughts, as we have construed them in this project, and the conscious lives of beings in the world.

We have seen again and again how the LTT posits a stark distinction between thinking beings, on the one hand, and non-thinking beings, on the other. Yet the LTT also posits a great deal of continuity between thinking and non-thinking animals. While the break between conceptual representations and non-conceptual representations is, I have argued, a difference in kind (of representation), there is still a connection between those beings that have conceptual representations and those that do not—if for no other reason than the fact that, according to the LTT, all conceptual representers were at one time non-thinking beings whose representations were all non-conceptual. But there are other reasons as well to suppose a level of continuity between thinking and non-thinking animals. For starters, there is the simple fact that many animals to whom the LTT would deny thoughts regularly exhibit very intelligent behavior in interacting with their environment. Also, we might safely suppose that just as thinking beings have sensations, so too non-thinking animals also have sensations. Finally, there is the fact (and I do take it to be a fact, though as we'll see, this is both a simple and a complicated point) that non-thinking beings as well as thinking beings are one and all conscious.

So, ultimately, I have two overarching goals for this section: (a) to explore the continuities between thinking and non-thinking beings as the LTT embraces them; (b) to go as far as we can at this point toward explaining the intelligent behavior of non-thinking animals in terms of consciousness, sensations, and non-conceptual representations. I will begin with the issue of the difference between sensations and thoughts, particularly in thinking beings. Having said something about what I take sensations to be, and how they figure into the cognitive and behavioral mechanisms of thinking, and to some extent non-thinking, beings, I will proceed to develop a focused

(though necessarily abbreviated) discussion of consciousness. As we will see, I believe that an appropriate understanding of sensations and consciousness will give us a good picture of the continuities between the thinkers and the non-thinkers. With those discussions in place we will be able then to say something about the intelligent behavior of non-thinking animals, which will put into place the final piece of the framework for the Linguistic Theory of Thought.

To begin, then. There is a broad worry about contemporary representationalist views that goes something like this. If we stick with the language of representationalism, having a conceptually informed mental state is a matter of the tokening of symbols in the mind or brain. Even if we think that this view can work when it comes to cases of thinking things over, contemplating this or that, etc., there appears to be a problem when we consider what seem to be much simpler cases: *viz.*, becoming aware of something in one's environment—seeing, hearing, touching, etc., some object. Consider a simple case, say seeing a blue flower. For thinking beings such as ourselves this event involves conceptual mental states—we see the blue flower *as* a blue flower. That means that, according to representationalists, the simple act of perceiving a blue flower involves the tokening of symbols. Yet consider the fact that there is all the difference in the world between thinking about a blue flower ('picturing' it in one's mind, say) and actually seeing one. Are we to believe (says the objector) that having a visual sensation of a blue flower, simply fixing one's eyes upon it and becoming aware of it, is just a matter of symbols being tokened in one's head? To put it another way, are we really supposed to believe that the involuntary, vividly real, forceful visual awareness of a blue flower is nothing but the triggering of abstract symbols? So the question is, How do

representationalists distinguish between the *thought* of a blue flower and the actual *perception* of one? That there *is* a difference is obvious and undeniable. Since both situations involve the tokening of symbols, however, there must be something that gives perceptions their unique and irrepressible character. The most likely candidate for the 'something' that distinguishes the two cases is *sensation*.

Notice, as a relevant aside, that I, at least, have put myself even more directly into the path of this worry by insisting, in Chapter Four, that the difference, on my view, between Jane's perception of a black ball, and her dog's perception of that ball rested in essence primarily on the fact that Jane saw the black ball *as* a black ball (her perception was conceptually informed) while her dog did not. This means that I not only need to be able to explain what separates thoughts from perceptions in thinking beings (a difficulty that faces all representationalists in one way or another), I also need to be able to explain what separates the *perceptions* of thinking beings from the *perceptions* of non-thinking beings.

Let me be clear, too, that this worry is aimed at the most basic of cognitive activities. It will not do, that is, to suggest that the objector is concerned with the higher cognitive functions of, say, conceptually informed beliefs (even perceptual beliefs) of the sort that thinking beings have. The objection is aimed at the simple act of perceiving. I'll illustrate with a quick example. Imagine that I am walking across campus, as I do many times every day. As I walk I perceive the environment around me. We can imagine that I am not taking a nature stroll, so my perception of the environment is almost in the background of my mind; its primary purpose (I imagine) is to allow me to navigate my environment successfully. As I walk, then, I am aware of many things; let's

pick one: a large tree off to the side of the path. I perceive the tree—I see it—but I am not reflecting on the fact that I see it. I notice it without really even being self-consciously aware that I have noticed it. Yet, despite that, I still notice it *as* a tree. My perception of the tree is (on my view), unlike that of a non-linguistic animal's perception of the tree—mine is conceptually informed. We may imagine that my perception of the tree is not yet even a belief. If stopped and asked to say what it is I see, I could say that there is a tree by the side of the path. But in the ordinary course of my wanderings my perception of the tree does not count as my having formed any beliefs about the tree (or even that there is a tree). Nevertheless, I claim that the perception is conceptually informed, and hence involves the tokening of linguistic symbols in my brain. And that is where the problem seems to lie, for it will strike many people as odd (I'd think) to suppose that the mere perception of an object in my visual field is a matter of symbol tokenings occurring in accordance with the criteria that the LTT places on conceptual representings.

Other representationalists, at least, won't have a problem with the idea that perception involves representations. What seems problematic is the idea that merely perceiving, say, a tree *as a tree* requires language and rules in the way that the LTT says it must. The worry ends up being two fold. First of all there will be an objection—basically the same objection to my denying that non-linguistic animals have thoughts—to the idea that non-linguistic animals' (and pre-linguistic humans') perceptions have none of the conceptual character that our perceptions have. This worry is disposed of as we dispose of its sister worry: that consequence is a bullet I'm willing to bite. The second objection is more telling, however, for here I think many will find it implausible (to say

the least) to suppose that even my own simple acts of perception require the complex apparatus of conceptual representation as spelled out by the LTT. Perceiving a tree as a tree will be a conceptual event, it will be conceded, but surely (says the objector) it's too simple an event to require language and rules and all that. Yet that is precisely what the LTT requires.

As I said above, the defining feature of perceptions (that set them apart from thoughts, for example) is that they involve sensations. So the worry that I need to respond to, and one of the issues that I must here address, involves clarifying the relationship between conceptual mental states and the simple impacting of objects in the environment on our sense receptors. That is a more complicated way of saying that what I need to talk about here is the relationship between thought and sensations.

Let's return to the blue flower. Unlike the example of Jane and her dog, from Chapter Four, here we are not concerned with the difference between a non-thinking animal's perception of the environment and a thinking being's perception of the environment (we will come to that later, but that is not the issue here). Rather, we are interested in two different states in the same thinking being: (1) its perception of an object in its environment; (2) its thoughts about that object when the object is not being perceived. Of course, even on a representationalist view it's not as though there aren't obvious significant differences between these two states. There are differences in the contexts of their respective occurrences, and in the various physical facts of each situation (e.g., the causal story that leads to the tokening of the representation), that we have every reason to suppose are important. Yet, in both cases, the LTT will claim that the conceptual representing of a blue flower involves the tokening of linguistic symbols

that mean 'blue flower.' What we want to know is if, and if so how, the production of this token in the case of visually perceiving the flower is different from the case of having a thought about the blue flower (without perceiving it). That the two cases seem quite different is hard to deny. Yet in each case we are supposedly (according to the LTT) dealing with the mental tokening of a representation (composed of linguistic symbols) that means 'blue flower.' Is there anything to distinguish the two cases, any way to account for the seemingly significant difference between perceiving a blue flower and merely thinking about one, besides the fact that the tokening of 'blue flower' in the case where I am perceiving the blue flower is caused by an object in the environment?

Again, I believe that if we think about this situation in fairly intuitive, straightforward terms, most of us would be inclined to say that the primary difference between the two cases we're considering is the fact that (1) involves a *sensation* while (2) does not. That is, the causal story is the important distinguishing factor, but it is not so much the fact of causation as it is the nature of the mental event that results from the causal connection to the environment, that seems important. I think that this is really the right way to go. The difference, on my view, *is* that perceiving involves a sensation, while thinking about an object does not. To see how I intend for this to provide a solution to our worry, though, I have to say something about what I take sensations to be, within the framework of the LTT.

Though I will here offer a simple definition (if it can be called that) of sensations, I'm not attempting to do empirical science from the armchair. That is, I'm not getting into the nuts and bolts of biology, etc. in giving the following definition of sensations. What I am doing is taking a concept that I think is relevant to the development of the

philosophical framework that I'm creating and explaining its place in the new paradigm. That said, here is what I take sensations to be. Ordinarily they are simply the activation of sense receptors (nerve endings, or whatever) in conscious beings. This definition is quite straightforward, though undoubtedly not entirely satisfying—in large part because it relies on the concept of consciousness, and we haven't yet said anything about that topic. I'm not actually very concerned to offer a thorough definition of sensations, however. What I really want to do is call attention to the fact that, on my view, *sensations are not representations*. The picture I have in mind is of sensations as the phenomenal aspects of conscious experience that are often triggered by the world impinging on our senses (and which *accompany* the representations that are triggered in by the same event, while not themselves being further representations). When I see a red object I experience (have a sensation of) red because the light waves reflecting off the object strike my eyes (the appropriate sense receptors), which in turn sends signals to my brain, etc., etc. Of course, there are many types of sensations that do not involve the external world impinging on our senses; for example, when I have a headache the pain that I feel is certainly a sensation, but there is no external stimulation of my senses. (There are other examples as well, such as the experience of various illusions or hallucinations; there are, as I said, many examples of sensations that do not involve the impingement of the world upon our senses.) Still, my senses (my nerve endings, say, or the signal sent to my brain) are activated in such a way as to cause the experience of pain. As a thinking being I will, of course, conceptualize that pain (or the experience of seeing red, or whatever), and that will involve representations. And sensations in non-thinking animals will undoubtedly trigger non-conceptual representations in those animals. These sensations *themselves*,

however, are not representations of anything. They are simply the brute activation of the animals', or my, senses.

Perhaps the most important feature (for my purposes here) of this definition of sensations is that is does not require that sensations be in any way part of a conceptual representational system, i.e., that we can talk about the sensations of non-linguistic animals. Having a sensation of something, that is, does not require having a *concept* of the thing. I think that it's fairly obvious that sensations require consciousness (since, after all, we don't want to end up with the absurd suggestion that rocks or the dead can have sensations). But as I indicated, I do not require that the conscious beings who have the sensations also have concepts. We might imagine, say, that a dog feels pain without having any concept of pain. It is a brute sensation, of the kind that thinking beings may very well have a hard time imagining, since sensations in thinking beings will (almost surely) always involve a conceptualizing of the experience. When I feel pain the sensation will be accompanied by, or will cause, a conceptual representing on my part. When the dog feels pain the sensation may be accompanied by, or cause, a non-conceptual representing. Yet the sensation may very well be the *same* sensation in me as in the dog—the sensation is not to be equated with the representation (or with any representation), conceptual or not.

This idea is important on my view because it speaks to the issue of the continuity that exists between thinking and non-thinking beings. I find the existence of this continuity plausible for various reasons which we might sum up as the 'similarities in physical structure' that exist across the spectrum of conscious animals. Again, though, the important point is just that sensations are 'experiential primitives'; that is, they are

non-representational features of conscious experience that cross the divide between the conceptual and non-conceptual representers.

Let's return now to consideration of the difference between *seeing* a blue flower and *thinking* of one. How does an appeal to the idea of sensations, as just defined, figure into an account of this difference? Quite simply, in fact: the sensation of a blue flower figures into the difference by being present in one case (seeing the flower) and absent in the other (thinking about the flower). This is a simple, but important, distinction. Perhaps the most crucial function of sensations is, surely, to allow animals, both thinking and non-thinking alike, to navigate and interact with their environments. The impacting of its sense receptors by objects in the environment is a conscious being's link to the world. Sensations are, thus, hardly trivial occurrences. So, I am suggesting, their presence or absence can make a profound difference in the total experience one has when mentally tokening a set of linguistic symbols. We needn't change anything about our account of how conceptual representations work or get their meaning according to the LTT in order to accommodate this point. We're simply noting that when I token, say, 'blue flower' there is a very large difference between the case in which that token occurs because of a sensation I'm having and the case in which that token occurs without a sensation of a blue flower.

While the idea of sensations provides us with an explanation of the primary difference between perceiving an object and merely thinking about that object, it also, as I've said, provides us with a link between thinking and non-thinking beings. I am arguing for a kind of continuum from simple animals to thinking beings such as ourselves. This is not a continuum of degrees of representational capacity (as we saw in

Chapter Four), but a continuum of physical attributes, of organs and hardwiring, and certain experiential capacities. These similarities lead to shared cognitive traits. In non-thinking beings sensations will trigger stimulus-response behavior patterns, and (in more advanced animals) non-conceptual representations. In thinking beings these same sorts of sensations will trigger conceptual representations. During the language learning process, which is (on the LTT) also the process of learning to think, we have every reason to suppose that the sensations had by the language learner will remain the same even after she has transitioned to being a language speaker, and hence a thinker. In the language learner, those sensations trigger pattern-governed behavior and the tokening of non-conceptual representations; once the learner begins to get the hang of things, those representations will become conceptual (as we have already discussed in Chapters Four and Five). Yet there is no reason, on my view, to suppose that anything about the sensations that give rise to these various representations will change during the transition from learner to speaker. I believe, that is, that it is reasonable to suppose that the functioning of sensations in the production of mental representations will remain undisturbed as a non-conceptual representer transitions to a conceptual representer.

This may all be well and good as far as it goes, but it doesn't really go all that far. The impacting of sense receptors produces nothing, does nothing, and serves no purpose, if the being whose sense receptors are being impacted isn't conscious. So we have to say something about consciousness to really flesh out the importance of this point. Consciousness, of course, is a very, very big topic. One could write books on the subject (as many philosophers have). So my discussion of consciousness here is rather modest in scope. I wish primarily to say just a little bit about the elements of consciousness that the

LTT embraces and that help us see the true significance of the above discussion of sensations. I also want to talk about an idea that I first encountered in an essay by Ned Block: that consciousness is not a single phenomenon; rather, there are many types of consciousness. This idea is important to the way that I envision non-thinking animals interacting with their environments—and, ultimately, in how the LTT deals with the difference between the functioning of non-conceptual representations and conceptual representations in mediating between organism and world.

To begin, then, I am going to assume that most of us are comfortable with the idea that (to borrow Thomas Nagel's phrase) there is 'something that it's like to be' a dog, a cat, an infant, etc.[139] (just as there is something it's like to be me, or you). After all, such beings are not rocks.[140] Now Ned Block distinguishes, in his essay "On a Confusion About a Function of Consciousness," between what he calls 'access consciousness' and 'phenomenal consciousness.' I am going to modify this distinction, and use Block's

---

[139] Of course, this famous bit from Nagel is often used in philosophy of mind to motivate at least a couple of claims that might be seen as contrary to the central thesis of the LTT. First of all, many people take Nagel's point (in "What is it Like to be a Bat?", *Philosophical Review* 83 (1974), pp.435-450) to be that physicalism must be false. Secondly, all this 'what it's like to be' something kind of talk is usually part of the argument in favor of qualia—those elusive elements of conscious experience that are supposedly missed by functionalist theories of mind (and physicalist theories generally). Allow me to briefly respond. With regard to the first point, I simply wish to point out that a careful reading of Nagel's essay reveals that his conclusion is not, in fact, that physicalism must be false; rather, his conclusion is the more modest one that we cannot as of yet see how physicalism could be true (because we do not understand the relevant phenomena and concepts—the physical world, objectivity and subjectivity, and conscious experience—well enough to truly make sense of the claim that mental states are physical states). As for the issue of qualia: that is without doubt a very tricky question, but I do not think that merely acknowledging that there is 'something it's like to be' a conscious creature commits us to the existence of 'qualia'. At the very least, we would have to say a lot more about what these 'qualia' are supposed to be. Regardless of how all that turns out, however, I don't think that the existence of qualia would in any way endanger the LTT, for reasons that will become apparent below.

[140] I suppose it may be less obvious that they are not automata—in fact, it's probably somewhat contentious (at least in some circles) to assert that they are *not* automata. And, we may suppose, there isn't really anything it's like to be an automaton. Nevertheless, for the most part I intend here to take for granted that non-linguistic animals and infants are more than just machines. I hope that what I say makes it at least appear plausible to suppose that such beings have some form of consciousness, but I don't intend to argue directly for the claim that they do. Rather, assuming that they do, I want to say what that amounts to.

general assertion that consciousness is not a unified concept, to argue that we can grant animals and infants consciousness—and a consciousness in many ways like our own—without granting them thoughts. The value of this, if it is not obvious, is that it will bring us closer to our goals of (a) accounting for the intelligent behavior of such creatures within the paradigm put forth by the LTT, and (b) further explaining the continuity between thinking and non-thinking animals.

According to Block, discussions (philosophical and otherwise) of consciousness suffer from the fact that those who discuss it fail to realize that *consciousness* is actually "a hybrid, or better, a mongrel concept: the word 'consciousness' connotes a number of different concepts and denotes a number of different phenomena" (1995, p.227). The main point of Block's article is to distinguish between what he calls 'phenomenal consciousness' (P-con) and 'access consciousness' (A-con). He argues that these two senses of consciousness (and others as well) have commonly been run together. The result (with respect to P-con and A-con) is that what Block considers an obvious function of access consciousness—*viz.*, making the content of our experience available for reasoning and the rational control of action—is illegitimately taken to be a function of phenomenal consciousness as well. Or, to be more accurate: since most people have not distinguished between P-con and A-con, this obvious function of Block's 'access consciousness' is taken to be the (or a) function of consciousness *simpliciter*. While acknowledging that, in thinking human beings, at least, phenomenal and access consciousness almost always interact and work together—thus, that the functioning of access consciousness will both affect and be affected by phenomenal consciousness—

Block still wants to insist that they ought to be kept distinct in our theorizing about consciousness. With this I agree.

I should say right up front, though, that while what I have to say about the different kinds of consciousness is, I believe, largely in keeping with Block's distinction, Block himself would not be particularly sympathetic to the use to which I am going to put that distinction. I am going to argue that there is a type of consciousness (something much like Block's 'phenomenal consciousness') that is possessed by non-thinking animals and helps to explain their intelligent behavior, while there is a different sort of consciousness (which is like Block's 'access consciousness) that comes into play when we're dealing with thinking beings and conceptual representations. Block, however, would deny both (1) that animals have only P-con and not A-con, and (2) that having only P-con would enable animals to exhibit the kinds of behaviors that they do. Thus, for the most part, one can assume that what follows is more inspired by Block than a using of his article in support of my own ends. Before we get to my own ideas of consciousness, however, we will be well served by looking briefly at the key points of Block's definitions of P-con and A-con.

Block finds A-con easier to define, and has this to say:

> A state is access-conscious ... if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous, that is, poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech. (1995, p.231).

He goes on to note that while he takes these three conditions as sufficient for A-consciousness, they are not all necessary. In particular, he claims that (3) is not necessary, because he wants "to allow that non-linguistic animals ... have A-conscious

states" (p.231). For our purposes, though, I want us to note that Block's A-con involves reasoning and inference. Right away, then, it should be obvious that A-con is not the sort of thing that the LTT will ascribe to non-linguistic animals—for reasoning and inferring are activities that require *thought*. We'll see below that there is a sense in which I am willing to grant that the non-conceptual representations of non-linguistic animals and pre-linguistic humans will figure into processes that are the non-thinking analogues of things like inferences. Nevertheless, reasoning is not something that non-thinking beings are capable of, so they are not going to be access conscious, on my view. To avoid confusion, I will refer to the kind of consciousness that the LTT ascribes only to thinking beings (and that is my own version of Block's A-con) as 'conceptual consciousness' (C-con).

Now, while Block's definition of A-con is quite straightforward, phenomenal consciousness, on the other hand, is much more difficult to define. (Yet it also seems to be the sort of thing that is most often meant by 'consciousness.') Block acknowledges at the outset that he "cannot define P-consciousness in any remotely noncircular way" (p.230). The best we can do, here, according to Block, is point:

> P-consciousness is experience. P-consciousness properties are experiential ones. P-conscious states are experiential, that is, a state is P-conscious if it has experiential properties. The totality of the experiential properties of a state are 'what it is like' to have it.... [W]e have P-conscious states when we see, hear, smell, taste, and have pains. P-conscious properties include the experiential properties of sensations, feelings, and perceptions .... (P.230).

As I said, it seems to me that P-con is really what most people generally mean by 'consciousness.' Block even borrows Nagel's phrase in calling P-con the 'what it's like' aspect of our experience. If an organism is phenomenally conscious, then there is

something it is like to *be* that organism.[141] But, as becomes evident more clearly as Block continues to talk about it, we need to realize that Block does not consider P-conscious states to be simply the experience of phenomenal features of the environment (despite the fact that he calls it 'phenomenal' consciousness). When Block here defines P-conscious states as involved in *seeing*, *hearing*, in short *perceiving*, these are more robust acts than they might at first appear to be. This comes out in a couple of examples that Block uses in his discussion and that we will look at momentarily. As with A-con above, however, I think it will be easier for us if we don't try to use Block's terminology. So, instead of P-con, I will talk of 'non-conceptual consciousness' (NC-con). NC-con and P-con will have many things in common, of course (just as C-con and A-con do), but I will put NC-con to broader use than Block puts P-con.[142]

Now, even though Block doesn't have the same use for his distinction (between A-con and P-con) as I do for mine (between C-con and NC-con), it is still instructive for us to ask what Block hopes to gain by making his distinction. The answer, simply put, is that he is trying to show that a legitimate function of A-con is often illegitimately taken to be a function of P-con (or, rather, of consciousness generally—since other theorists make no distinction between A-con and P-con). This leads Block to the discussion of two examples that illustrate the problem—and these two examples will be important for our purposes as well. The first example has to do with the phenomenon of blindsight; the second with certain epileptic episodes. Here is what Block has to say about the first:

---

[141] Nagel, of course, only talks about consciousness generally, not having made the distinction that Block is trying to make.

[142] Perhaps most importantly, even though I think it would be a mistake to treat Block's P-con as just another way of talking about the having of sensations (a mistake that is invited by the fact that Block calls this kind of consciousness 'phenomenal'), I still want to give broader scope to NC-conscious states, so that they include the phenomenal aspects of experience as well as the having of non-conceptual representations. And this may very well be a more robust notion of consciousness than is captured by Block's P-con.

> Patients with damage in their primary visual cortex typically have "blind" areas in their visual field. If the experimenter flashes a stimulus in one of the blind areas and asks the patient what he saw, the patient answers "nothing." The striking phenomenon is that some (but not all) of these patients are able to "guess" reliably about certain features of the stimulus, features having to do with motion, location, direction (e.g., whether a grid is horizontal or vertical). In "guessing," they are able to discriminate some simple forms. If they are asked to grasp an object in the blind field (which they say they cannot see), they can shape their hands in a way appropriate to grasping it, and there are some signs of color discrimination. (1995, p.227).

Given that there seems to be some sense in which consciousness is missing in cases of blindsight, Block notes that some theorists have concluded from such cases that "consciousness must have a function of somehow enabling information represented in the brain to be used in reasoning, reporting, and rationally guiding action" (p.228). The reasoning that Block wants to target thus goes like this:

> [W]hen a content is not conscious—as in the blindsight patient's blind field perceptual contents, it can influence behavior in various ways, but only when the content is conscious does it play a rational role; and so consciousness must be involved in promoting this rational role. (P.228).

Given his distinction between P-con and A-con, though, it should be obvious that while Block would agree that the rational control of action is a function of *access* consciousness, it is not a function of consciousness *simpliciter* (since there is no such thing) nor of phenomenal consciousness. It's not, then, the reasoning *per se* that Block has a problem with, but rather the fact that those who have so reasoned have failed to distinguish between the different types of consciousness, and have thus reasoned too broadly.

The second case is used by certain theorists in much the same way as the phenomenon of blindsight. Block talks about van Gulick's and Searle's discussions of Penfield's

> observations of epileptics who have a seizure while walking, driving, or playing the piano. The epileptics continue their activities in a routinized, mechanical way despite, it is said, a total lack of consciousness. Searle says that because consciousness as well as flexibility and creativity of behavior are missing, we can conclude that a function of consciousness is somehow to promote flexibility and creativity. (P.228).

The problem here is much the same as it is above: The reasoning might be acceptable if it's taken to apply to A-con, but it cannot be legitimately applied to consciousness generically nor to P-con. These cases do not, of course, demonstrate that there *are* two (or more) kinds of consciousness; rather, they are 'springboards' for thinking about consciousness.

Block, of course, is going to use such cases to suggest that there are in fact different types of consciousness—in part because distinguishing between A-con and P-con will give us a more fruitful way of accounting for such phenomena as blindsight and epileptics continuing routine behavior during epileptic episodes. I, on the other hand, want to use these two examples to explore the distinction between conceptual and non-conceptual consciousness.

I take it that researchers generally agree that the epileptic in the middle of a seizure is not what Block would call access conscious (since those who talk about such research in the context of consciousness cite these cases as evidence for the idea that a (or the) function of consciousness is to allow for the flexibility and creativity of conscious behavior, as opposed to the routine behavior exhibited by the epileptics). The suggestion

is that the behavior of these (supposedly unconscious) epileptics (say, driving a car) continues uninterrupted precisely because it is routine, and hence does not require consciousness. If any sort of rational decision-making were required of the epileptic while having his seizure, he would not be capable of doing it. Nevertheless, the routine behavior can be successfully performed. The epileptic can continue to make his way home, driving his car successfully (without crashing). Now, Block's target in his article is the reasoning that goes from talking about these cases of epileptics continuing to carry out routine behavior to the suggestion that the epileptics are actually completely unconscious. Block disagrees, and writes:

> [N]either Searle nor van Gulick nor Penfield give any reason to believe that P-consciousness is missing or even diminished in the epileptics they describe. The piano player, the walker, and the driver don't cope with new situations very well, but they do show every sign of *normal sensation*. For example, Searle, quoting Penfield, describes the epileptic walker as "thread[ing] his way" through the crowd. Doesn't he *see* the obstacles he avoids? Suppose he gets home by turning right at a red wall. Isn't there something it is like for him to see the red wall—and isn't it different from what it is like for him to see a green wall? Searle gives no reason to think the answer is no. Because of the very inflexibility and lack of creativity of the behavior they exhibit, it is the *thought processes* of these patients (including A-consciousness) that are most obviously deficient; no reason at all is given to think that their P-conscious states lack vivacity or intensity. (Pp.239-240).

I want to focus for a bit on a couple of parts from this long quote, for I think that Block has here described the very differences between C-con and NC-con that are essential to the LTT.

First of all, notice that Block sees no reason to suppose that the epileptics in question lack normal *sensations*. But more than that, the walker can 'thread his way' through a crowd—that is, he can successfully navigate his environment in response to

visual stimuli. Block imagines one of these epileptics turning right at a red wall. Granted, this behavior would take place because it is routine. Nevertheless, if it is the presence of a *red* wall that triggers the habitual turn to the right, shouldn't we suppose that the epileptic *sees* the red wall, in some sense of 'see'? If we believe the interpretation of these episodes by the researchers who study them, that the epileptics lack any sort of conscious states that would allow them to *reason*, it would seem that the epileptics in question are lacking what I am calling 'conceptual consciousness.' Simply put, they are not having anything thoughts. Nevertheless, they are not *un*conscious: that is what the example is supposed to show. They lack reason, but they retain the ability to navigate their environments, etc.

Notice, though, what it would mean to grant that these epileptics do respond to the visual stimuli in their environment: it would mean that responding to one's environment does not require that one be C-conscious. Animals are often quite responsive to stimuli in the environment, and that is sometimes taken to indicate that they must be conscious. I would agree that this responsiveness indicates a kind of consciousness. But why suppose that it's thoughtful, *conceptual* consciousness that these animals possess? The cases of the epileptics we're considering suggest to me that there is no reason to go this far.[143]

Notice, too, that Block says that the lack of flexibility and creativity in epileptics having seizures indicates that their *thought processes* are deficient (or, I would add, perhaps absent altogether), but that this gives us no reason to suppose that their NC-

[143] I'm not suggesting, of course, that animals are the same as epileptics having seizures. In fact, a normally functioning animal may be much more capable than an epileptic—a lot of normal animal behavior is really quite flexible, in a way that the epileptic's behavior is not. (Though, granted, 'threading one's way' through a crowd seems like fairly flexible behavior, within certain bounds. Must this be substantially different from an animal threading it's way through the trees?) What's important here, though, is simply the suggestion that responses to one's environment do not seem to require C-con.

conscious states are in any way lacking. Perhaps the epileptic who sees the red wall forms no memory of having seen it, and hence could obviously not *report* that there was something it was like to see it. But does that mean that there is not *in fact* something it's like for the epileptic to see the red wall? If the wall had that very day been painted green, would the epileptic still have been able to find his way home? Similarly, though conversely, why suppose that the fact that there is something it is like for an animal to see a tree or a predator requires that the animal has thought processes or conceptual consciousness?

Recall, though, that we are not here concerned to *argue* that animals lack thoughts (and C-con). Rather, *assuming* (as the LTT must) that animals lack thoughts, I want to suggest a way in which this could be the case while still accepting that animals are conscious. The cases of the epileptics, and Block's analysis of what such cases show, I think provide a plausible way of ascribing consciousness to animals that lack thoughts (conceptual representations): animals are NC-conscious, but not C-conscious.[144] That is, there is something it is like to be a cat, or a dog, but none of their conscious states are conceptual, nor available to them for use in inferences, reporting, reasoning, or the rational control of action. I think that it is at least *prima facie* plausible to suppose that the non-conceptual conscious states of any animal (e.g., its sensory awareness of a tree) play a role in controlling the behavior of the animal (e.g., its moving around the tree rather than running into it).[145] Yet, such control need not suggest the presence of

---

[144] Again, though, remember that I am not equating animals with epileptics. See previous footnote.

[145] I limit myself to a *prima facie* plausibility here, rather than something stronger, because I do not think it is in any way *obvious* that the conscious states of an animal play a causal role in its behavior. After all, why couldn't it be the case that the 'what it's like to be'-consciousness of animals (and humans, too, I suppose) is a mere by-product: that the pathways from sensory input to behavioral output are causally

reason—and certainly does not need to suggest that the animals have concepts. Nor, of course, do I see any reason to suppose that any of an animal's conscious states are available for use in reasoning.

Let us contrast, briefly, these points and the cases of epileptics carrying out routine behaviors during seizures with the phenomenon of blindsight. I take the epileptic cases to be instructive because I think that they suggest a way in which we can use the C/NC-con distinction to both explain how we can ascribe consciousness to animals without also ascribing thoughts to them, and also begin to explain how non-thinking animals can exhibit the intelligent behaviors that they so often do. The blindsight cases also suggest that the C/NC-con distinction might be useful to the LTT in the way that I am proposing, but for a different reason. The epileptics having seizures seem to lack C-consciousness, but we have seen that there's no immediate reason to suppose that they lack NC-consciousness. And because they are able to exhibit fairly complex behaviors in response to their environments, it seems (to me, at least) somewhat plausible to suppose that NC-consciousness can enable such behaviors. The blindsight patients, on the other hand, do not exhibit very complex behaviors. On its face, the really amazing thing about blindsight is that (some of) the patients can make fairly accurate 'guesses' as to a certain range of limited features of objects in their blind field (in response to the experimenter's prompting, of course). What's interesting for our purposes, though, is just the opposite

---

closed? In that case, it would certainly be the case that animals could navigate their environments in response to sensory stimulation, but the fact that there was 'something it's like' to have such sensations would not figure into the causal story. Or it might even be the case that behavioral responses to stimuli could take place in the complete absence of any consciousness. In fact, modern experiments in robotics (e.g., with Honda's famous ASIMO robot) seem to provide evidence that this can in fact be the case—assuming, of course, that we do not believe the robots have any form of consciousness. With respect to most animals, however, I do think that it is likely that the phenomenal states of the animal are the same states that trigger motor-control mechanisms.

point: the blindsight patients' responses to stimuli are so utterly limited as to suggest that consciousness is missing altogether. Block himself simply claims that "in the blindsight patient, both P-consciousness and A-consciousness ... are missing," though he notes that this claim is "just an assumption. I decided to take the blindsight patient's word for his lack of P-consciousness of stimuli in the blind field" (p.242). Of course, if C-consciousness is missing in the blindsight cases (C-consciousness of the objects in the blind field, of course—the patients are still C-conscious in general), we really can't take the patient's 'word' for it that the objects in the blind field are NC-unconscious as well—since, if the presence of those objects are C-unconscious, then *of course* the subject cannot *report* their presence.

Nevertheless, I think that we actually may have some reason for supposing that the perceptual contents of the blind field are NC- as well as C-unconscious. If those perceptual contents were NC-conscious, I think that we'd expect to see much more sophisticated behavior on the part of the patients—just as we observe in the epileptics.[146] Blindsight patients, however, do not exhibit very sophisticated behavior: they barely respond to the objects in their blind field at all. Not only is the content of the blind field unavailable for use in reasoning and the rational control of action, but it seems to be unavailable even for most *non*-rational control of action. Of course, the content of the blind field does appear to control some actions: the blindsight patient can shape his hand in a way appropriate to grasping the object presented in the blind field, and his 'guesses' about certain features of the object can sometimes be surprisingly accurate. Why suppose, though, that any of this requires NC-consciousness? What the phenomenon of

[146] We might expect to see behavior like that of Block's 'superblindsighters' (p.233).

blindsight suggests to me is just that our perceptual mechanisms might sometimes by-pass consciousness altogether in sending information to our motor control (and, in the case of C-conscious beings, rational) systems. At any rate, the contrast (in terms of the complexity of the behaviors available in each case) between the epileptics and the blindsight patients is at the very least, I think, suggestive of the idea that NC-consciousness is (in part) responsible for the (non-rational) control of some actions.

All of this gives us reason, I think, to suppose that the distinction between NC-con and C-con can be fruitfully used by the LTT to explain the consciousness of non-linguistic (and hence, non-thinking) animals. But there is even more to say. One thing that Block notes is

> that P-conscious content is phenomenal, whereas A-conscious content is representational. It is of the essence of A-conscious content to play a role in reasoning, and only representational content can figure in reasoning. Many phenomenal contents are *also* representational, however, so it would be better to say that it is in virtue of its phenomenal content or the phenomenal aspect of its content that a state is P-conscious, whereas it is in virtue of its representational content, or the representational aspect of its content, that a state is A-conscious. (P.232).

What this adds to our earlier understanding of the difference between P-con and A-con, is this: A-conscious states are essentially representational, whereas P-conscious states are not. Yet, while it is not of the essence for P-conscious states to be representational, Block says that many of them are *in fact* representational (in addition to being essentially phenomenal, of course). To see the value, for our purposes, of this point, consider what Block says a few paragraphs later about "a perceptual state" of seeing a square:

> This state has a P-conscious content that represents something, a square, and thus it is a state of P-consciousness *of* the square. It is a state of P-consciousness of the square even if it doesn't represent the square as a

> square, as would be the case if the perceptual state is a state of an animal that doesn't have the concept of a square. (P.232).

The idea here, I take it, is that the perceptual state of seeing a square can be *representational* without being, as it were, *conceptual*—one can be phenomenally conscious *of* a square without being conscious of it *as* a square. This is the essence of my distinction between C-conscious and NC-conscious states: the latter can have a great deal of content, but that content simply isn't conceptual content. Thus, it seems to me, if NC-conscious states can represent features of the world without representing them under particular descriptions, such states are promising candidates for the conscious states of animals under the LTT. Hence (we might speculate), the representational features of an animal's NC-conscious states are what allow it to successfully navigate its environment, even though the animal has no thoughts.

This actually seems to fit pretty well with the way Block understands his point about representationalism. He writes:

> The paradigm P-conscious states are sensations, whereas the paradigm A-conscious states are "propositional attitude" states such as thoughts, beliefs, and desires, states with representational content expressed by "that" clauses (e.g., the thought that grass is green).... [H]owever, thoughts are often P-conscious and perceptual experiences often have representational content.... Pains [for example] often represent something (the cause of the pain? the pain itself?) as somewhere (in the leg). A number of philosophers have taken the view that the content of pain is *entirely* representational .... I don't agree with this view, so I certainly don't want to rely on it here, but I also don't want to suggest that the existence of cases of P-consciousness without A-consciousness is a trivial consequence of an idiosyncratic set of definitions. To the extent that representationalism of the sort just mentioned is plausible, one can regard a pain as A-conscious if its representational content is inferentially promiscuous, and so on. (P.232).

In my terms, then, a pain is C-conscious if it is 'inferentially promiscuous,' so animal pains are not C-conscious. They are, nevertheless, still *sensations*, still *phenomenal* states, and still *non-conceptually* conscious.

There is clearly a lot more that could be said about animal consciousness under the LTT, and about my C/NC-con distinction; consciousness is a huge topic, and my treatment of it in this project must remain necessarily brief. What I've tried to do here is just begin to show how a more sophisticated understanding of the phenomenon of consciousness can benefit the LTT.

Furthermore, while I obviously think that there are sharp and drastic differences between linguistic and non-linguistic animals, between the thinkers and the non-thinkers, I also believe that there are marked and important similarities—continuities—between us (the only thinkers we're sure of) and other animals. This is where the C/NC-con distinction comes in. If consciousness is not a single, unified concept, but rather an ambiguous concept (which sometimes means NC-consciousness, sometimes C-consciousness, sometimes self-consciousness, etc.) then there is room to draw apart the different senses, the different types, of consciousness and say which are common between us and the other animals, and which are not. I see vast potential in this, particularly in discussing the continuities between thinking and non-thinking animals.

The picture we end with is this. NC-conscious beings (the non-linguistic animals and pre-linguistic humans) will have non-conceptual representations of their environments (along with, I am happy to suppose, whatever phenomenal contents accompany such representations in virtue of such beings' sensations). These representations will first and foremost allow the NC-conscious beings to navigate their

environments, but we needn't suppose that these representational states are simplistic, stimulus-response events. In more sophisticated animals (and, most importantly, in pre-linguistic humans) they will be the analogues of those first-order, perceptual conceptual representings that figure into the C-conscious awareness of one's environment enjoyed by all thinking beings. There is no need to rehash the differences between non-conceptual and conceptual representations. What is important is to recognize that while the LTT posits a sharp divide between thinkers and non-thinkers in virtue of the fact that the former, but not the latter, have conceptual representations, the LTT also endorses a strong sense of continuity in the 'natural order,' a continuum not of representations, but of biology and function. What begins as a simple stimulus-response capacity (e.g., the input-output mechanism found in Dretske's magnetotactic bacteria) develops into a capacity for non-conceptual representing. This latter capacity will surely admit of degrees, from the seemingly hard-wired behavior of the digger wasp (which, though hard-wired in many respects nevertheless is not so simple as the non-representational responses of the bacteria), to the flexible, adaptive, intelligent behavior of higher mammals, whose representations, though non-conceptual, allow for complex interactions with the environment, and the development of complex, behaviorally conditioned activities. To these beings we will ascribe consciousness of a distinct type (NC-consciousness, as distinct from C-consciousness). And then there will be those beings whose conditioned, complex, (and importantly, in fact crucially) linguistic behavior has finally shifted into the conceptual representing of the C-conscious thinking being—who has, at last, taken her rightful place within the space of reasons.

6.3 Closing Comments

The ability to think is perhaps the single most defining feature of *homo sapiens*. I have long believed that to understand anything about us and our place in this world, we must understand this amazing ability: what it is, how it works, and how we come to have it. The theories that we develop to address these issues are, to my mind, fundamental to understanding all aspects of those various enterprises that make anything about us interesting, from our scientific strivings to know the world around us, to our attempts (both as a species and individually) to know ourselves. Our epistemological, ethical, aesthetic, metaphysical, etc., inquiries are only made possible because we can *think*, because we can go beyond merely being in the world to actually conceptualizing that world, and ourselves.

And though my current project must end here, this is really only the beginning. The goal of this piece of work, as indicated by its title, has always been simply to set us upon the path toward what I take to be the most promising way to understand the phenomenon of thought. I have attempted to lay down a foundation, nothing more. This has, of course, been no small task. We have followed in the footsteps of giants, but this has only made more subtle the obstacles our theory must overcome, the questions it must answer, and the solutions it offers. I have been able to draw upon the work and insights of a great many philosophers who have come before me, but at the same time I have been therefore obligated to say that much more about why I believe we must move in the particular direction I suggest. I am, as I have already mentioned, attempting a synthesis. I locate the Linguistic Theory of Thought within the broad playing field of contemporary representationalism, but I bring to that field a firm belief that we will not move forward,

that we cannot succeed in understanding the human mind, without adding to representationalism the insights of philosophers such as Sellars, Quine, Wittgenstein, and those who continue to carry on the work they began.

My work here is, I believe, a step in the right direction. I have tried to clear the ground for the LTT, to argue its most fundamental virtues, and to suggest what the final superstructure might look like. I have shown that a linguistic conception of thought, generally, is both coherent and promising. I have also argued that the LTT specifically appears to have the best chance of any theory thus far of delivering on that promise.

Yet this is, as I've said, merely the beginning. The possibilities lie stretched before us—as do many remaining challenges. So how does one end a project of this magnitude which, for all that it does, is still only a beginning? If I can be forgiven a bit of a rhetorical flourish, I would like to end my project with a grand observation and prediction. In the end, the task of developing a proper and fruitful theory of mind will likely stay with us for as long as we care the think about it—but as minded creatures, as *thinking* beings, we are, by our very nature, equal to that task.

Bibliography

Armstrong, D.M. The Nature of Mind. Cornell, 1980.

Austin, J.L. How to do Things with Words. 2nd Edition. Harvard, 1962.

----. Sense and Sensibilia. Oxford, 1962.

Baier, Annette. Postures of the Mind. Minnesota, 1985.

Baker, Lynn Rudder. "On a Causal Theory of Content." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.351-363.

Bilgrami, Akeel. "On McDowell on the Content of Perceptual Experience." *Philosophical Quarterly*, vol.44, no.175, 1994. Pp.206-213.

Block, Ned. "Advertisement for a Semantics for Psychology." Reprinted in Mental Representation: A Reader, eds. Stephen Stich and Ted A. Warfield. Blackwell, 1994. Pp. 81-141.

----. "On a Confusion About a Function of Consciousness." *Behavioral and Brain Sciences*, vol.18, 1995. Pp.227-287.

Block, Ned and John Campell. "Functional Role and Truth Conditions." *Aristotelian Society: Supplementary Volume*, 1987. Pp. 157-181.

Boden, Margaret. "Escaping From the Chinese Room." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.379-388.

BonJour, Laurence. "Is Thought a Symbolic Process?" *Synthese*, vol.89. Kluwer, 1991. Pp.331-352.

----. "Analytic Philosophy and the Nature of Thought." Unpublished Paper: http://faculty.washington.edu/bonjour/Unpublished%20articles/UBCPAPER.html.

Boroditsky, Lera. "Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time." *Cognitive Psychology*, vol.43, 2001. Pp.1-22.

----. "Linguistic Relativity." Encyclopedia of Cognitive Science, ed. Lynn Nadel. Nature Publishing Group, 2003.

Boroditsky, Lera and Dedre Gentner. "Individuation, Relativity, and Early Word Learning." Language Acquisition and Conceptual Development, eds. Melissa Bowerman and Stephen C. Levinson. Cambridge, 2001. Pp.215-256

Boroditsky, Lera and Lauren Schmidt, et al. "Sex, Syntax, and Semantics." Language in Mind. MIT, 2003. Pp.61-79.

Brickhard, Mark H. "Some Notes on Internal and External Relations and Representations." Online Paper: http://www.lehigh.edu/~mhb0/Int.ExtRelations.pdf.

Carnap, Rudolf. Meaning and Necessity. Chicago, 1947.

Carruthers, Peter. Language, Thought and Consciousness. Cambridge, 1996.

----. "The Cognitive Functions of Language." *Behavioral and Brain Sciences*, vol.25, 2002. Pp.1-69.

Chalmers, David. "Can Consciousness be Reductively Explained?" *From* The Conscious Mind. Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.588-598.

Churchland, Patricia Smith. "Language, Thought, and Information Processing." *Nous*, vol.14, no.2, 1980. Pp.147-170.

Churchland, Paul M. "Eliminative Materialism and the Propositional Attitudes." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.184-197.

----. "Reduction, Qualia, and the Direct Introspection of Brain States." *The Journal of Philosophy*, vol.82, no.1, 1985. Pp.8-28.

Churchland, Paul M. and Patricia Smith. "Functionalism, Qualia, and Intentionality." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.163-177.

Cole, David. "Hearing Yourself Think: Natural Language, Inner Speech, and Thought." (1997). Online Paper: http://www.d.umn.edu/~dcole/hearthot.htm.

----. "I Don't Think So: Pinker on the Mentalese Monopoly." *Philosophical Psychology*, vol.12, no.3, Sept. 1999. Pp.283-295.

Cummins, Robert. "Functional Analysis." *The Journal of Philosophy*, vol.72, no.20, 1975. Pp.741-765.

----. Meaning and Mental Representation. MIT, 1989.

----. Representations, Targets, and Attitudes. MIT, 1996.

Davidson, Donald. "Thought and Talk." Reprinted in The Nature of Mind, ed. David M. Rosenthal. New York : Oxford, 1991. Pp.363-371.

----. "Mental Events." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.430-442.

Dawkins, Marian Stamp. Through Our Eyes Only?: The Search for Animal Consciousness. Spektrum, 1993.

Delaney, C.F., Michael Loux, et al. The Synoptic Vision: Essays on the Philosophy of Wilfrid Sellars. Notre Dame, 1977.

Dennett, Daniel C. Content and Consciousness. Routledge, 1969.

----. Brainstorms. Bradford, 1978.

----. Consciousness Explained. Little, Brown and Company, 1991.

----. "Real Patterns." *The Journal of Philosophy*, vol.88, no.1, 1991. Pp.27-51.

----. "True Believers: The Intentional Strategy and Why It Works." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.226-242.

----. "The Myth of Original Intentionality." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.389-400.

----. Brainchildren. MIT, 1998.

deVries, Willem A., and Timm Triplett. Knowledge, Mind, and the Given. Hackett, 2000.

Dretske, Fred. "Misrepresentation." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp. 329-340.

Field, Hartry H. "Mental Representation." Reprinted in Mental Representation: A Reader, eds. Stephen Stich and Ted A. Warfield. Blackwell, 1994. Pp.34-78.

Fodor, Jerry A. Representations. MIT, 1981.

----. "The Mind-Body Problem." *Scientific American*. Vol.244, no.1. January, 1981. Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.118-129.

----. "Why Paramecia Don't Have Mental Representations." Midwest Studies in Philosophy, vol.10, eds. Peter A. French, et al. Minnesota, 1986

----. Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT, 1987.

----. "Fodor's Guide to Mental Representation." Reprinted in Mental Representation: A Reader, eds. Stephen Stich and Ted A. Warfield. Blackwell, 1994. Pp.9-33.

----. In Critical Condition. MIT, 1998

Fodor, Jerry and Ernest Lepore. "Why Meaning (Probably) Isn't Conceptual Role." Reprinted in Mental Representation: A Reader, eds. Stephen Stich and Ted A. Warfield. Blackwell, 1994. Pp.142-156.

Geach, Peter. Mental Acts. Routledge, 1957.

Ghils, Paul. Language and Thought: A Survey of the Problem. Vantage, 1980.

Gibson, Roger F., Jr. Enlightened Empiricism. South Florida, 1988.

Gleitman, Lila. "The Structural Sources of Verb Meanings." Language Acquisition, vol. 1. Lawrence Erlbaum Associates, 1990. Pp.3-55.

Godfrey-Smith, Peter. "A Continuum of Semantic Optimism." Reprinted in Mental Representation: A Reader, eds. Stephen Stich and Ted A. Warfield. Blackwell, 1994. Pp.259-277.

Grice, H.P. "Meaning." (1957). Reprinted in The Philosophy of Language, 4th edition, ed. A.P. Martinich. Oxford, 2001. Pp.92-97.

----. "Utterer's Meaning, Sentence Meaning, and Word-Meaning." *Foundations of Language*, vol.4. Pp.225-42

Griffin, Donald R. Animal Minds. Chicago, 1992.

Hamlym, D.W. "Perception, Sensation and Non-Conceptual Content." *Philosophical Quarterly*, vol.44, no.175, 1994. Pp.139-153.

Harman, Gilbert H. "Three Levels of Meaning." Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology, eds. Danny D. Steinberg and Leon A. Jakobovits. Cambridge, 1971. Pp.66-75.

Henser, Steve. "Thinking in Japanese? What Have We Learned About Language-Specific Thought Since Ervin Tripp's 1964 Psychological Tests of Japanese-English Bilinguals?" *Nissan Occasional Paper Series*, no.32, 2000.

Horgan, Terence. "Computation and Mental Representation." Reprinted in Mental Representation: A Reader, eds. Stephen Stich and Ted A. Warfield. Blackwell, 1994. Pp.302-311.

Hunt, Earl and Franca Agnoli. "The Whorfian Hypothesis: A Cognitive Psychology Perspective." *Psychological Review*, vol.98, no.3, 1991. Pp.377-389.

Hylton, Peter. Russell, Idealism, and the Emergence of Analytic Philosophy. Oxford, 1990.

Imai, Mutsumi and Dedre Gentner. "A Cross-Linguistic Study of Early Word Meaning: Universal Ontology and Linguistic Influence." *Cognition*, vol.62. Elsevier Science, 1997. Pp.169-200.

Jackson, Frank. "Epiphenomenal Qualia." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.556-563.

----. "What Mary Didn't Know." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.577-580.

----. "Mental Causation." *Mind*, vol.105. Oxford, 1996. Pp.377-413.

Kant, Immanuel. Critique of Pure Reason. Trans. Norman Kemp Smith. Macmillan, 1929.

Kay, Paul and Willett Kempton. "What is the Sapir-Whorf Hypothesis?" *American Anthropologist*, vol.86, no. 1, March, 1984. Pp.65-79.

Kim, Jaegwon. "What is 'Naturalized Epistemology'?" *Philosophical Perspectives*, vol.2, 1988. Pp.381-405.

----. "Naturalism and Semantic Normativity." *Philosophical Issues*, vol.4, 1993. Pp.205-210.

----. Philosophy of Mind. Westview, 1998.

----. "The Myth of Nonreductive Materialism." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.452-464.

Kitcher, Patricia. "In Defense of Intentional Psychology." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.215-225.

Knowles, Jonathan. "The Language of Thought and Natural Language Undertsanding." *Analysis*, vol.58, no.4, 1998. Pp.264-72.

Kripke, Saul A. Naming and Necessity. Harvard, 1972.

----. Wittgenstein on Rules and Private Language. Harvard, 1982

Lee, Penny. The Whorf Theory Complex. John Benjamins, 1996.

Leeds, Stephen. "Qualia, Awareness, Sellars." *Nous*, vol.27, no.3, September, 1993. Pp.303-330.

Lewis, David. "Mad Pain and Martian Pain." Readings in the Philosophy of Psychology. Vol.1, ed. Ned Block. Harvard, 1980. Pp.216-222.

Libet, Benjamin. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *The Behavioral and Brain Sciences*, vol.8, 1985. Pp.529-566.

Loux, Michael J. Substance and Attribute. D. Reidel, 1978.

Lucy, John A. Language, Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis. Cambridge, 1992.

Lycan, William. Consciousness. MIT, 1987

----. "Ideas of Representation." Mind, Value and Culture: Essays in Honor of E.M. Adams, ed. David Weissbord. Atascadero: Ridgeview, 1989. Pp.207-228.

----. "A Deductive Argument for the Representational Theory of Thinking." *Mind & Language*. Vol.8, no.3. Blackwell, 1993. Pp.404-422.

----. "Consciousness as Internal Monitoring, I." *Philosophical Perspectives*, vol.9, 1995. Pp.1-14.

----. Consciousness and Experience. MIT, 1996.

Marks, Charles E. Commissurotomy, Consciousness and Unity of Mind. MIT, 1980.

Marras, Ausonio. "Sellars on Thought and Language." *Nous*, vol.7, no. 2, May, 1973.

----. "On Sellars' Linguistic Theory of Conceptual Activity." *Canadian Journal of Philosophy*. Vol.II, no. 4, June 1973. Pp.471-483.

McDermott, Michael. "Quine's Holism and Functionalist Holism." *Mind*, vol.110. Oxford, 2001. Pp.977-1025.

McDonough, Richard. "Wittgenstein's Reversal on the 'Language of Thought' Doctrine." *Philosophical Quarterly*, vol.44, n.177, 1994. Pp.482-494.

McDowell, John. Mind and World. Harvard, 1994.

----. "The Content of Perceptual Experience." *Philosophical Quarterly*, vol.44, no.175, 1994. Pp.190-205.

----. Mind, Value, and Reality. Harvard, 1998.

Millikan, Ruth Garrett. "Biosemantics." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.401-411.

Nagel, Thomas. "What is it Like to be a Bat?" Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.534-542.

Neander, Karen. "The Division of Phenomenal Labor: A Problem for Representational Theories of Consciousness." *Philosophical Perspectives*, vol.12, 1998. Pp.411-434.

Nelson, Lynn Hankinson and Jack Nelson. On Quine. Wadsworth, 2000.

O'Shea, James R. "On the Structure of Sellars' Naturalism with a Normative Turn." Draft version, www.philosophy.sas.ac.uk/EPM_OShea.doc.

Osherson, Daniel N. and Howard Lasnik, eds. Language: An Invitation to Cognitive Science, vol.1. MIT, 1990.

Papineau, David. "The Teleological Theory of Representation." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.416-424.

Pinker, Steven. The Language Instinct. William Morrow, 1994.

----. How the Mind Works. Norton, 1997.

Putnam, Hilary. "The Nature of Mental States." Originally published as "Psychological Predicates," in Art, Mind, and Religion, eds. W.H. Capitan and D.D. Merrill, 1967. Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.102-109.

----. "The Meaning of 'Meaning'." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.279-285.

----. The Threefold Cord: Mind, Body, and World. Columbia, 1999.

Pütz, Martin and Marjolijn H. Verspoor, eds. Explorations in Linguistic Relativity. John Benjamins, 2000.

Quine, W.V.O. From a Logical Point of View. Harvard, 1953.

----. Word & Object. MIT, 1960.

----. The Ways of Paradox and Other Essays. Harvard, 1966.

----. Ontological Relativity and Other Essays. Columbia, 1969.

----. Pursuit of Truth. Revised Edition. Harvard, 1992.

----. From Stimulus to Science. Harvard, 1995.

Ristau, Carolyn. "Cognitive Ethology: Past, Present and Speculations on the Future." *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol.2, 1992. Pp.125-136.

Rorty, Richard. "In Defense of Eliminative Materialism." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.178-183.

Rosenberg, Jay F. Linguistic Representation. D. Reidel, 1974.

Rowlands, Mark. "Teleological Semantics." *Mind*, vol.106. Oxford, 1997. Pp.271-303.

Rumelhart, David E., et al., eds. Parallel Distributed Processing. MIT, 1986.

Ryle, Gilbert. The Concept of Mind. Chicago, 1949.

Sapir, Edward. Language: An Introduction to the Study of Speech. Harcourt, 1921.

Searle, John R. The Rediscovery of the Mind. MIT, 1992.

----. "Animal Minds." Midwest Studies in Philosophy, vol.19, 1994. Pp.206-219.

Sellars, Wilfrid. "Language, Rules and Behavior." John Dewey: Philosopher of Science and Freedom, ed. Sidney Hook. The Dial Press, 1949. Pp.289-315.

----. Empiricism and the Philosophy of Mind. Originally published in Minnesota Studies in the Philosophy of Science, vol. 1, eds. Herbert Feigl and Michael Scriven. Minnesota, 1956. Reprint, Harvard, 1997.

----. "Abstract Entities." *Review of Metaphysics* 16, 1963. Pp.627-671.

----. "Empiricism and Abstract Entities." The Philosophy of Rudolf Carnap. Ed. Paul Schilpp. La Salle : Open Court, 1963. Pp.431-468.

----. "Philosophy and the Scientific Image of Man." Science, Perception, and Reality. Ridgeview, 1963. Pp.1-40.

----. "Some Reflections on Language Games." Science, Perception, and Reality. Ridgeview, 1963. Pp.321-358.

----. Science, Perception, and Reality. Ridgeview, 1963.

----. Science and Metaphysics : Variations on Kantian Themes. Humanities, 1968.

----. "Language as Thought and as Communication." *Philosophy and Phenomenological Research*. Vol.29, 1969. Pp.506-527.

----. "Reply to Marras." *Canadian Journal of Philosophy*. Vol.II, no.4, June 1973. Pp.485-493.

----. "Meaning as Functional Classification." *Synthese* 27. D. Reidel, 1974. Pp.417-437.

----. "The Structure of Knowledge: (1) Perception; (2) Minds; (3) Epistemic Principles." Action, Knowledge and Reality: Studies in Honor of Wilfrid Sellars, ed. Hector-Neri Castañeda. Bobbs-Merrill, 1975. Pp.295-347. Presented as The Matchette Foundation Lectures for 1971 at the University of Texas.

----. Naturalism and Ontology. Ridgeview, 1979.

----. "More on Givenness and Explanatory Coherence." Justification and Knowledge, ed. George Pappas. D. Reidel, 1979. Pp.169-182.

----. "Behaviorism, Language and Meaning." *Pacific Philosophical Quarterly* 61, 1980. Pp.3-30.

----. Pure Pragmatics and Possible Worlds. Ed. Jeffrey F. Sicha. Ridgeview, 1980.

----. *Selection From* "The Structure of Knowledge." Reprinted in The Nature of Mind, ed. David M. Rosenthal. Oxford, 1991. Pp.372-379.

Sellars, Wilfrid and Roderick M. Chisholm. "The Chisholm-Sellars Correspondence on Intentionality." Intentionality, Mind, and Language, ed. Ausonio Marras. Illinois, 1972. Pp.214-248.

Shahan, Robert W., and Chris Swoyer, eds. Essays on the Philosophy of W.V. Quine. Oklahoma, 1979.

Skinner, B.F. *Selection From* About Behaviorism. Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.59-67.

Smart, J.J.C. "Sensations and Brain Processes." Reprinted in Problems in Mind, ed. Jack S. Crumley II. Mayfield, 2000. Pp.81-90.

Sperry, R.W. "Hemisphere Deconnection and Unity in Conscious Awareness." *American Psychologist*, 1968. Pp.723-733.

Stalnaker, Robert. "What is the Representational Theory of Thinking? A Comment on William G. Lycan." *Mind and Language*, vol.8, no.3. Basil Blackwell, 1993. Pp.423-430.

Steiner, George. After Babel. Oxford, 1975.

Sterelny, Kim. The Representational Theory of Mind. Basil Blackwell, 1990.

----. "Basic Minds." *Philosophical Perspectives*, vol.9, 1995. Pp.253-270.

Stich, Stephen. "What is a Theory of Mental Representation?" Reprinted in Mental Representation: A Reader, eds. Stephen Stich and Ted A. Warfield. Blackwell, 1994. Pp.347-364.

Stich, Stephen and Ted A. Warfield. "Introduction." Mental Representation: A Reader. Blackwell, 1994. Pp. 1-8.

Strawson, P.F. "Meaning and Truth." (1970). Reprinted in The Philosophy of Language, 4th edition, ed. A.P. Martinich. Oxford, 2001. Pp.110-121.

Tye, Michael. "A Reprsentational Theory of Pains and Their Phenomenal Character." *Philosophical Perspectives*, vol.9, 1995. Pp.223-239.

Vinueza, Adam. "Sensations and the Language of Thought." *Philosophical Psychology*, vol.13, no.3, 2000. Pp.373-392.

Weller, Cass. "Bonjour and Mentalese." *Synthese*. Vol. 113. Kluwer, 1997. Pp.251-263.

Whorf, Benjamin Lee. Language, Thought, and Reality: Selected Writings. Ed. John B. Carroll. MIT, 1956.

Wittgenstein, Ludwig. Philosophical Investigations. 3rd Edition. Trans. G.E.M. Anscombe. Basil Blackwell, 1958.

----. On Certainty. Ed. G.E.M. Anscombe and G.H. von Wright. Basil Blackwell, 1969.

Yablo, Stephen. "Mental Causation." *The Philosophical Review*, vol.101, no.2, 1992. Pp.245-280.

Ziff, Paul. "On H.P. Grice's Account of Meaning." Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology, eds. Danny D. Steinberg and Leon A. Jakobovits. Cambridge, 1971. Pp.60-65.

VITA

Ben Stenberg received his Bachelor of Arts degree in Philosophy and English from Whitman College in 1997. He earned his Doctor of Philosophy degree in Philosophy from the University of Washington in 2006.