INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

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 bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI 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.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

**Crossing Divides:
Avoiding Incompatibility in Semantic and Cognitive Theories
by
Alex Taylor, B.Sc.**

**A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of**

Master of Arts

**Carleton University
Ottawa, Ontario
September 13th, 1998
© 1998, Alex Taylor**

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-48413-0

Canada

The undersigned recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis

“Crossing Divides:
Avoiding Incompatibility in Semantics and Cognitive Theories”

submitted by Alex Taylor, B.Sc.
In partial fulfillment of the requirements for
the degree of Master of Arts

Chair

Department

Thesis Supervisor

Abstract

The purpose of the thesis was to explore methods of avoiding the apparent incommensurability between 1) semantic atomism and semantic holism and 2) connectionism and classical artificial intelligence. The argument depends on first, reconciling connectionism and classical artificial intelligence, and second, exploiting 1) some features of the structure of the representations in connectionist networks and in classical artificial intelligence and 2) certain facts about supervenient properties in cognitive science, to suggest that atomism and holism are necessarily compatible.

Keywords: semantics, mental representation, artificial intelligence, neural networks, supervenience.

Acknowledgements:

The author would like to acknowledge the generous help of the following: First, his thesis supervisor, Robert Stainton, and also the following people: Andrew Brook, Director of the Institute of Interdisciplinary Studies, István Berkeley at the University of Southwestern Louisiana, Don Ross at the University of Cape Town, and John Poland in the School of Mathematics and Statistics. The project would have been much more difficult without them.

Table of Contents:

Introduction	1
Chapter 1: Semantic Holism and Semantic Atomism	9
1.1 Atomism	9
1.11 Three Claims	9
1.12 Four Cases	9
1.13 Definitions	13
1.14 Problems	17
1.2 Radical Holism	27
1.21 Quine	27
1.22 Davidson	33
1.23 Definitions	36
1.24 Problems	38
1.3 Mild Holism	40
1.31 Devitt	40
1.32 Block	43
1.33 A Definition	45
1.34 Problems	46
1.4 Conclusions	54
Chapter 2: Connectionism and Classical A.I.	56
2.1 Characterization of the Two Architectures	57
2.11 Classical A.I.	57
2.12 Connectionism	60
2.13 Cognitive Levels	64
2.2 The Fodor/Pylyshyn Dilemma	73
2.3 Responses to the Dilemma	75
2.31 A Geography of Replies	75
2.32 Beating the Dilemma	77
2.33 Reply 1: Anti-Representationalism	78
2.34 Reply 2: Distributed Representations	86
2.35 Reply 3: Classical Representations	110
2.4 Conclusions	113
Chapter 3: Reconciling Connectionism and Classical A.I.	119
3.1 Hybrid Systems	119
3.2 One Consequence of Hybrid Systems	126
3.3 Connectionism and Semantics	131
Bibliography	139

Introduction

There are two main discussions in this thesis. The first discussion is about holism and atomism in semantics, particularly the semantics of mental representations. It is about why atomism and holism are distinct and seemingly incompatible approaches to semantics. The second discussion has to do with whether or not the mind ought to be viewed like a classical computer, or alternatively if a connectionist approach to modelling cognitive activity might be more appropriate. What connects these two stories is the idea of cognitive levels and supervenience. Using Pylyshyn's nomenclature, both connectionism and what is meant by a classical computer model can be described as entities of the *algorithmic* level of cognition. Semantic entities are so called *cognitive* level entities. They are supervenient on the algorithmic level, which in turn is supervenient on the implementational level, i.e. the brain. The central claim of the thesis is that: First, we don't have to pick between the classical approach to cognition, and the connectionist approach to cognition, indeed it would be inappropriate to do so, as neither alone constitute an adequate cognitive model. Second, because we don't have to choose between classical models and connectionist models of cognition, it would be wrong to claim that atomism and holism are exclusive alternatives for semantics.

In Chapters I and II, I establish a variety of claims and premises. In Chapter III I argue for the central claim. For reference, I will briefly outline the argument for the main claim. The first premise of the argument is that subvenient entities and properties entail the properties and entities that supervene on them. Thus algorithmic entities entail semantic

entities. Second premise: The semantics entailed by connectionism is holistic and the semantics entailed by classical models is atomistic. Third premise: Both classical models and connectionist models are true of human cognition. Conclusion: Both holism and atomism with respect to semantics are entailed. Thus it is possible for them to be compatible.

Chapter I concerns itself with semantics. This chapter had two main objectives. The first objective was to define what is meant by holism and atomism, and intermediate positions between these two. This wasn't straightforward. What different authors mean by these words are often quite different things, worse, many authors don't provide any explicit explanation at all, relying instead on metaphor to explain their views. Yet without good definitions, it is much more difficult to either assess or make arguments about holism and atomism as classes of semantic theories, even though many such arguments have been offered. Fodor and Lepore (1992) do provide explicit definitions of holism and atomism, but in my view they are unsatisfactory.

However holism and atomism are ultimately defined, the consequences that are attributed to holism or atomism, good or bad, ought to follow from the definition in a reasonably obvious way. This entails that those consequences be described and justified. I spent some time doing this, particularly for atomism. Furthermore, we want those definitions to encompass all of the people that are normally thought to hold a particular view. Otherwise talk about holism, in general, may simply amount to changing the subject of debate when applied to a particular author. Thus, a definition of holism ought to apply equally to the semantic theories of Quine and Davidson, for example. The consequences

that Fodor and Lepore attribute to holism and atomism are not generally thought to follow from the definitions that they provide. I also argue that Fodor and Lepore's definition doesn't meet the second requirement either. It is too permissive. There is a class of views — those held by Block, Devitt and Field, in particular — that could fall under Fodor and Lepore's definition of atomism, even though neither Fodor and Lepore, nor the authors in question, would call these views atomistic. My own definitions were constructed with these two desiderata in mind.

The second task of the first chapter was to assess the claim that holism and atomism are exclusive alternatives. There are two obvious ways to reconcile these positions. One is to split a semantic theory into parts, where some of the parts are holistic, and some of the parts are atomistic. Block tries to do this with his two-factor theory, and we might imagine other ways of doing this as well. For example, we might use an atomistic approach for some kinds of words and a holistic semantics for others. The second way of reconciling atomism and holism is to have a hybrid theory that is half-way in between. This is what Devitt tries to do. Fodor and Lepore claim that these "middle ground" positions amount to full-blown holism, at least as far as they inherit all the bad consequences of full-blown holism. The central issue here is the requirement that semantics employ compositional explanations of sentence meaning. Holism, it seems, must give up on compositionality.

While I agree with Fodor and Lepore on this point, I make a further claim. None of the grounds so far given, either for or against holism or atomism, constitute adequate grounds for choice between the two. This is important because, in spite of their disclaimer

to the contrary¹, Fodor and Lepore really *are* making a disjunctive argument for atomism, as they seem to believe all three of these claims: 1) There are valid arguments against semantic holism 2) There are no valid arguments *for* semantic holism (the stated thesis of “Holism: A Shopper’s Guide”) and 3) There is no such thing as only a little holism, and hence no middle ground between holism and atomism. An argument for atomism is immediately entailed, constructed as follows: either semantics is holistic or atomistic. Holism isn’t plausible, therefore atomism is the correct approach to semantics. Unfortunately, a similar disjunctive argument may be as readily made against atomism. Clearly, both cannot be right. What seems to make each of these arguments plausible, and therefore a choice of theory plausible (i.e. holism or atomism) is that in each case, the author makes their point only by restricting the evidence that they attend to. Fodor and Lepore, for example, pay a great deal of attention to compositionality, and to certain kinds of words, such as nouns, while ignoring other kinds of words and semantic phenomenon. When the evidence is taken as a corporate body, both approaches to semantics seem to be inadequate. This make reconciliation between the two a worthwhile problem.

Chapter II focusses on the debate between connectionism and classical computation as models of psychological explanantion. First, the classical approach and connectionism are described. With respect to connectionism I employ Smolensky’s definition of a connectionist network. The reason for this is that first, a lot of the philosophical literature implicitly or explicitly refers to this kind of network, and second because Smolensky is trying to develop connectionism as an alternative to classical

¹In the Introduction to “Holism: A Shopper’s Guide”

computation as a psychological model. Smolensky argues that others formulations of connectionism are not alternatives to classical computation. The key difference between connectionist neural networks, and non-connectionist neural networks is the style of representation employed by the network. That is, whether it is distributed or local.

Having stipulated this definition two issues follow. First, the notion of a cognitive level is central to the main claim of the thesis, in particular, as it is important for me that 1) cognitive level entities (most semantic entities) supervene on algorithmic level entities (connectionist networks and/or classical computers), and 2) that connectionism and classical machines are compatible algorithmic entities. Smolensky insists that connectionist networks are not algorithmic level entities. I claim that they are. If Smolensky is correct, then reconciliation between connectionism and classical computation is a moot point. Rather, the question is can connectionist networks implement classical computation. Second, Smolensky and others talk about these levels as descriptions. As I later made use of the fact that semantics supervenes on the algorithmic level, talk about levels as descriptions won't do. I therefore needed to argue that Smolensky ought to be a realist about cognitive levels. By implication this also applies to anyone that endorses Smolensky's definition of connectionism. I argue that supervenient cognitive properties requires realism about cognitive levels. As Smolensky exploits the idea of supervenience he ought to endorse the realism as well.

Having described the two approaches to psychological modelling, the next thing was to describe the dispute about the two approaches. In this case the dispute is between Fodor and Pylyshyn, and connectionists who endorse what Smolensky says. Fodor and

Pylyshyn claim that either connectionist networks merely implement classical architectures, or they are neural (hardware) level description. In either event they are of not useful as alternative cognitive models. As for those connectionist networks that do not implement classical architectures, Fodor and Pylyshyn claim that lack the syntactically structured representations required for phenomenon such as systematicity, and are therefore inadequate as models. Establishing connectionism as a distinct and viable, but compatible, alternative is a necessary premise in supporting the main claim, therefore Fodor and Pylyshyn's conclusions must be resisted.

In the philosophical literature, there are a variety of negative responses to Fodor and Pylyshyn's attack on connectionism. In the main, these responses focus on the issue of representation in the connectionist network. There were three positions on this issue. The first and perhaps most radical position is that connectionist networks don't have and don't need representations. The second position is that connectionist representations exist, but they don't have constituent structure. There are two varieties of this position. Authors such as Waskan and Bechtel don't feel that constituent structure is required. Others, such as Hinton, believe that structured representations are indeed necessary, and are seeking ways to create connectionist networks with distributed and structured representations. Lastly, some authors (e.g. Clark) hold that the representations in networks are precisely the same as those in classical machines.

I contend that most of these responses to Fodor and Pylyshyn fail, save for the position that connectionists must find ways of making networks that have appropriately structured representations. And this position, to some extent, acknowledges Fodor and

Pylyshyn's point, but without drawing as pessimistic a conclusion as they do. Where I think Fodor and Pylyshyn go wrong is that they fail to recognize that while there are problems with connectionist models, these models can model certain cognitive processes more naturally and robustly than classical models can. Furthermore, the classical model has serious problems of its own. An argument against a cognitive model on the grounds that it lacks certain capabilities, is an argument against both connectionist and classical models, not one alternative alone. The truth, I think, is that both models are inadequate by themselves.

If I am right so far, I have established that connectionism is a viable and distinct alternative to classical computation as a model of cognition, and that it is indeed an algorithmic level explanation. The next step is to show that connectionism and classical systems aren't incompatible. The easiest and best way to do this is with examples of systems that employ both architectures or are true hybrids. There are examples of both, and while these hybrid systems are limited in scope, there has been fairly impressive progress on this front, and no reason to believe that the problems that remain are insurmountable. As this thesis is also to a large extent a commentary about Fodor's philosophy of mind, I would be remiss if I didn't point out at this juncture that one consequence of a hybrid architecture is that we must give up on the modularity thesis, because the modularity thesis insists in peripheral and central components to cognition, each of which possesses a pure classical, and pure connectionist architecture respectively.

In the remainder of Chapter III, I establish the last two premises that I need to argue for the central claim of the thesis. First I argue that supervenient properties are

entailed by subvenient entities. Thus, for example, quantum mechanics entails thermodynamics via the correspondance principle. With respect to cognition, this means that algorithmic level entities entail semantic level entities. Given that both connectionism and classical systems seem to be both viable algorithmic entities, what kind of semantics do they entail? We know from previous discussion that connectionist systems don't easily support constitutive representations, therefore it is unlikely that an atomistic semantics, which requires constituent structure, is an emergent property of a connectionist system. An analogous argument is constructed for classical systems and holistic semantics. I then give reasons to suggest that an atomistic semantics would be most naturally implemented in a classical architecture and a holistic semantics would be most naturally implemented in a connectionist system. This suggests that if classical architectures are true of human cognition, the semantics they entail are atomistic. Similarly for connectionism and holism. Given that both connectionism and classical computation seem to be true, it follows that both kinds of semantics are entailed. This contradicts the conclusions of Chapter I, which suggested, that holism and atomism are mutually exclusive. So which conclusion is correct? As a general methodological principle for cognitive science, I endorse the following: That no claim about a supervenient property may be inconsistent with the subvenient entities. So for example, no claim about vision could be inconsistent with the known physics of the eye. The claim that semantic holism and semantic atomism are incompatible is inconsistent with the claim that their subvenience bases are compatible. So even though we cannot see how to make a semantic level system that reconciles atomism and holism, it is entirely plausible that the sought for "middle ground" exists.

Chapter 1: Semantic Holism and Atomism

1.1 Atomism:

Holism being as tricky a concept as it is, I propose to start a discussion of holism and atomism in semantics, by first talking about atomism. While holism and atomism are each burdened by hard problems, atomism, I think, is intuitively easier to understand as a concept. The purpose of this discussion is threefold. First, the discussion will hopefully clarify what is meant by holism, which I first define in contrast with atomism, second it may elucidate what might motivate a holistic approach to semantics, and finally, it may clarify some of the problems with each of these approaches to semantics.

1.1.1 Three Claims:

Atomism is committed to three things. First, atomism is committed to the idea that the meaning of a word is determined by its reference. Second, atomists are committed to the idea that words may be learned and understood independently of what other words mean. Third, word meaning is logically prior to sentence meaning.

1.1.2 Four Cases:

The first claim must be defended in four cases. First, contemporary atomists adopt a view of word meaning that is either 1) Russellian (1919, 1956; see also Neale 1990) or 2) Fregean (1892, 1918; see also Evans, 1980). Second, these views must be examined in cases where 3) there is a standard reference, and 4) those cases where there is not.

Russell's view of word meanings is initially straight forward. Word meaning consists in the extension or reference of the word. Suppose that I want to know what the

word “ethology”² means. I might say that the word means just *that thing* that biologists do when they study animal behavior. Alternatively, the extension, and therefore meaning of the word “red” is the set of red objects. Thus, meaning is a relation between words and things in the world, with which, Russell says, we are directly acquainted.

Alternatively one might suppose that it is not the thing per se, that determines what “ethology” means, but rather a *mode of presentation*, or a sense, which in turn determines reference. The classical example of a mode of presentation is owed to Frege, of course, who came up with the idea of sense (Frege, 1892). Frege’s example talks about “the morning star” and “the evening star”. Both words have the same referent, but each has a different sense; so we learn something new and interesting when astronomers discover that both ‘stars’ are really just Venus. It would be wrong therefore, to claim that “morning star” and “evening star” are synonymous, as Russell might have to do. Evans explains sense in the following way (1982, p. 16):

Frege’s idea was that it may be a property of a singular term as an element of a public language that, in order to understand utterances containing it, one must not only think of a particular object, its Meaning, but one must think of that object *in a particular way*: that is, every competent user of the language who understands the utterance will think of the object in the same way.

The basic style of explanation is similar to Russell’s, in the sense that the semantic value of the sentence ultimately depends on a specific thing in the world. For Frege, however, semantics isn’t directly denotational: meanings depend upon senses, which in turn depend upon objects. Thus Frege’s is a *mediated* theory of reference.

²Mention is indicated by double quotes. Use is indicated by single quotes.

Words and phrases may also be non-denoting. Russell's favorite example is 'the present King of France'. This singular term does not refer to any thing in the world, because there is no 'King of France'. Russell proposes to treat expressions with no obvious reference, such as empty singular terms, as quantifiers.

Frege, on the other hand, was willing to ascribe a sense to empty singular terms. He therefore ascribed a sense to an expression *with no reference*. Again I turn to Evans. Evans points out that the notion of empty singular terms having senses is inconsistent on any classical Fregean explanation of what sense is. Such an explanation must establish the relationship between sense and reference. Frege's having established that the semantic value of a sentence is its truth value as one of the anchors of his semantics (the other anchor being the idea that singular terms refer to objects), there exists a tension between the former claim, and the claim that there is a fragment of intelligible discourse which has no truth value (e.g. sentences containing empty singular terms) and therefore no semantic value. The tension is to be resolved by adopting the Russellian view on empty singular terms, namely they are in fact descriptions, so that sentences containing them do indeed possess a truth value, e.g., that they are false. What follows is that sense can now be cashed out as an aspect of, that is, a way of thinking about, a reference, whether this reference is a particular object (i.e. for a singular term), or a truth value (for a sentence)(Evans, 1982: sections 1.3, 1.6, 1.7). This explanation makes clear how sense and reference interrelate, while the alternative, treating sense as wholly independent from reference, makes this interrelation difficult to fathom, particularly in the context of Frege's larger views.

Given this as background, an important point about atomism can be made:

Atomism is ultimately neutral with respect to the issue of denotational (Russellian) or non-denotational (Fregean) semantics. Both accounts of meaning are (so far) atomistic, because on both accounts meanings determine a unique reference (though, as Frege showed, references needn't determine a unique "meaning": One referent, e.g. the planet Venus, can have multiple senses). Meanings, however, *are* supposed to determine unique references, whether they are mediated through a sense or directly to the world.

Another essential feature of atomism is that words can be understood in isolation. That is, meanings come from a single word, single thing relation, not from a) a word, other word relation or b) a relation between a cluster of words and a cluster of things (and in the extreme case, a whole language, whole world relation). Thus, the semantic atomist claims that I can understand perfectly well what you mean by "cat" with out knowing what you mean by "mammal". I can even understand what you mean by "morning star" without knowing what you mean by "evening star", for instance, if I consistently go to sleep at three in the afternoon, and haven't ever seen an 'evening star'. Alternatively, a word might be understood as denoting a specific function, for example, taking two truth values as arguments and outputting a truth value. I don't, however, need to know what "cow" or "if" means in order to understand "and", even though words that denote functions operate only in the context of sentences.

It is also characteristic of atomism that not only are words understood in isolation, but word meanings are logically prior to sentence meanings (claim iii). The meaning of a complex expression like a sentence, defined as a truth value, is determined by the

meanings of the individual words that it contains, and their method of arrangement. I don't *need* to know what other sentences mean, but I do need to know what the words that compose the expression mean, and how their arrangement contributes to the meaning of the whole. In an expression like "the cat is on the mat", the hearer needs to know how the positions of "mat" and "cat" etc. contribute to the meaning of the expression; for example, if the position of these two particular words are exchanged, the sentence meaning changes substantially. Some word arrangements would render the expression incomprehensible. So not only are word meanings logically prior to sentence meanings, but so are their lawful syntactic relations. The principle of compositionality, so called, is at first blush, necessarily atomistic³.

1.13 Definitions:

One recent definition of atomism concerns the possibility of punctate languages. This definition originates with Fodor and Lepore. A punctate language is a language with only one word in it. So a semantics is atomistic if, within that semantic theory, a punctate language is possible. This definition is important, because the book that it appears in, *Holism: A Shopper's Guide*, has been the *locus classicus* for much of the discussion regarding holism and semantics since. However, I think the definition has serious flaws⁴.

What motivates Fodor and Lepore's definition are the same sorts of considerations

³I discuss this at length in part iii of this chapter

⁴

I am not alone in thinking this. See, for example, Devitt (1996), ch. 1, Boghossian (1994), Block (1993), and for a particularly trenchant remark about this definition, Heale (1994). As I am not trying to defend holism, my issues with Fodor and Lepore with respect to atomism are quite different, however.

that I've been discussing, that word meanings consist in their reference, and so forth. If these sorts of considerations are right-headed, then it seems that we can infer that, necessarily, the possibility of a punctate language follows. However, I think they have taken a sufficient condition to be necessary. Specifically, while I would agree that any punctate language is thereby atomistic — an un-exciting conclusion if there ever was one — I don't think that, necessarily, any atomistic semantics must allow the possibility of a punctate language. There are a host of other constraints on semantic explanations that might preclude punctate languages, at least in semantics for natural languages (I include mentalese), without really impacting on the extent to which a theory is atomistic. For one thing, one of the things that we might want to account for in a semantic theory is the meaning of complex expressions. This presumes that we are already talking about a language with multiple words, but this shouldn't alarm anyone — most semantics do presume this. The question is whether an atomistic semantics allows the *possibility* of a punctate language, not whether the language already actually possesses more than one word. The fact that a language has multiple words won't rescue Fodor and Lepore's definition of an atomistic language, however. The reason why relates directly to the need for explaining the meaning of complex expressions.

Typically, the principle of compositionality is employed to explain the semantics of complex expressions. But the principle of compositionality seems to preclude the possibility of a language even with a single sentence, let alone the more radical possibility of a language with a single word. Any compositional explanation of what "The dog bit the man" means also automatically entails the possibility of the expression "The man bit the

dog”, which contains precisely the same elements, only arranged differently. The lawful arrangement of the latter sentence is in fact prescribed and predicted by the syntax. It would seem then, that any semantic theory employing compositionality is already non-atomic. Ironically, the topic of systematicity, of which this is an example, is one of Fodor’s favorite topics of discussion.

At an even more fundamental level, while we might consider the possibility of a punctate language consisting in the word “cow”, it seems unlikely that anyone would allow punctate languages consisting of “and” or other sentential connectives. Nouns for anatomic concepts such as numbers, e.g. “five” probably won’t work either. Does this then mean that any semantic theory that provides an account of the meaning of logical connectives (or indexicals, or particles, or ...) is automatically holistic? Probably not. More plausibly, to say that a semantics is atomistic may merely mean that it is consistent with the possibility of a punctate language for at least one word in the language. The trouble is, if we accept this we aren’t left with much of a criterion for a definition. While such a definition would preclude the kind of radical holism advocated by Quine or Davidson, who in some sense deny individual word meanings altogether, it’s unlikely to exclude milder holists such as Field (1977), Block (1986) and Devitt (1996), who do think that it is both useful and meaningful to talk about the meanings of single words. The criterion of Fodor and Lepore, thereby revised, is so weak that Field et al. could comfortably swallow Fodor and Lepore’s definition whole and claim that their semantics are as atomistic as that of Fodor or Dretske! But this statement would be obviously false.

Fodor and Lepore’s definition of atomism, has very little consequence for the

central arguments of their book⁵, and almost seems to miss the point. So I will put it aside and use my own criteria. To revisit: The essence of atomism is two-fold. First, the semantic atomist postulates that words are semantically primitive, and that there is some kind of a function that determines their reference. Second, the meaning atoms obtain their primary semantic significance *only* in virtue of this function; this also applies to complex expressions. For Russell for example, the meaning of a name is its reference. “Cat” means cat because it refers to that kind of a thing over there on the mat, not because it plays a certain conceptual role, or licences a certain T-sentence as the case may be. The word cat may enter into other semantic relations but these are dependant on the primitive relations. Call this the independence criterion. During lexical acquisition, I can later learn what mammals are, or that this particular cat is Weeble and the other one over there is Stir Fry. Indeed this may even change some of the semantic relations that I entertain when I say “cat”, but if we are good Russellians, say, this is irrelevant to determining the reference, and therefore meaning of the word “cat”. It might be worthwhile to formalize these notions in a definition:

Def’n (1): For every symbol S in a language L, the semantics of L are atomistic iff:

- 1) Complex meanings derive from simple meanings, and not *vice versa*.**
- 2) Primitive meanings come from reference: a relation between single words and a single kind of thing.**

By a simple meaning, what is meant is that the meaning has no sub-components. Complex meanings, in contrast, can be split up into smaller units. A primitive meaning identifies

5

Fodor and Lepore’s definition of atomism figures little in their book, but this is definitely not the case for the definition of holism they use.

which semantic items are first meaningful. Thus words must have meanings before sentences do, and sentences before paragraphs, and so on. The above definition, unlike Fodor and Lepore's, seems to capture the sorts of theories that are generally called atomistic, while excluding conceptual role semanticists of either Block or Devitt's stripe, as well as the radical holists.

1.14 Problems With Atomism:

Atomism has some problems, and it is these problems which, in part, ought to motivate a holistic semantics, or at least militate in favor of it. One of the main problems turns out to be that meaning does not exclusively determine reference in a number of non-trivial cases. Yet on either a Russellian or Fregean account of meaning, a unique reference is entailed. The classic paper on this theme is Putnam's "The Meaning of 'Meaning'", but I think that the basic problem goes well beyond the scope of Putnam's discussion of it. For example, in the broader area of philosophy of language and mind, this problem also crops up in the semantics of mental representation under the rubric of misrepresentation (Stich and Warfield, 1992: section 2.2). One example of misrepresentation in semantics proper occurs in instances of false speech (in fact it might be the only unproblematic example of misrepresentation in semantics, as distinct from the semantics of mental representations). Suppose an utterer says 'the dog over there has gray hair' when in fact 'the dog' is actually a wolf. The "dog" token is supposed to designate dogs, and in an atomistic semantics it should do so uniquely, that is, for example, only dogs should be in the extension, but in this instance the designation fails. It seems that the extension of "dog" contains not only dogs, but perhaps also wolves, coyotes and so on. What then does

“dog” mean? “Dog” and “wolf”, *pace* Russell/Frege do not mean the same thing, because they have different extensions. But if meaning is cashed out in terms of extension, then it seems that they might mean the same thing.

One way of avoiding this trouble is to make a distinction between the extension of the term itself, and those things to which that term is reasonably applied — the ‘appliees’ of the word. The appliees of ‘dog’, for example, are those things that prompt a reasonable agent to token ‘dog’. While it is clear that such a distinction can be made, it is not clear, however, that the atomist is entitled to it. What makes this distinction doubtful for an atomistic semantics is that it must have some kind of conceptual resource on which to base such a distinction, and it is very hard to imagine what this conceptual resource might be (for the atomist). One possible candidate are rules of use, but this will turn out to be problematic, for reasons I describe below. The effect of all this is that the concept of an appliee is undefined in an atomistic theory of semantics, and therefore the concept ought not to be used within the context of the formal semantic theory, or in defense of it, rather the notion of an appliee should be restricted to pragmatics.

The claim then, is that misrepresentation is a problem for the atomist, and not the holist. Usually, though, misrepresentation is seen to be a problem for everyone. I have described a quick example, but by itself, it isn’t very compelling; so, here’s an argument. First premise (**P1**): If meanings are individuated ‘individualistically’ then atomism invariably entails that meanings have exactly one reference (remember that without additional conceptual resources, I am claiming that Russellians or the Fregeans ought not make use of the concept of an appliee in their formal semantic theories). Second premise

(P2): atomism entails individualistic individuation. First conclusion (C1): atomism therefore entails that meanings have unique references. However (P3), if misrepresentation occurs, then it is not the case that meanings have unique referents. As misrepresentation does in fact occur, specifically, false speech occurs (P4), the second conclusion follows. The second conclusion (C2) is that atomism and misrepresentation are therefore inconsistent. Now, (P5) holism does not entail that meanings have unique extensions, thus (C3) holism and misrepresentation *are* thus far consistent.

Some of the premises in the above argument are going to generate argumentative heat, so I turn to defending them. The first premise was that if meanings are individuated individualistically then a consequence of atomism is that meanings entail unique referents. What makes this pressing is that every well-known contemporary semantics that might be called atomistic is essentially either Fregean or Russellian (e.g. Fodor, Dretske, Millikan, Stampe etc.), and all of them individuate meanings individualistically. Thus individualistic atomists are all either Russellians or Fregeans, for which statements may have either a standard reference, for example singular terms, or a non-standard reference, for example quantifier phrases. If these two things, taken all together, all entail a unique reference, then individualistic atomism entails unique referents. It is pretty clearly the case that in all four possible combinations (Russell or Frege, standard or non-standard reference) a unique reference *is* entailed, especially when Frege is interpreted via Evans.

So, why must atomists hold that meaning is individuated individualistically (the second premise)? First, if anti-individualism is true, then meaning specification makes reference to rules of use. This is Putnam's position, among others. Rules of use, in turn,

are anatomic — they involve a network of interrelated rules and facts about words, society and the world. If this is true, then, if meaning-specification for concepts makes reference to rules of use, and rules of use are anatomic, then concepts are anatomic also. Therefore, if anti-individualism is true, then concepts are anatomic. Finally, if we assume that concepts are atomic, then it follows immediately from this assumption and the previous result, that the concepts must be individualistic in an atomistic semantics.

The third premise (P3) follows because false speech is a kind of misrepresentation, indeed perhaps the only unproblematic kind of misrepresentation in semantics, and false speech entails that a meaning may have not have a unique reference.

Now, I have conceded that it is possible to avoid the consequences of the example I have given, if we can distinguish between the reasonable applications of a word, and the extension of the word. However, I maintain that an atomist cannot use the notion of an applyee without additions to the semantic scheme. One possibility for making the distinction is to exploit rules of correct use, but this won't do for the atomist, as long as holistic consequences result from recruiting such rules into the formal semantic theory.

Another way of avoiding the consequences of examples like false speech is to say that this is a question for pragmatics. To be sure, there are a number of problems about meaning that can legitimately be left to pragmatics to solve. This isn't one of them. The reason is that this would require pragmatics to possess the theoretical means to discriminate between the extension of a word, and the applyees of the word. If pragmatics could do this, however, then it could, in theory, completely specify the extension of a word, because, the extension is a proper subset of the applyees. And if it could do *that*,

there would be no need of semantics at all. Now in practice there are examples where there the extension of a word may not be a subset of the applyees, because there may be members of the extension, that for example, that aren't perceptible etc., and therefore aren't applyees. However, in the context of a theory, where statements apply in counterfactual circumstances (if I could see that meteor, then it *would* be an applyee) this isn't an issue. Note that I'm not claiming that pragmatics cannot help us determine the applyees — I'm quite sure that it can — what I'm claiming is that it cannot be used to tell the *difference* between the extension and the applyees; that must be left up to the semantic theory.

The fourth premise (P4) hardly needs defending — I have already provided an example — and neither does (P5). With respect to holism and mis-representation, one simply need recall authors such as Davidson, who insists that reference is inscrutable, or Quine, who questions the utility and perhaps even existence of individual word meanings. It is not obvious to me that misrepresentation is therefore a valid concept in this kind of a semantic scheme. Its hard to know what counts as getting a reference wrong, when reference is inscrutable, or what counts as getting the reference of a particular word wrong, when (if Quine is right) neither the reference or meaning of an individual word can reasonably be said to exist.

To sum up, misrepresentation entails a failure of meanings to determine a unique reference, yet the Frege/Russell conception of meaning seems to imply that there cannot be any such phenomena. What the atomist needs in order to rescue either the Fregean or the Russellian picture, is some sort of naturalistic account explaining these sorts of failures

(Stich and Warfield, 1994: Introduction). In *The Meaning of 'Meaning'* Putnam describes a number of systematic cases of error, some of which may be described as misrepresentational, and provides a solution to the problems he raises. Specifically, he suggest that the answer lies in anti-individualist rules of use (Putnam, 1975: p. 234):

Whenever a term is subject to the linguistic division of labour, the 'average' speaker who acquires it does not acquire anything that fixes its extension. In particular, his individual psychological state *certainly* does not fix its extension; it is only the sociolinguistic state of the collective linguistic body to which the speaker belongs that fixes the extension.

However, this does not look like its going to work for the atomist. A holist, however, can be an individualist and just not care about Putnam cases or any other form of semantic misrepresentation, because a holist does not derive meaning from reference, and therefore does not need to distinguish the referent of the term from the applyees. Thus the kind of contradiction that Fodor et al. face, namely that on their picture of semantics misrepresentation just shouldn't happen, but does happen, just does not face the holist, at least not so far as this argument goes. Indeed some holistic accounts, such as Davidson's, insist on the inscrutability of reference, and are plausibly read as individualistic. The problem, it seems, is atomism *in combination with* individualism. Atomism entails on the one hand that meanings have unique referents. Individualism, on the other hand, seems to entail that the meaning-reference connection can come apart, because words can reasonably be applied to objects that are not in their extension. And atomism entails individualism. Thus, each supplies one of the conditions that make some of Putnam's examples effective, and more generally, the conditions that make misrepresentation such a problem. As it happens, all contemporary atomists *are* individualists, which raises two

questions. First, does the individualist atomist have an effective means of dealing with misrepresentation, both for Putnam cases and more generally? Second, can the atomist abandon individualism? I argue that the answer in both cases is no. I have already argued the latter. Now to the former.

There have, to be sure, been a number of attempts to fix the kinds of problems that misrepresentation raises, of which Putnam cases are simply a special case. Millikan (1994) and Dretske (1981) both appeal to idealization, for example. For Dretske, when a particular syntactic item is acquired, special circumstances obtain. For Millikan, evolutionary idealizations are employed. Using her example, it is wrong to say that a frog really sees either flies or small black pellets, so it cannot misrepresent either. What a frog sees is a small black object, that when swallowed, reliably provides nourishment. For Dretske, learning environments are carefully constructed so that, as a learner, when I token “horse in a pasture”, there really is a horse in the pasture. Millikan explains error by appealing to the idealized function of the system that tokens the representation. Dretske explains error by appealing to idealized circumstances under which the representation is tokened. However, many atomists themselves remain unconvinced. Fodor, among others, has criticized both Dretske and Millikan’s proposed solution (Fodor, 1987: p. 102-6). Fodor himself has a solution on offer, asymmetric dependance (Fodor, 1994), but it too has been greeted with skepticism (Adams and Aizawa, 1994).

Another problem for an atomistic semantics is pointed out by Quine (1960: p. 141ff, p. 161ff). The problem, in a nutshell, is that certain sentences exhibit both *de dicto* and *de re* readings, and neither version of atomism can handle both: Russellian semantics

misses *de dicto* readings, while Fregean semantics misses *de re* readings. An example will help make the point. Suppose that Ralph says '*that man* is a spy'. Julie later reports 'Ralph says that Ortcutt is a spy'. The latter sentence is ambiguous. First, there is the *de dicto* reading (i.e. opaque): If Ralph does not know that the name of *that man* is "Ortcutt", the sentence is false. In the *de re* (i.e. transparent) reading, however, "Ortcutt" is simply used to designate some particular individual that Ralph happens to believe is a spy. In this case, substitutivity holds for any term referring to *that man*, and the sentence is true, even if Ralph does not realize that the man is named Ortcutt. If one holds, as atomists do, that sentence meanings depend entirely on the reference properties of the words that compose them, then clearly there is a problem. Here we have two possible meanings and it does not look as though the one resource that atomism has to account for it — syntactic variation — can generate both meanings.

The particular kind of ambiguity illustrated here can be handled with the resources of a holistic semantics. I describe one proposal in more detail in discussing Devitt's semantics later in the chapter. The lesson for now is that this is another example of a semantic problem that militates against the possibility of an atomistic semantics that completely describes meaning in a natural language.

What motivates the semantic atomist, like Fodor, is that atomism provides a powerful and intuitive tool for describing how certain classes of words (or concepts in the case of mentalese), manage to have meaning. Atomism seems especially well suited to describing many kinds of nouns and verbs. Indeed atomists, Fodor, Millikan and Dretske among them, rely very heavily on examples drawn from these classes of words or concepts

(perhaps *too* heavily). It also accounts well for certain generally observed features of language. Compositionality, for example, seems to be a central fact about language. The connection between compositionality, productive language use and systematicity is quite striking, and as Fodor points out, it is exceedingly difficult for the holist to account for it. This is because these three facts or observations about language depend on the interchangeability of a finite number of fixed meaning units — words — within a fairly rigidly specified syntactic structure.

What motivates the semantic holist, is conceptually quite a bit like what motivates the atomists. First, holism is also quite good at explaining certain categories of words and certain linguistic facts. In particular, holists are good at explaining syncategorematic expressions like particles e.g. ‘up’ in ‘shut up’, and the “functional categories”, such as determiners, case markers, and inflections. Second, holists are concerned about certain hard problems that face the atomists, and think that holism provides the answer. These types of words interact with contexts, so it seems *prima facie* difficult to describe what they mean except in relation to both the environment and the meaning other words. What the “that” points at in “that pig” depends on the content of “pig” in a specific context. Contrast this with how a logical connective does its work.

Having claimed that misrepresentation is a problem for atomists, it is however necessary to describe why this is not the case for semantic holists. Or, why holists think they have the answer to this problem. Largely, semantic holists don’t have to worry about how to map individual referents to words, or individual bits of information to specific representations. This makes it quite difficult to even pin down even what we might mean

by misrepresentation with reference to the holist. Certainly dog/fox worries don't arise because holists don't suggest that 'dog' gets its meaning from dogs. This is because word meanings are not independent functions, (from the world to a particular word, for instance) that uniquely specify that meaning. Generally, what a word might mean is something like the inferential or conceptual role that it plays in the whole language, or in mentalese if we happen to be talking about mental representations.

Where we are so far: First, a definition of atomism was provided. Second, I have tried to show generally the sorts of problems that atomism is good at solving, and the kinds of problems it is not good at solving. I then contrasted these features of atomism with holism. There are two things to note. First, by way of conclusion, it is important to see that both atomistic and holistic approaches provide an inadequate account of semantics thus far. Given that both approaches can lay claim to some relatively powerful explanatory apparatus, and both suffer from fairly serious explanatory deficiencies, it is really difficult to choose between them. Second, this constitutes only a first pass — of four — at holism. Even though this first attempt fairly reflects some of the standard mantras about holism in the literature, it is inadequate. As Christopher Peacock comments, much depends on precisely what is meant by "holism", and this turns out to be several different things. Where this is leading: In the second part of the chapter, I define a certain kind of holism. Specifically, the kind of radical holism endorsed by Davidson and Quine. Again, this is the logical place to begin, because this form of holism is the polar opposite to atomism, and therefore easier to compare with it. I also talk about some of the generally accepted conclusions about radical holism, in particular, why many interlocutors

in the debate — perhaps even most of them — find radical holism so problematic. Largely, I accept these conclusions. Given that radical atomism and holism are so problematic, why not take a moderate position? The purpose of section three is to evaluate the prospects of the middle position.

1.2 Radical Holism

For many uses, I am concerned here with the work of Davidson (1967, 1984) and Quine (1953, 1960). Both start from the premise that understanding the semantics of a language places the listener in the role of a radical interpreter. A radical interpreter has two very special features. First, a radical interpreter begins the process knowing nothing about their informants' language. Second, the only evidence that is available to the radical interpreter is their informants' behavior and environment. Setting aside for the moment the arbitrary nature of this latter assumption, I will turn first to how the radical interpreter accomplishes her goal, and what this has to do with holism.

1.21 Quine

For Quine, understanding the semantics of a language is of a piece with constructing a scientific theory. So, first point: The job of the theory builder or the radical interpreter, as the case may be, is to maximize the consistency of the theory with all of the available evidence, as a group. Quine likens a theory, or a language (I will use the two interchangeably while talking about Quine) to a field or a network, in which each available piece of evidence is a node, and the connections of the nodes are like the relations among each piece of evidence. If Quine is right, the Popperian (and also Positivist) notion that theory falsification or confirmation, respectively, occurs one

statement at a time, is untenable. In this metaphor, a node, or statement, is connected to any number of others, which in turn foster a host of connections of their own. To add or dislodge any particular node from the net requires a rearrangement of the whole net. This rearrangement may be drastic or minor, depending on how central the node is. Hence, Quine's famous slogan about the un-revisability of any statement "come what may".

Second point: Quine denies that there are any fixed relationships or fundamental components in his network. Thus, languages and theories are fundamentally unstructured objects. This must be explained in the relation to the position of the Logical Positivists, to whom Quine was responding. So, following a brief Positivist interlude, I will return to the point above.

For the Positivists, there are two kinds of possible statements in a theory. First, there are synthetic sentences which can be confirmed by experiment or observation. "The sky is blue" is a synthetic statement. Additionally, it was claimed that the scientist could know what kind of evidence or experiment was required for a particular statements' verification. In the example that I gave, it is relatively obvious what this experiment would be: go outside on a clear day and look. There were thought to be unique defining experiments to confirm or falsify every observational statement. Returning to Quine's metaphor, these synthetic statements made up the periphery of the theoretical network. The positivist position on science was also a position on language. The meaning of a synthetic statement was its conditions of verification.

The second kind of statements are analytic statements. Analytic statement are statements that are true solely in virtue of the meaning of the words they contain, and the

logical structure of the sentence. Analytic statements, while (in earlier versions of Positivism) they themselves have no content, are the interior structure of a theoretical edifice or network. The usual example of an analytic statement is “Bachelors are unmarried men”. Analytic statements have the special job of determining where in the network new evidence fits into a theory. The logical structure acts as a constraint on evidence interpretation⁶. It also was supposed to provide the inductive mechanism necessary to determine conditions of verification or falsification. Analytic statements also have the special feature that they are immune to revision on the basis of observation or experiment.

The Positivist take on language meanings and scientific theories was atomistic. There was a primitive semantic relation — conditions of verification — that mapped the world to the meanings of individual statements. Moreover, the conditions of verification were unique to each statement. This was called reductionism, and I will refer to it as epistemological reductionism to distinguish it from the kind of ontological reductionism that concerns some philosophers of science currently⁷. Lastly, statements were confirmed or rejected on a case by case basis. They are independent of one another, and this independence and uniqueness of meaning was supposed to be guaranteed by the internal

6

Actually, it was initially supposed to completely define it. But as Carnap says, this was both “...too strong, and too weak.”. The shortfalls of this thesis, particularly with respect to verificationism and reductionism, are discussed in Carnap (1947) and (1951).

7

The idea, for example that temperature can be reduced to molecular motion, or the possibility that psychological concepts might be completely reduced to neuroscience with no loss of information.

analytic structure.

Now I return to Quine. Quine's famous paper "Two Dogmas of Empiricism" (1953) is an attack on both the analytic-synthetic distinction, and on the epistemological reductionism that depends on it. In this paper, Quine did not so much reject the a/s distinction as claim that there were no truly analytic statements. Therefore, no special epistemic weight was to be put on the logical relations among sentences. These relations, as well as any other, might be subject to revision given the appropriate evidence. What follows from this conclusion for the Positivist philosopher of science is that the structured nature of scientific theories is put into question. As Quine (1953: p. 37, in Martinich, 1990) says:

Science is like a field of force whose boundary conditions are experience. A conflict at the periphery occasions adjustments in the interior of the field. [...] But, the total field is so under-determined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the light of any single contrary experience. *No particular experiences are linked with any particular statements in the interior of the field, except indirectly through considerations of equilibrium affecting the whole.* [emphasis added]

So we have moved from an atomistic position about theory and language — where particular experiences are linked to particular statements — where the interior of the field is a logical structure, to a holistic position, where experience necessitates adjustments that are distributed across a network. The interior of the network is generated on a pragmatic basis. We choose certain structures because they are simple or elegant, and because they maximize consistency, but this structure or arrangement of nodes is optional and revisable. The argument for holism is essentially a disjunctive one. Quine sees holism and atomism as

exclusive alternatives. The latter depends on an untenable distinction between the analytic and the synthetic. We are left with holism as the only viable alternative.

Quine's thesis in "Two Dogmas" is also a linguistic thesis about meaning. His argument against analyticity starts with the idea that analyticity depends on substitutivity. What makes "bachelors are unmarried men" true is the fact that "unmarried man" may be replaced with "bachelor" without altering the truth value of the sentence. Substitutivity in turn depends on two things having the same definition. What Quine shows, or rather is usually thought to have shown, in the first four sections of the paper, is that there is no useful formulation of the concept of definition that is not circular. The only way to tell if two words mean the same or not is empirical — we *check* their conditions of verification. These conditions aren't knowable *a priori*; they are determined after the fact in the whole context of the language. The task of the language interpreter is likened to the scientist. The interpreter is a radical interpreter because they cannot epistemically count on any inner structure of language to limit the hypotheses about what words or sentences mean. Both the semantics and the syntax are ascribed holistically on the pragmatic basis of consistency and parsimony only. Ultimately, the radical interpreter fails in at least one sense because no definitive semantic theory can be ascribed to any given language. What ensures this failure is the under-determined nature of the theoretical enterprise.

So what import do Quine's conclusions have for semantics? To begin, consider Fodor's relentless commentary about the holy triad of systematicity, productivity and compositionality. One of the things that motivates his atomistic approach to language is that holists, according to Fodor, have a hard time dealing with these observed features of

language. For Quine, however, language structure (which is really what the holy trinity are all about) is not a fact that needs to be explained. Language structure is merely what is observed in behavior. With respect to productivity, for example, we cannot in fact observe that there are an infinite number of sentences. The syntactic structure of language is therefore part of the explanatory apparatus (a compositional, recursive syntax explaining, for example, why there are so many sentences), and is both indeterminate, and subject to modification. Let me illustrate. Suppose we start with a statement and you want to figure out what the individual words mean. Linguists probably won't care much about this kind of a scenario if they want to know what "dog" means, but they will certainly care if they want to know what "and" means. No one is going to figure out the meaning of a logical connective in isolation. Frege tells us that the way to go about doing this is to see how that word contributes to the meaning of the sentence. Now, if we limit ourselves to verbal behavior as evidence, Quine points out that there are, in principle, a large number of possible explanations that might satisfy the data, because the data underdetermine the theory. Each of these explanations will account for the data equally well, but each might also attribute a different role to any given word. While more information (as more sentences are examined) will certainly rule some of the possibilities out, we will never be able to examine all of the sentences in which "and", for example, occurs. Thus we will never, in principle, have a unique solution for what 'and' means, even though in practice there may well be only one credible hypothesis. So there is no "fact of the matter" about what 'and' means — no unique role in determining the sentence meaning — save what we ascribe to it. One semantic consequence is that this sort of an argument leads to

nihilism about meaning, because there is, in a real sense, no fact of the matter about word or sentence meanings.

1.22 Davidson:

Davidson also starts from the assumption that language users are radical interpreters. However, unlike Quine, Davidson assumes that radical translation actually occurs successfully. Davidson begins by identifying certain criteria of a successful semantic theory. First, it must explain how words contribute systematically to the meanings of sentences. Second, it must be extensional — no intensional concepts are to be employed in explaining sentence meaning. Davidson (p. 66 in Geirsson and Losonsky, 1996) complains that semantics in the style of Frege aren't up to snuff:

...the switch from reference to meaning leads to no useful account of how the meanings of sentences depend on the meanings of words (or other structural features) that compose them. Ask, for example, for the meaning of "Theaetetus flies". A Fregean answer might go something like this: given the meaning of "Theaetetus" as argument, the meaning of "flies" yields the meaning of "Theaetetus flies" as value. The vacuity of this answer is obvious.

When the Fregean switches from explaining word meanings to sentence meanings, it seems that he/she become trapped in an intensional idiom that is useless in explaining what "means" means. What Davidson also wants is a completely non-referential theory of meaning like Tarski's theory of truth. This would involve describing an equivalence function, where on the one side we have a structural description of a sentence, and on the other the sentence itself. An adequate theory of meaning would entail all such equivalences.

Davidson then proposes that, in semantic theory, we abandon the troublesome

notion of meaning altogether. Rather, a theory of meaning will be adequate if it entails all true equivalences. So, the job of the radical translator is to construct a Tarskian truth theory, or T-theory for a language. A successful T theory entails all equivalences of the following form: x is true iff p . Thus 'grass is green' is true if and only if grass is green.

Now for Davidson's holism. Davidson's semantic program is holistic in three ways. First, he exploits both holism and compositional syntax to solve a certain well known problem with his proposal. The problem is that extensional truth theories, i.e. Tarski's, only guarantee material equivalence. A successful T-theory entails that 'grass is green' is true iff grass is green, but it also entails sentences like 'Grass is green' is true iff 'snow is white'. Indeed we could replace the right side of the equivalence with any true sentence. Clearly more is desired from a theory of meaning. Where compositionality comes in is that the parts of the sentence, 'grass', 'green' and what have you, are used in many sentences. Thus the T-theory must entail the correct equivalences for every sentence with 'grass' in it, and every sentence with 'green' in it. The way to do this, says Davidson, is to exploit the compositional syntax of the language. Most commentators agree that this will actually work as a constraint, for example Evans and McDowell (1976: p. xv) remark:

The fact that each axiom of a truth theory has its impact on an infinite number of T-sentences does indeed have the consequence that it is difficult for counterfeit theories to pass the test.

The solution is holistic because this solution only acts as a constraint if it is applied over a large number of sentences⁸.

⁸

I propose to pass over the reasons — I am not concerned here whether or not Davidson's holism succeeds or fails, rather I want to describe it and what it entails for semantics.

Second, Davidsonian semantics are holistic because, like Quine, Davidson denies that individual words have meaning. Davidson is quite explicit about this (in Geirsson and Losonsky 1996: p. 67):

We decided a while back not to assume that parts of sentences have meanings except in the ontologically neutral sense of making a systematic contribution to the sentences in which they occur. Since postulating meanings has netted nothing, let us return to that insight. One direction in which it points is a certain holistic view of meaning. If sentences depend for their meaning on their structure, and we understand the meaning of each item in the structure only as an abstraction from the totality of sentences in which it features, then we can give the meaning of any sentence (or word) only by giving the meaning of every sentence (or word) in the language.

On this conception of meaning, word meanings are merely derivative of their context. I gave an example of this kind of concept earlier with respect to particles and indexicals and also with respect to logical connectives. The meanings of these kinds of words really are apparent only in the context of sentences. Davidson's holism is more radical than just this however. He is also proposing that all word meanings are inter-dependant. The meaning of "grass" depends on the totality of the sentences in which it appears. This is similarly the case for the meanings of the other words in "grass is green" too. This can be literally true, because ultimately the set of sentences that we must understand to comprehend even this simple little sentence embraces the totality of the language.

The third holistic feature of Davidson's theory results from his view that language interpreters are radical interpreters, and that one of the hallmarks of a good semantic theory is that it will work for a radical translator. In the translator's role as a T-theory

Interested readers can consult Evans and McDowell (1976), particularly the papers by Foster and Davidson. For a different take on the issue see Fodor and Lepore (1992), chapter 3.

builder, one of the main tasks is to maximize the consistency of the informer's beliefs. Clearly this is not achieved by picking and choosing which points of their verbal behaviour we will attend to. Rather, when the radical translator ascribes meaning to the utterances of an informer, the best tactic is to gather as many instances of, say, the utterance "grass", as possible. Thus the ascription of content only takes place within a holistic context, from which it is rapidly inferred that meaning is holistic. Note that Davidson feels that this is *necessarily* the case. As Fodor and Lepore point out (Fodor and Lepore, 1992: Ch. 3), the argument is metaphysical: For radical translation to work, content ascription must be holistic. In fact radical translation does happen. Therefore meaning is holistic.

1.23 Definitions:

Up to this point I have been using a number of metaphors to describe holism, particularly with respect to Quine. This was unavoidable in describing the position of Davidson and Quine, because they themselves never go beyond the metaphor. Now it's time to tighten things up. What I want to do first is to define holism in a way that captures the semantic theories of Davidson and Quine.

There are a number of explicit definitions of holism floating about in the literature already, however, they are generally unsatisfactory. The primary one is owed again to Fodor and Lepore. They start by describing the property of being anatomic. The idea is that an anatomic property is one such that if one thing has it, then others do also. The property of being a sibling for example, entails the existence of at least one other thing — a brother or a sister — that is also a sibling. Anatomism comes in degrees. In the sibling example, it is quite limited, in other cases, such as with numbers, there is a high degree of

anatomism. Very high degrees of anatomism mean holism.

There are a number of problems with this definition, so much so that it seems every single paper commenting on the *Holism: A Shopper's Guide* rejected it. The central problem seems to be that the unwholesome consequences of holism that Fodor and Lepore are concerned about don't seem to follow from the definition (e.g. Harrel, 1996), even where many commentators agree that the consequences Fodor and Lepore allude to do in fact follow from radical holism (Devitt, 1996: Introduction, Ch. 1). A second problem is that the definition overreaches itself. It tries to capture too many disparate theories in one net. As I am concerned only with Davidson and Quine for the moment, I want a narrower definition. Most of the remaining definitions in the literature concern themselves with the kind of holism embodied by conceptual role semantics. These will become of relevance later on.

Radical holism embodies two things. First, there is the idea that there is no fundamental semantic unit smaller than whole languages. Second, neither is there any word-by-word meaning determining relation. What something means is a function of the meanings of the entire body of statements in the language, with no one group of statements being the foundation.

Def'n: A language L (including theories) is holistic iff

- 1) The meaning of the set of all statements S in L, is logically prior to the meanings of words and individual expressions. Word or statement meanings are derived from relations with all of the other expressions in L.**
- 2) Primitive meaning, which is the meaning of all S in L, is derived from the relation between all L, and the whole world.**
- 3) It follows from 1) that the meaning of an expression in L can change if L changes (for example if expressions or statements are added or subtracted from L), because the relations among S may change.**

4) There is no *a priori* method of determining how the meaning of an expression S will change as L does.

A note about the definition: First, the definition does not guarantee that S will necessarily change with L. In fact, the change could be zero. No statement, however, is immune to potential revision in content or meaning. Compare this definition with the definition of atomism. In an atomistic semantics, simple meanings are prior to complex meanings, thus it is simple meanings that are primitive. Primitive meanings, in turn, are derived from relations between single words and single things, and are therefore independent from other meanings. Holism, so defined, is exactly the opposite. The primitive unit is the entire language. Simple meanings are derived from it, and therefore depend on the meanings of other statements, taken as a corporate body. Reference, such as it exists, is dependant upon the relations between the whole language, and the whole world.

1.24 Problems With Radical Holism:

This kind of holism faces several problems. First, there seem to be some words that *are* immune to revision as we add or remove statements from L. Even supposing that the meaning of a word like 'atom' might change for me as I take succeeding courses in physics (so that my meaning of this word could end up being very different from that of the lay person), it is not clear that any number of new statements, no matter what they are, will change the meaning of a word like 'or'. Indeed the idea is somewhat absurd. Changes in this kind of a word would soon make my language incomprehensible to everyone else, so in at least one important sense they are immune to revision. Imagine if the truth table for 'or' was replaced with the truth table for 'and'. Two people with different beliefs

about the meaning of 'or' would come to different conclusions about the meaning of sentences containing the connective. Worse still for the holist, we can identify these words in a principled and systematic way. The second problem is that language acquisition and comprehension becomes mysterious. If your L contains a different set of S, then what you mean by a particular statement might be quite different from what I mean. Indeed unless we share a large number of meanings, it may be that we don't share any meanings at all. Similarly, first language acquisition is problematic. An infant shares so few of my beliefs that it is unlikely that it could ever learn anything from me. And if it did, Quinian indeterminacy says that there is no guarantee that it would be the same language. Third, if no two people share the same beliefs then the prospects for a scientific psychology are dim (Fodor and Lepore, 1992: chapter 1). If there is no regularity about the beliefs and desires people have, then it is hard to see how laws of a predictive or explanatory nature could be formed about such things. With respect to some of these issues both Quine and Davidson are prepared to bite the bullet. Neither think the possibility of a scientific psychology is very good, for example. On other issues they are silent.

This concludes my second pass at holism. To this point I have provided an explanation and a definition of both radical atomism and radical holism. In both cases I have tried to capture the essentials of the position in such a way that both the explanations and the problems on offer flow from the definition in a reasonably obvious way. The purpose of the next section is to examine milder versions of atomism and/or holism. This sort of a position is captured by people advocating what Fodor calls semantic molecularism. It includes people calling themselves semantic localists as well as people

interested in conceptual role semantics. The middle position seeks to retain the explanatory insights of both extremes while avoiding the problems. At face value, this is an obvious and highly sensible move. It is questionable, however, if conceptual role semantics, for example, can in fact accomplish this end. Also, the middle positions have some unique problems of their own. The section on the middle positions is followed by a more extended discussion of compositionality, systematicity and productivity with respect to holism, and some concluding remarks for the chapter.

1.3 Mild Holism: Semantic Localism and Inferential Role Semantics

More or less, the central idea behind the more mildly holistic semantic theories (semantic localism) is that the meanings of a substantial number of words depend on *some* of their inferential properties. For example, part of the meaning of “bachelor” has to do with the fact that the following sentence is true of every satisfier of “bachelor”: “...is an unmarried man”. Contrast this view with Quine’s: the meaning of *all* words depend on *all* of their inferential properties. Semantic localists come in at least two main flavours. The first, single factor theories, are those who argue that tokens have a single meaning, depending on inferential role. The second, two-factor theories, argue that types in fact have two kinds of meanings. An example of the first kind of semantic localist is Devitt. An example of a two factor theory of semantics would be that of Block.

1.31 Devitt

The main idea of Devitt’s semantics is that some of the inferential properties of a token determine a designational chain, or d-chain. A d-chain in turn determines a reference. A designational chain is very similar in principle to Kripke’s causal-historical

chains (Kripke, 1980: 88-97). A d-chain borrows all the familiar Kripkean ideas about a word having an initial baptism via a direct perception of an object, and subsequent reference borrowing by other users of the word. What is special about a d-chain is that a word may have several of them, and that the d-chains are linked into a network. So, a d-chain is like Kripke's causal chains, but it is also like a Fregean mode of presentation.

Devitt (1996: p. 165) says:

[...] I claimed in effect that Frege was right about thinking that a name had a meaning or sense, but wrong in thinking that this meaning was descriptive (except for a few attributive names like Jack the Ripper). I identified the meaning of a name with the property of referring by a certain type of d-chain, each name by a different type. Similar remarks can be made about the meanings of pronouns, demonstratives and descriptions.

A d-chain, for example, may be referential, or attributive (Donnellan, 1966). Thus, using the familiar example from Donnellan, suppose that we have Joe who does not have a view on who murdered Smith, but believes that whoever murdered Smith is insane, on the basis of the bizarre means of the death. He says "Smith's murderer is insane". This is an example of attributive use. "Smith's murderer" refers to the object satisfying the description, if there is one. In fact, it was Jones who murdered Smith. In contrast, suppose, Deborah, who knows that Jones is on trial for the murder says "Smith's murderer is insane", meaning to refer to Jones. This is an example of referential use. The reference of "Smith's murderer" is to the specific object the speaker has in mind.

Devitt employs the machinery I have described to try to solve two problems: (1) en-rapport reference and (2) opaque contexts. I shall only discuss the latter. Recall that opaque contexts are problematic for atomists, so it will be a major point in favor of

Devitt's semantics if he can do this and yet avoid the pitfalls of radical holism. Let us return to Quine's opacity example, "Ralph said that Ortcutt is a spy". One solution to the transparent construal of the phrase is to export the singular term out of the opaque context. From this we get "Ortcutt is such that Ralph said him to be a spy". The latter phrase is transparent, and the inference from the first sentence, read transparently, to the second seems valid. However, the latter does not follow from the former, when read opaquely, as is now well known. Consider Devitt's counterexample. "Ralph believes that the shortest spy is a spy". Following exportation we get: "The shortest spy is such that Ralph believes him to be a spy". The two sentences are not equivalent. So, there has to be some sort of restriction on exportation that distinguishes between valid and non-valid uses. This is what Devitt provides. Exportation of singular terms outside of opaque contexts is valid where the d-chain is referential, otherwise it is not.

Can this work? The immediate reaction to this proposal might be that Devitt is in fact simply re-stating the problem. For one thing, one needs to know in a systematic way which d-chain applies in which situation. Devitt does not address this issue, but he perhaps might reply that this is a question of pragmatics rather than a question of semantics. Once again, however, my motive is not to address specific shortcomings in detail, rather it is to convey the essence of the middle positions in semantics with respect to holism and atomism.

I stated earlier that Devitt's view was that some of the inferential properties of a token determine its d-chain. Which ones? Importantly, Devitt does *not* want wholesale holism (Devitt, 1996: p. 10ff). Primarily Devitt is concerned with the inferential properties

of a token that determine reference — although presumably this relationship is mediated by way of the network of relevant d-chains. These inferential properties are related to the token analytically. Thus, it is in virtue of what the properties are, and what the token means, that we can distinguish between relevant inferential properties and every other inferential property.

1.32 Block

Ned Block also advocates a conceptual role approach to semantics (Block, 1994), although he doesn't have a specific semantic proposal. Rather, Block is arguing for a certain class of semantic theories. The main difference between the type of theory that Block wants and the kind of theory that Devitt wants is that Block advocates a two-factored approach to semantics. So on this proposal a token in fact has two kinds of meaning or content. The first kind of content factor is *wide*, meaning that the content consists in information that is reliably caused by some sort of world to mind relation. Wide content is neutral on the issue of holism and atomism, conspicuously, however, almost all of those that endorse an *information theoretic* approach to meaning are semantic atomists, perhaps necessarily so. The second kind of content is narrow content. This is the holistic part of the program. Narrow content is entirely "in the head"; Block defines it as the inferential role played by a particular token. Like Devitt, Block does not want full blown holism. Therefore his picture must include some sort of mechanism to systematically limit which inferential properties are relevant to a particular token. The mechanism he chooses is precisely the same as Devitt's. The relevant inferences are analytically related (Block, 1994: p. 93-93):

A crucial component of a sentence's conceptual role is a matter of how it participates in inductive and deductive reasoning. A word's conceptual role is a matter of its contribution to the role of sentences [...] Conceptual role is the *total causal role*, abstractly described [...] Conceptual role abstracts away from all causal relations, except those that mediate inferences, inductive or deductive, decision-making, and the like.

Fodor and Lepore note that two-factor theories inherit the specific problem of how to keep both factors in register. What guarantees that an object that reliably causes 'dog' tokens will have anything to do with inferences about 'vet', 'dog food', 'leash' and the like? Another problem, one raised by Block, is that it is not at all clear what the conditions of individuation are for conceptual roles. Radical holists don't care about this question, because they don't believe that single statements or words are independently contentful. Conversely, atomists have a number of solutions on offer, and they are, at least superficially, relatively intuitive and obvious. Nevertheless, Block suggests, and Fodor concedes, that these problems aren't knock down arguments against conceptual role semantics.

On the positive side, a two-factored theory does seem to be able to avoid some of the problems that confront atomism with respect to Putnam and Frege cases. One way of interpreting certain kinds of misrepresentational problems, including Putnam's and Frege's examples, is to interpret them as problems with the grain of representations (Stich and Warfield: p. 7). Frege cases show that there are instances in which a representation is not finely grained enough. Denotational semantics of the Russell/Mill tradition don't distinguish well between "morning star" and "evening star", where they should do so. Conversely, Putnam cases show that some meanings are *too* fine-grained. What a two-

factor theory permits is a way of understanding how it is that a semantic system can make both coarse and fine distinctions at the same time. Wide content does the former job, while narrow content does the latter⁹. Note that this also overcomes the problem of capturing both readings of *de dicto/de re* ambiguities described earlier.

1.33 A Definition:

If either Block or Devitt's position is tenable, then there are powerful reasons to adopt them. Because they postulate contents for individual concepts they retain the explanatory power of an atomistic approach to certain kinds of words. On the other hand, the additional resources they bring to a semantic theory are helpful in disentangling certain other kinds of problems that atomists find difficult to cope with. Having described two versions of what I call the middle position in semantics, as well as some of their specific strengths and shortcomings, what I would like to do next is to provide a definition for this position along the lines of what I have provided for atomism and radical holism. Following this, I will explore some general problems, specifically the possibility that the middle position inherits the shortcomings of full-blown holism.

In defining the middle position I'm going to adopt Devitt's less awkward terminology and refer to this cluster of theories as localist:

Def'n (3): The semantics of a language L are localist iff:

- 1) Meaning comes from word-word relations in some cases and from single word, single entity relations in some other cases.**
- 2) Some words have meaning in isolation, others do not.**

⁹

I am not suggesting that Putnam is arguing for a two-factored semantics. I am suggesting that an advantage of Block's approach is that it solves semantic problems that arise because of the grain of meanings.

- 3) In some cases, word meaning is logically prior to sentence meaning, in some cases the reverse is true.**
- 4) Some meanings in L can be altered by making alterations elsewhere in L. But it is not the case that all meanings will be altered, and there is usually a canonical method for determining the effect of the alterations.**

One thing that is apparent from the three definitions that I have provided is that they lie on a continuum. In an atomistic semantics, none of a symbol's inferential properties are meaning determining. In a localist semantics, some of the inferential properties are relevant. In a holistic semantics, all of the inferential properties are relevant. Where a particular theory lies on the continuum depends on how structured or organized one takes semantics to be. Here again I am regrettably forced to rely on metaphor to explain this idea. A holistic semantics can be likened to a drop of water. The position of a molecule inside the drop depends on the motion of every other molecule. An atomistic semantics is more like an ice crystal. The position of each and every molecule is rigidly specified by the arrangement of the crystal. Adding or removing molecules of ice makes no difference to the position of any other molecule. Localism is of course somewhere in between, but like the middle ground between a liquid and a crystal, it's actually very hard to describe what this middle state might be. Is localism tenable?

1.34 Problems with Semantic Localism:

There are two central complaints against semantic localism. The first is a general worry that semantic localists depend on an analytic/synthetic distinction in order to limit the scope of inferential roles. Quine, as I have described, provides arguments against the

analytic/synthetic distinction. If these arguments apply to the localist, then the principled distinction they need is lacking, and the position collapses into full-blown holism. The second problem, explicitly raised by Fodor and Lepore, is that the holy trinity of compositionality, systematicity and productivity may be as hard to explain for the localist as they are for the holist.

The first argument depends on a certain interpretation of Quine, specifically, of “Two Dogmas”. The interpretation is that Quine’s conclusions about science also apply to semantics. This interpretation is questionable. What Quine is uncontroversially thought to have shown is that analytic statements aren’t epistemically privileged. Analytic statements face the tribunal of experience like everything else. It is open for the semantic localist to claim that nothing epistemological turns on drawing the distinction between an analytic and a synthetic statement for semantic purposes. Fodor and Lepore reply:

On our view, Quine brought into question the possibility of *any* reconstruction of ‘meaning constitutive inferences’, in terms of ‘inferences accepted’. The crucial point is that there is, according to Quine, no way of distinguishing the inferences X accepts *because of what he believes*, from those that he accepts *because of what he means*; the failure of an a-prioricity criterion for what X means (hence a/s) is just a special case of this more general collapse. If Quine is right about this, it strongly suggests that there can be no a/s distinction that is accessible to an *inferential role semanticist* — none that the IR semanticist can appeal to to stop his localism from degenerating into holism.

The argument here is that the use of the a/s distinction, even in semantics, in fact has epistemic implications. Quine is arguing against analytic a-priori statements, which supposedly have two features. First, they cannot be empirically verified/rejected. Second, they cannot be given up without change in meaning. Fodor and Lepore’s point is that

analytic *beliefs* cannot be revised without change in meaning, and if they cannot be given up without a change in meaning then they cannot be empirically rejected either.

Suppose I convince someone on the basis of some linguistic evidence that what 'bachelor' means is actually a married man. They must now give up on the belief that bachelors are unmarried men. If 'bachelor' is believed to mean the same as 'unmarried man', then the statement that 'bachelors are unmarried men' is unrevisable, unless there is a change in meaning, and now we are back to Quine.

The point is well taken. Nevertheless it may still be too strong a reading of analyticity. Jane Heal, for example, points out that the constraint that IR semantics needs to impose is no more than the constraint that language users are rational thinkers. I think that she is right. Rey is more explicit in explaining this (Rey, 1993). Analytic beliefs (the kind that are meaning constitutive for an IR semantics) may simply be the result of automatic cognitive processes, about which we have no choice. However, there is no guarantee that these processes are themselves infallible, correspond to reality and so on. Thus we may have analytic beliefs in virtue of the nature of human cognition, but that these beliefs bear no special epistemic weight, because neither introspective access, nor an investigation of the cognitive apparatus responsible guarantees that they are true (at least not without further empirical investigation into how well these cognitive systems actually work). This is surely innocuous, indeed it would be quite remarkable, even ironic, to invoke Quine in contention of this naturalistic view. What I want to do next is assume for the moment that Fodor and Lepore are right on this point. Wouldn't some other principled distinction do the same job? Fodor scoffs at the possibilities ("Maybe its their length, or

whether they are tokened on Tuesdays...?”), but in fact there are some serious proposals. Quine’s point applies to deductive inferences, but it does not apply to, say, counterfactual robustness, or perhaps inductive inferences. There may even be a mix of less sweeping criteria, such as the two just mentioned, that could get the job done. Nevertheless I think Fodor is successful here. Any such criterion (or set thereof) would almost certainly have to include at least some deductive inferences at some point to retain any explanatory plausibility, which brings us right back to Quine.

The second major problem for IRS, compositionality, is more difficult. It has also received a more sparing treatment in the literature. The problem is that compositionality etc., seems to presuppose atomism. Peter Pagin (1997: p. 11) remarks:

On the face of it, the principle of semantic compositionality flatly contradicts the principle of semantic holism. The former implies that the meaning of a complex expression depends only on the meaning of its parts, while the latter implies that it also depends on the meanings of other expressions in the language.

Fodor and Lepore give an example of the consequences of trying to have both together. Suppose a person believes that brown cows are dangerous. If we take IRS to be true then part of the meaning of ‘brown cow’ is that brown cows are dangerous. Next, add the requirement of compositionality. Compositionality requires that the meaning of ‘brown cow’ be composed of the meaning of both words and an appropriate piece of syntax to connect them, in this case the modifier-head construction. The following inferences are now licensed: (1) brown entails dangerous and (2) cow entails dangerous (Fodor and Lepore, 1992: p. 177ff). Plainly this is absurd. Pagin agrees that this consequence necessarily follows, but he suggests that there is a fix for this. What the fix has to involve

is a way of blocking the inference to (1) and (2). How Pagin does this is in principle similar to Davidson's move with respect to the problem of extensionality, namely he makes use of the fact that words take part in multiple expressions, and then imposes a consistency requirement on uses of the words. Now, how IRS works is it maps the relevant global role¹⁰ of an expression onto the expression meaning. Inferences (1) and (2) follow from the premises because we have been talking about the role of 'brown' and 'cow' in isolation. However, Pagin points out that a token's global role involves mapping across multiple expressions, over which inferences must remain consistent. Unless all the expressions in which 'brown' is a part of licence an inference to 'dangerous', then none of them do. So (1) does not follow, and similarly the argument applies for 'cows' and (2). In fact, there are only two permissible inferences for 'brown cow', taken in isolation: that brown cows are brown, and that brown cows are cows. This is as it should be.

So far so good. Compositionality, however, is intricately linked with explanations of how systematicity and productivity work. And it looks like IRS might have to give up on these explanations. Pagin (1997: p. 29) summarizes the reasons:

Given that Holistic Inferential Role determines the meaning of an expression regardless of its earlier meaning, and given that a decision concerning a new sentence [added to a language] can change the meaning of that sentence (and many others with it), can compositionality [in the context of IRS] really explain how we can get new beliefs [about expressions] through our understanding of new sentences? [remarks added].

The worry is this. In an atomistic semantics, one of the ways that we get new sentences, and thereby also explain productivity, is by recombining the parts of existing expressions.

¹⁰Pagin defines this notion technically, but I lack the space to go into it in detail.

The meaning of the new sentence is stable because the meanings of the simple expressions that compose it are primitive and more or less invariant. However, in a localist semantics, the meanings of sentence parts aren't necessarily primitive, they can depend on the meanings of other expressions. So if we use compositional mechanisms to generate new sentences, the following consequence occurs. The new sentence might change the meanings of already existing expressions, and therefore also change the meanings of the parts of which the new expression is composed, because these parts are themselves expressions in the language. Thus it may be that IRS entails that the meanings of new sentences cannot be stable, because the new expressions, simply by existing, can constantly change their *own* meaning (effectively changing them to different statements), by acting on the meanings of existing expressions that contain the parts of the new expression. If this crazy result were true, then holism would of course be doomed.

Pagin contends that yes, one can rescue IRS from this problem. How this works is that the interpretations of new sentences, and therefore the interactive effect that they have on the meanings of other sentences, are also constrained by a consistency requirement. If I believe that 'dogs are small and black holes are dense' and my belief that 'dogs are small' changes from true to false, then the meaning of the compound sentence must also automatically change to false within my web of beliefs. Pagin is forthcoming about the entailments of this kind of repair. First it requires that the meanings of complex expressions cease being holistic — the above example is no different from, say, a Fregean treatment of sentence meaning. The meanings of the parts, e.g. individual words, however, can retain their holistic character, with one important exception. The meaning of any

words that performs syntactic functions, such as ‘if/then’ constructions, must remain invariant.

This is a pretty weak holism, though. In fact, it might cease being holistic altogether. Holism about individual words means that the contents of one word alter the contents of another. The limited holism of inferential role semantics requires that the way in which one belief impacts on another be constrained or structured in some way. So how do individual word meanings inferentially impact on one another cognitively? Presumably either via the same sorts of logical-syntactic operations that Pagin uses to constrain sentence meanings¹¹, or maybe also via some other mechanism, perhaps some kind of associationism. The former won’t help the cause of the IR semanticist any. Suppose I learn that dogs are mammals, and mammals have kidneys. This changes my beliefs about the meaning of ‘dog’ in the following way: I now assent to the belief that dogs also have kidneys. But this is a trivial sort of holism. The atomist who endorses direct reference can reply as follows: The new inference does not change the extension of ‘dog’, therefore it has no impact on the meaning. It tells us only that the extension of ‘dog’ is contained in the extension of ‘mammal’. Furthermore, the atomist will say, recall that most IR semantics *themselves* require that the meaning-relevant inferential properties specify the reference of an expression.

The second possibility, that perhaps some form of association is involved, is

11

And note it does no good to complain that this presupposes LOT. Pagin’s repairs to IRS require that we retain the logical syntactic apparatus for the constraints he introduces on sentence meanings.

reminiscent of radical holism. But it is difficult to see how associationistic mechanisms could limit the interaction of word contents in any principled way¹². Also, is it really plausible that our word contents impact on one another via some sort of associationism, but that the contents of our complex expressions don't? I'm not quite sure of what to make of the plausibility of the latter proposal. What would such a cognitive system look like? It strikes me that an account like Pagin's needs to flesh this out. As it stands, the repairs are pretty leaky, and it looks like the IR semanticist might indeed have to give up on compositional explanations of systematicity and productivity. Even if Fodor and Lepore's arguments about the a/s distinction fail, and I think that they do, this would be a pretty persuasive reason to be suspicious of semantic localism.

While I'm on the subject of syntax, it's worthwhile to mention again that even the semantic atomist might have to endorse holism about syntax, and consequently, be a holist about words that perform syntactic functions. Words like 'and' and 'or' are meaningless outside the context of a complex expression. Their only meaning is the role that they play in establishing the meaning of a complex expression. Once again, though, this is a trivial kind of holism. Words like 'bachelor' change their inferential role depending on the inferential properties they possess. What makes holism useful and interesting is that these properties can change, depending on the other meanings of other words. The inferential role of a sentential connective like 'and' however, *never* changes, regardless of our other

¹²

It is of course an open possibility that some other mechanism can bridge the gap here, but I have no idea what it might be, and so far there are no proposals. As I have already remarked, Fodor makes similar complaints.

beliefs. It is entirely possible, even if entirely speculative, that the kinds of beliefs that would be required to change the meaning of 'and', 'or', and the like, would be entirely incomprehensible to us as human beings. Some words just don't face the tribunal of our beliefs as a corporate body, nor does it appear, semantically speaking, that any meaning is revisable, come what may.

1.4 Concluding Remarks

The situation as it stands is pretty grim. We have two varieties of semantic accounts, atomism and holism, that provide partial explanations of certain semantic phenomenon. If one chooses among these semantic phenomenon carefully, emphasizing compositionality etc., an atomistic picture of semantics follows. If other phenomena are selected, eg. indexicals and so on, then a holistic picture might follow. This leaves little to choose between them. What would be nice is to occupy a localist position that recruits the explanatory resources of both. Such a position takes the view that atomism and holism are a smooth continuum, and that we can perch on this continuum wherever we like. One reading of the alternatives is that they are really questions about how much internal structure there is to a semantic system. For an atomist there would be quite a lot, for the holist, there would be very little, and semantic localism would again occupy a position somewhere in the middle. However, it looks like the middle position may be untenable. As described by Pagin, the problems of using compositionality to explain productivity in a localist semantics suggest that either it is driven to wholesale holism, or it is driven to wholesale atomism. So we might be stuck with either a lot of structure or almost none. Is this a consequence of compositionality? Fodor seems to think it is. This suggests that

atomism and holism may not be a continuum, like the continuum from hot to cold, but in fact might be true dichotomous opposites, like 'alive' and 'dead'. What is unfortunate about the notion that holism and atomism are true opposites is that it leaves very little hope of a theoretical reconciliation. Thus, if one is skeptical about the prospects of atomism or holism solving the fundamental problems that face them, and I am, then a semanticist gets by with their choice of either one or the other only by ignoring the bits of semantic ontology that inconvenience them. For example, Fodor's choice of examples. Its always 'cow', and never 'that'. This does not strike me as a good way to go about doing things.

Chapter 2: Classical Approaches to Cognition and Connectionism

In semantics, I described two general approaches to assigning meaning, a holistic approach, and an atomistic approach. Similarly, there are two general answers for what kind of model should be used to describe thinking. The first, from classical AI work, is that the mind is a physical symbol system. The second is that the mind is a neural network, and goes about its business as a parallel distributed processor. With respect to cognition, the discussion on semantics dealt with the content of words and mental representations. In contrast, the discussion in this chapter deals with how those representations are structured and processed. The first part of this chapter describes the two approaches, and why they are seen by some as distinct alternatives to modelling cognition. It also describes Marr's and Pylyshyn's (and others) notion of the cognitive level. The idea of cognitive levels figures prominently in the second part of the chapter. The second part of the chapter discusses Fodor and Pylyshyn's dilemma for the connectionist. Their argument suggests that the connectionist account is either inadequate or irrelevant. It may also imply that the two approaches are incompatible. Part three discusses responses to Fodor and Pylyshyn's dilemma, that is, what the options are for the connectionist. Some of these options also imply that connectionism and physical symbol systems are incompatible alternatives, and that physical symbol systems are in turn either irrelevant or inadequate for cognitive science. My conclusion to this discussion is that i) while importantly distinct, the two models are not necessarily incompatible. Their distinctness is quantitative, rather than qualitative and ii) While both are inadequate as theories at present, there is no valid

scientific reason to abandon either as approaches to modelling cognition.

2.1 Characterization of the Two Cognitive Architectures:

2.1.1 Classical Machines

Physical symbol system models generally make three claims (Pylyshyn, 1983).

First, the bearers of mental content — mental representations — are physical symbols.

This is important because mental representations are intentional; they carry information about things in the world. Beliefs and desires are also intentional, and there are arguments that any psychological account of cognition will necessarily resort to explanations in terms of beliefs and desires (Davidson, 1980; Fodor, 1987). Thus, the proposition that mental representations are tokened as symbols gives us an explanation of how beliefs and desires might be implemented in a physical system. To have a belief or desire consists in having an appropriate mental representation, and to be in a certain functional state. As the metaphor goes, believing that foxes are red means having a ‘foxes are red’ token in a “belief box”.

Second claim: Mental representations are tokened as special kinds of symbols.

Specifically, they are discrete structures with a syntactic organization. Thus, mental representations can be likened to sentences in a language. Importantly, this is more than a metaphor. Several formulations of the Physical symbol system model of cognition propose that thinking takes place in a language of thought (the LOT hypothesis, see Fodor, 1980 and Fodor, 1975). This second claim permits an explanation of systematicity.

Systematicity is the idea that if you understand the meaning of the constituents of a sentence (or a thought), and the manner in which they are lawfully arranged, then you automatically understand and are able to produce all the other well formed arrangements

of the constituents. If thoughts are sentence like, and the sentences have parts which can be arranged according to recursive syntactic rules, then we have an explanation for how thinking can be productive and systematic. This facet of the classical model is seen as especially important by Fodor, Pylyshyn, and McLaughlin (1988, 1990).

Third, cognition occurs by way of formal operations on these syntactically structured representations. Thinking is like solving a problem in predicate logic. It occurs in distinct steps, with distinct procedures, and need pay no attention to the content of the symbols. Furthermore, the symbols and the rules for conducting formal operations on symbols are explicitly represented — they are like a computer program. This third claim is required because of the necessity of explaining process. The computer program/formal operation metaphor gives us a convenient analog for cognitive processes.

Physical symbol system theories entail two things. First, they require that symbols and their contents be discrete and identifiable. A particular piece of content is uniquely attached to a certain kind of symbol. These kinds of explanations therefore entail semantic atomism with respect to the LOT. Second, Physical symbol system entail von Neumann-like architectures, with input and output devices, memory structures, and a central processor. This follows from the requirement that both the symbols and the rules for operating on symbol structures be themselves explicitly represented in the system, in a form analogous to data and programs. A Von Neumann architecture provides the resources for doing so. Input devices token the symbols, memory stores them, and the central processor performs operations on them according to an instruction set, which is also contained in the memory.

Physical symbol systems are extremely good at explaining and implementing a variety of cognitive phenomena. For example, they are good at explaining deductive reasoning. One example of such a system is Rips' ANDS system (1983). Rips programmed ANDS with a set of rules to solve a series of 32 problems in symbolic logic, and made proposals to extend the system to probabilistic reasoning. The system was able to solve the problems with the same degree of accuracy as a group of subjects that were untrained in symbolic logic. This suggests that deductive reasoning in humans is based on rules sensitive to the syntax of the logic problem, even if the rules are not precisely those found in logic texts. Second, for fans of belief/desire psychology, the physical symbol system implementation seems to be a natural home, given that there are (relatively) easy ways of cashing out the intentionality of mental representations. Several examples are to be found in Stich and Warfield (1994). Most of these explanations exploit both discrete symbols and discrete functional states, which are generally considered to be properties of a physical symbol system, indeed Smolensky (1988), among others, stipulates that the kinds of connectionist systems of interest (for my purposes) do not have such properties, by definition. Lastly, Physical symbol systems are good at explaining systematicity and productivity, for the very same reasons that an atomistic account of semantics is good at explaining these things: Physical symbol systems postulate complex symbol structures — sentence analogs in the LOT — with constituent structure and a recursive syntax. Thus, because Physical symbol systems have symbol structures with discrete, atomic components, and explicit rules for how the combination of these parts contributes to symbol content, they are systematic for the same reasons that an atomistic semantics with

a recursive, combinatorial syntax is. I described these reasons in detail in chapter I.

The classical model, embodied by the Physical symbol system, is however, quite poor at explaining a variety of cognitive phenomena (Rumelhart, McClelland, and Hinton, 1986). Classical implementations have a hard time with pattern recognition and deformed input, to name just two items. Physical symbol systems are also brittle. If the physical system is damaged, it generally ceases to function altogether. Human “hardware”, in contrast, can sustain damage with only a partial loss of function. Similarly, a Physical symbol system that is given inadequate memory or other computational resources to perform the tasks on hand tends to crash, or just not work. Human cognition exhibits graceful degradation of performance under these conditions. That is, cognitive processes will continue to function, but with reduced efficiency under conditions of either damage, or cognitive overload. Lastly, Physical symbol systems are rigid and inflexible because the rules which govern formal cognitive operations are exceptionless. Human cognition, in contrast, seems fluid and full of exceptions.

2.12 Neural Networks:

The second general approach to modeling cognition is on a neural network. In general, all neural networks possess three kinds of things. First, there are layers of nodes. Nodes are simple processing elements that take the sum of their inputs, perform a mathematical operation on those inputs and then transmit the output of that operation to other nodes that are downstream to them in the network. Nodes are frequently arranged in three layers, an input layer, a ‘hidden’ layer, and an output layer. The reason for three layers is that two layered networks are not sufficiently complex to perform interesting

cognitive tasks, and more than three layers are generally computationally unnecessary for most purposes.

The second element of a neural network are the connections between the nodes. These connections have a certain strength or weight that is applied to any throughput. Thus, if the connection is transmitting a node output of value 2, and has a weight of 0.5, the throughput — what the target node will ‘see’ from that connection — will have a value of 1. Nodes and connections can be arranged so that networks are modular, that is, so that there can be a variety of sub-networks that are relatively isolated from the rest of the net.

The third element of many neural networks is some form of training algorithm which changes the weights of the connections (and sometimes makes more extensive modifications of the neural network also). One of the most common learning or training algorithms is the generalized delta rule, implemented by a technique called back propagation. This technique compares the output values of the network with the desired output, and then alters the weights of the connections of the network, based on the difference between the actual and the desired output. There are also networks that don’t require training algorithms like back propagation, but instead make use of concepts and techniques such as Hebb synapses and competitive learning techniques to modify the network.

It is important to state from the outset that there are a variety of different styles of connectionist computing, and a wide variety of neural networks. I am interested, however, in a particular class of neural networks described in Smolensky (1988). So in all

subsequent discussions of neural networks, I am referring explicitly to this kind of a neural network. In “The Proper Treatment of Connectionism” Smolensky stipulates that a connectionist neural network possess three characteristics. First, the representations are distributed. The content of a distributed representation is carried by the pattern of activation of the entire network. One consequence of this is that the content of a particular representation can never be uniquely associated with any particular network unit, or even set of units. A particular unit, or node (or sometimes neuron), for example, may be active as *part* of the representation of ‘fish’ one moment, and *part* of the representation of ‘shark’ at another. There is therefore superposition of semantic content (van Gelder, 1990). Furthermore, for more complex representations, such as a representation of ‘a cup of hot coffee’, Smolensky is particular about the following point: No particular node or set of nodes explicitly represent the parts of the representation in a semantically systematic way. If we want to know what part of the network is representing ‘coffee’, the best approximation is to take the pattern of activity for ‘a cup of hot liquid’ and subtract it from the pattern that stands for ‘a cup of hot coffee’. The residual pattern approximates the pattern that means ‘coffee’ in the complex expression, and there is no guarantee that this pattern will be identical to that produced by coffee alone. So what do individual nodes or connections represent? What they are said to represent are the micro-features of the representation. Although Smolensky gives examples of micro features in a variety of places (see Smolensky 1988 or 1990), its not altogether clear what counts as a micro feature, and what does not, so it is difficult to know explicitly what kind of an entity a micro feature is, or how these are assembled into larger groupings that mean whole things.

The second feature of neural networks that interests Smolensky concerns rules. In a classical architecture rules play an explicit causal role and are also explicitly represented. Furthermore, both the rules and the symbols are best described by the mathematics of discrete structures; set theory, for example. Smolensky's interest lies in networks where the rules are implicit. They aren't stored explicitly, instead the rules are contained in the sum of the weights and processing elements of the network. Moreover the rules don't seem to play the same sort of causal role that they do in a classical architecture, but rather describe the evolution of the activation vector through a phase space, which is an epiphenomenon of node-level causal forces. Lastly, the mathematics that describe the evolution of the activation vector aren't discrete mathematics, but rather are the mathematics of dynamic systems — sets of differential equations¹³.

Lastly, Smolensky says that the kind of neural network he describes operates sub-symbolically. Thus he suggests it is closer to Marr's implementation level than standard symbolic architectures, which generally operate at the algorithmic level¹⁴. Two claims follow. First, Smolensky suggests that network architectures will more smoothly reduce to neuronal architectures than classical architectures because they are more obviously like neurons. Second, Smolensky claims that symbolic processes are emergent properties of

¹³

Some authors, such as Bishop, use matrix notation to analyze networks, but as linear algebra is also a branch of continuous mathematics this is not a worry. Smolensky is quite adamant about this mathematical point and has gone to considerable effort to show that tensor-product analysis of networks have wide applicability to a variety of networks (Smolensky, 1990), including those characterized by Bishop.

¹⁴

In contrast to McClelland, Rumelhart and Hinton, who are of the opinion that networks are algorithmic level descriptions (1986, p. 124)

sub-symbolic architectures. Presumably a network that is big enough and complex enough will exhibit behavior that approximates the behavior of classical architectures. How, precisely, this all comes about is also a bit of a mystery, but more on this later.

I am interested in this class of neural networks for two reasons. First, Smolensky argues that his definition of connectionism is preferable for a research program, because many of the other connectionist systems may very well turn out to be merely implementing classical architectures; a point which will be important shortly. Smolensky is in effect claiming that new and interesting capabilities of connectionist systems are only likely to be found with this kind of a network.

The second reason for using the class of networks identified by Smolensky is that many, if not a majority, of the philosophical debates about connectionism refer explicitly to this style of connectionist network (see Ramsay, Stich and Rumelhart, 1989 and Horgan and Tienson, 1989 for many examples).

Before proceeding into the debate between Fodor, Pylyshyn and Smolensky, one more piece of background is necessary, and that is a discussion about the notion of cognitive levels. This is required because the notion of a cognitive level is pervasive in both subtle and obvious ways in most of the literature surrounding the classical vs. connectionist debate.

2.13 Cognitive Levels

The notion of a cognitive level originates with the classical approach to cognitive science, initially with 1) David Marr (1982) in particular, but it has also been used/described by 2) Pylyshyn (1984) and 3) Horgan and Tienson (1989). The notion of

levels are consequently almost always described in the vocabulary of the classical approach, although the latter authors do attempt to generalize the framework, so that it encompasses connectionist architectures as well. Whether the vocabulary of cognitive levels is appropriate, adequate and relevant is part of the dispute between classicists and connectionists. For purposes of this discussion, I use Marr's terminology, because that is the usual terminology in the connectionist literature. However, the interpretation of the notion of levels that I personally favor is that of Pylyshyn (1984), because Pylyshyn is specific about what his ontological commitments are. Pylyshyn is a realist about the algorithmic level. So am I. That said, all three of the numbered approaches to the idea of levels do share large similarities. First, there are three levels¹⁵. The highest, least reductive level is the cognitive or semantic level. This is the level at which environmental inputs and behavioral outputs are relevant; cognitive tasks, such as navigation about city streets, are described at this level, and their requirements within a functioning cognitive architecture are described. For instance, navigation requires perceptual mechanisms to identify terrain features, some sort of memory, and so on. This level is always considered to be semantically interpretable — and it had better be, because semantics is largely (but not exclusively) a *cognitive level* description about what words mean!

The middle cognitive level is the algorithmic level. The algorithmic level consists in

15

Some descriptions of cognitive levels refer to more than three (e.g. Newell, 1982). However, Newell's knowledge level is clearly analogous to the cognitive level, his program and register transfer level are analogues of the algorithmic level, while his circuit/device levels are clearly implementational.

transitions from one functional state to another. These transitions are mechanistic, formal, abstract and mathematically governed. Algorithms are mechanistic because they contain symbol structures that are causally effective in producing changes of state in the system. Part of the connectionist debate is about what these symbols are like (distributed or local, symbolic or subsymbolic etc.). The algorithmic level is formal. What constitutes a formal system is that what is causally relevant to the system is the syntax (sometimes described as the *shape*) of the symbols. It is not necessary for this level to be semantically interpretable — it *can* be semantically evaluable¹⁶, but the semantics of a symbol are not what is causally efficacious. Thus the algorithmic level abstracts away from specific content. An analogy of this are procedures for solving problems in logic. Once the problem is ensymbolized, the content is not relevant to the mechanics of solving the problem. Lastly, the causal properties of a symbol are due to the mathematical nature of the symbol structure, including its syntax. If a symbol is a discrete structure c.f. the classical symbol, it may have certain properties, such as a recursive, combinatorial syntax. If however, the symbol is a *tensor product* (see Smolensky, 1990), it may exhibit different properties. A tensor product for example, does not have the strict combinatoric elements of a classical symbol, save in the limiting case where all the input elements of the network are linearly independent. Linear independence requires that the appearance of a particular microfeature in the input should not alter the probability of that another microfeature is represented. This constraint would be violated, for example, with microfeatures of the

¹⁶Nor does it matter for my purposes if this is the case or not.

letters of the alphabet¹⁷.

The lowest, most reductionist cognitive level is the *implementation* level. This level is the hardware of the system. In the case of computers it consists in the circuitry inscribed on the microchips, the hard drive and so on. In the case of human beings, the implementational level is the brain.

One issue about levels that must be addressed immediately is the issue about whether these levels are merely descriptions, or whether they constitute an ontological claim about the structures of cognition. This is important because the validity of certain kinds of positions in debates about connectionist networks depend on whether levels are descriptions or whether they are real entities. For example, some people argue that a connectionist system possesses syntactically structured representations if it can be *described* as having them. These sorts of instrumentalistic answers are forbidden if a case can be made that cognitive levels are or ought to be treated as real entities.

The issue about the status of levels as objects must be addressed one level at a time. In the case of the implementational level the answer is clear and undisputed by any of the participants of the debates I'm interested in: implementations (i.e brains and computers) are real objects, furthermore, provided that we are materialists, they are physical objects too.

¹⁷

The appearance of a diagonal slash, for instance, will very much change the possibility of other microfeatures being represented. We get this feature in 'A', 'V', 'W', 'X', and 'Z' (and maybe 'Y' depending on the font). So the appearance of this microfeature eliminates the probability of, say, a curved line from being in the same letter representation as the slash feature, conversely there is a higher than normal probability of another diagonal slash.

The top level, the cognitive level, however, is considered by some philosophers to be a description only (Dennett, Clark) — a useful shorthand for describing implementation level processes. Dennett calls this description of the mind the *intentional* stance, and contrasts it with other descriptions, such as the design stance. Eliminative materialists, such as the Churchlands, go further and suggest that cognitive level descriptions are in fact not only descriptions, but false descriptions, like alchemical descriptions of chemical reactions, and that cognitive level terminology like ‘belief’ and ‘desire’ should be eliminated forthwith from the vocabulary of cognitive science. Lastly, there are a significant number of philosophers who contend that the cognitive level is real (Fodor, Pylyshyn).

Most important, however, is whether or not the algorithmic level is real, or simply a description, because many of the arguments further along use realism about the algorithmic level as a premise. Smolensky, and indeed many connectionists, contend that the algorithmic level is only a description, while Fodor, Pylyshyn, Stich and many others take the view that it is real. I must side emphatically with the realists, and against Smolensky in particular, because I believe, for one thing, that the only coherent reading of Smolensky’s position entails that he ought to be a realist about the algorithmic level also. The reasons follow. First, Smolensky is already ontologically committed to what he calls the sub-symbolic level of description. This follows immediately from his claim that sub-symbols have causal properties, in contrast to classical symbols, which according to him, do not have causal properties. And the thing described can only have causal powers if it exists. Smolensky also concedes that cognitive level phenomena, such as some semantic

properties of natural languages, require an explanation, and more importantly, stand in need of some mechanism of implementation in the sub-symbolic paradigm. And if you believe that something needs a mechanism of implementation then you are *de facto* committed to its existence. It follows that Smolensky is committed to realism about the cognitive level. Contrast Smolensky's position with Dennett, who does not feel that any reductive explanation or implementational strategy for the intentional stance is required, or perhaps even possible.

Smolensky makes two further claims, and this is where he must commit himself to realism about the algorithmic level. First claim: sub-symbols aren't semantically evaluable, but a tensor product algorithm (all connectionist algorithms, by stipulation) can be defined over them. Second claim: sub-symbols are constitutive elements of something like classical symbols, which *are* semantically evaluable, but over which a tensor product algorithm cannot be precisely defined (Smolensky, 1990). Thus he proposes what he calls a 'split-level' architecture¹⁸, in which a classical symbolic structure cannot be properly said to constitute a component of the algorithmic description. This second claim is made, in part, because Smolensky concedes that the sub-symbolic level, by itself, is inadequate to explain/implement all of the kinds of semantic properties that preoccupy Fodor et al., which is why he proposes a reduced role for the symbolic paradigm, rather than outright replacement or elimination. His solution to the implementational inadequacies of connectionism with respect to certain semantic properties is to propose that *descriptions*

¹⁸Indeed, Smolensky uses the architectural features of his model in order to argue that he is not merely implementing a classical system.

of the classical symbolic level can be built up out of *descriptions* at the sub-symbolic level, at least in an approximate way (Smolensky 1988). The problem is this: What Smolensky is in effect proposing is that a classical algorithmic description, i.e. the cognitive level at which classical symbols reside, can be an emergent property of the sub-symbolic level, which recall must be real, because it has causal properties. And Smolensky (1995, p. 185) is very explicit about the supervenience relation. He says:

Constituency is one illustration of a central component of the general PTC approach to connectionism: the relation hypothesized between connectionist models based on continuous mathematics and classical models based on discrete symbolic computation. That relationship might be called the *cognitive correspondence principle*: when powerful connectionist systems are appropriately analyzed at higher levels, elements of symbolic computation appear as emergent properties.

Note that in the preceding quote Smolensky makes reference to the correspondence principle of statistical thermodynamics, which is used to explain how the higher level thermodynamic properties emerge from the quantum mechanical properties of molecules; this will provide a useful analogy momentarily. Now, not only does Smolensky suggest that symbolic descriptions emerge from sub-symbolic states, but worse, he suggests that certain natural language properties are *implemented* by algorithms using these symbolic level descriptions (Smolensky does not dismiss symbolic architectures for all purposes, they merely have a reduced role). In the first case, it is confused to suggest that descriptions supervene on properties (Kim, 1993). In the case of the correspondence principle proper, it is not a *description* of thermodynamic properties that emerges from quantum mechanical properties, its the actual properties of real aggregations of molecules that emerge. In the second case, its just not the case that descriptions implement anything.

Descriptions of thermodynamic properties don't implement Boyle's Law — actual collections of molecules in a gaseous state do. In order for one thing to implement another, it has to have causal properties — at the very least those required for the implementation. As anything with causal properties is normally thought to be something that really exists, Smolensky is ontologically committed to the existence of the algorithmic level. Therefore, with respect to Smolensky, all future reference to cognitive levels presumes that there is an ontological commitment. What is an open question is whether or not the correct algorithmic description is the classical description, or a connectionist, presumably tensor product description.

The argument against Smolensky can be turned into an argument against instrumentalism about levels in general. Both instrumentalists (e.g. Dennett, Clark) and anti-reductionist realists (e.g. Fodor and Pylyshyn) appeal to algorithmic level explanations. Thus neither believes that a reduction of algorithms to hardware is useful or possible. The problem for the instrumentalist is that some observed patterns of behaviour are *caused* by algorithmic level entities. Productivity and systematicity in natural language speech, for example, are the result of features of the algorithmic level, specifically, they result from the characteristics of the symbols structures and rules of the algorithm. Now if reduction of the algorithm to an implementation is not possible, as both anti-reductionist positions must concede, then it cannot be said that symbols and rules exist at the level of implementation. This becomes problematic for the instrumentalist because, again, descriptions don't have causal properties, only objects have causal properties. This means that the only plausible anti-reductionist story that can be told appeals to real emergent

properties; i.e. the semantic level emerges from the algorithmic level which in turn emerges from the implementation level. The only other move available is to concede that the reduction is possible, in which case nothing motivates instrumentalism about cognitive levels in the first place.

So why not adopt a reductionist position like eliminative materialism? For a long time after all, cognitive scientists have emphasized the top two cognitive levels and have tended to ignore the implementation level. The justification for this pattern of behavior has been the issue of multiple realizability (see Horgan, 1997 and others). Multiple realizability is the conjecture that many different hardware systems are capable of implementing the same algorithms. Thus the program that runs the word processor that I am using is, at the time of writing, being implemented on an Intel chip by the Windows 95 operating system. However, the same program could be run on a physically different machine — perhaps one with Motorola chips — and a different operating system, such as UNIX. Similarly a Martian might think the same thoughts as I, and even think them in the same algorithmic way, yet have a completely different brain structure. There are two consequences of multiple realizability. The first is that it preserves the autonomy of psychology and probably also linguistics as disciplines independent from neuroscience, which tends to focus on the implementational level. This follows because it isn't clear how findings from neuroscience could provide meaningful reductive explanations in psychology or linguistics. The second consequence, which is closely tied to the first, is that the multiple realizability argument has fostered an attitude amongst cognitive psychologists and linguists that neuroscience may be safely ignored, because the relevant explanations are to be had at the

algorithmic level.

2.2 The Fodor/Pylyshyn Dilemma:

Perhaps the main argument against connectionism is a question of cognitive levels, representations, and rules. The argument is owed to Fodor and Pylyshyn (1988), and has proved to be a durable feature of the debate between connectionists and advocates of GOFAI. Fodor and Pylyshyn pose the following dilemma for the connectionist: either the connectionist is providing an algorithmic explanation of cognition which is empirically inadequate, or the connectionist is providing an explanation at the level of implementation, in which case connectionist explanations cannot generate any novel explanatory features in algorithmic or cognitive level explanations, which are the relevant levels of explanation with respect to cognitive science. The first horn of the dilemma is entailed by two observations. The first is that cognition seems to have certain features such as systematicity and productivity, which we would like our theories to account for. The second is that connectionism, taken as an algorithmic-level account, does not seem to have the resources to explain how these cognitive features work, and this lack of appropriate resources seems to be a principled privation. In order for a system to explain systematicity and so forth, that system would have to possess appropriately structured representations (representations with a compositional semantics) and explicit syntactical rules that are causally efficacious in operating on the representations. Connectionist systems don't have these things, and therefore cannot — in principle — account for a certain class of relevant cognitive features. The second horn of the dilemma follows from the multiple realizability argument. If connectionist explanations are merely about an alternative method of

implementation, then those explanations simply don't add anything to cognitive science, because the algorithmic and cognitive level explanations of a connectionist system are no different than those of classical explanations, and therefore don't represent an alternative to the classical model.

This dilemma furthermore implies that connectionism and physical symbol systems are not compatible alternatives. Because the physical symbol system is a formal system, then necessarily physical symbol systems have syntactically structured representations (and *only* this kind of symbol — nothing else can be processed). Connectionism, by definition, does not have this kind of symbol and cannot (according to Fodor and Pylyshyn¹⁹) process syntactically structured symbols. Similarly, connectionist symbols are no good for classical architectures. Vectors whose transition matrices are system of differential equations, for example the activation vectors in connectionist networks, are necessarily defined over $\mathbb{R}^n$, where n is the dimension of the vector space. The objects of classical computational mathematics however, are the well formed formulae for a language. Its hard to see how an operation defined over a discrete object such as a language would also work on an activation vector. It follows that its *prima facie* unlikely that either system can process (perform operations on) the kind of information stored in the other, because their respective symbol types are useless to the other type of machine.

¹⁹Much more on this later. More or less, Fodor and Pylyshyn turn out to be correct.

2.3 Responses to Fodor and Pylyshyn:

2.3.1 A Geography of Replies:

Fodor and Pylyshyn's argument is generally attacked in one of the following ways. The first overall strategy is to go after the notion of representation that is being employed. These arguments come in three varieties. First, some argue that connectionist systems aren't representational at all (i.e. Smolensky is wrong about the proper treatment of connectionism, and therefore Fodor and Pylyshyn miss the mark) and mental representations aren't a necessary feature of human cognition²⁰. So the connectionist has nothing to worry about. The second position is that connectionist systems are representational, and while these representations cannot have constituent structure of the sort that Fodor and Pylyshyn are fond of, this shouldn't worry us, because, it is claimed, human cognition just isn't systematic in the way that Fodor and Pylyshyn suggest²¹. As the purpose of postulating constituent structure is simply to support productivity and systematicity, syntactically structured representations are therefore just not required. Alternatively, Smolensky (1987 and 1990) argues that the required features that Fodor and Pylyshyn insist on are emergent properties of connectionist networks. Similarly, a number of researchers, including Smolensky (see Hinton, 1991) are trying to make connectionist networks where the representations can be said to have constituent structure in virtue of being able to represent information in graphs and tree structures. Third, it is

²⁰

See Brooks, 1991; Beer, 1990 and 1992, Churchland and Sejnowski, 1992; and Waskan and Bechtel, 1997 for examples.

²¹The original PDP volumes adopt this position in places, as do Waskan and Bechtel, 1994

argued that neural networks really do have the appropriate kind of representational features, and that Fodor and Pylyshyn are just wrong about the properties that they attribute to networks (e.g. Cummins, 1994 and Clark, 1989). These options are listed in the table below.

Representational Options for Connectionists:	Some Examples:
1) No representations.	Brooks, Beer, the Churchlands.
2) Representations, but the representations are different from classical rep.'s. Two versions: a) No constituent structure, and this is no problem. b) No constituent structure, but want to show how such can emerge or be implemented.	2a) Waskan, Bechtel, Horgan/Tienson, van Gelder. 2b) Smolensky, Hinton, Touretsky, Pollock
3) representations are like those in classical systems.	Clark (circa 1989), Cummins, Toribio, Berkeley et al. (1995).

A second class of arguments for attacking Fodor and Pylyshyn's charge against connectionism is over the issue of rules. As mentioned, the classical model specifies the presence of explicit rules. Fodor and Pylyshyn maintain that explicit rules, of the recursive variety, to be specific, are required for systematicity and productivity, and that connectionist networks don't have these sorts of rules. Three arguments are made against Fodor and Pylyshyn on the issue of rules. The three are rough analogs of the three positions with respect to the issue of representations. The first position is that neural networks basically don't have rules of any interesting sort — and neither do people — thus, this feature is in fact a desirable one if we are interested in modeling cognition (Horgan and Tienson, 1987). Therefore the connectionist should not concern themselves

over a lack of rules. The second position is that neural networks do indeed have rules, but the rules of neural networks are implicit (McClelland, Rumelhart, and Hinton, 1986, Smolensky, 1987, and others). What is meant by this is that there is no place where the rules can be read out of the machine, in the way that rules in a classical system can be. It is hoped that implicit rules will be able to do everything that an explicit rule can, and it is argued that implicit rules may also have valuable characteristics of their own. A third position is that neural networks have rules, and they are as explicit as anything in GOFAL (See Aizawa, 1994, and Berkeley et al., 1995, for examples).

2.32 Beating the Dilemma:

In all three cases (for both rules and representations) there are two conditions for an argumentative strategy against the dilemma to succeed. First, it must provide an accurate description of the network properties, that is the premises about networks must be true. Second, it must be empirically adequate within those confines — it must really address Fodor and Pylyshyn's concerns. Now one strategy here is to claim that Fodor and Pylyshyn have inaccurately described human cognition (option 1), so a part of the second condition includes conforming to generally accepted facts about human cognition. The results of an examination of these strategies to defeat the dilemma will be: i) Option 2b is the one that both best describes network properties and cognition — the others either don't accurately characterize neural networks or make mistakes about human cognition. However, as we have seen, 2b implies incompatibility of the two cognitive architectures ii) Nevertheless the connectionist may (even must) accept one horn of the dilemma, that their approach is inadequate, without the consequence that they are irrelevant to cognitive

science. The reason is that Fodor and Pylyshyn also favour an empirically inadequate account of cognition. The best that can be said for either view is that both approaches are good at doing some kinds of tasks and poor at doing other kinds of tasks. Neither can account for the full range of observed cognitive capabilities, as required by any adequate theory. So we can only choose between the theories by making an arbitrary choice about which cognitive capacities are more central or more important to explain. This is clearly not adequate grounds for theory choice.

2.33 Reply 1: Anti-Representationalism:

To begin, I will give a brief explanation about why representations might be important to cognitive scientists. The importance of representations motivates a detailed discussion of some of the arguments about representations with respect to Fodor, Pylyshyn and Smolensky, and the debate between them.

To some extent, how one feels about the issue of representationalism depends on what one thinks about intentional psychology. The reason for this is that the physical symbol hypothesis is used, in part, to explain how things like beliefs and desires can be cashed out as physical states and processes. If you don't believe in belief, as for example the Churchlands don't, then there isn't any need for the physical symbols and states to instantiate them. A person taking this position won't much care about representational issues, indeed they will likely as not simply ditch representation as part of their explanatory apparatus, if for no other reason than to avoid the various problems with it.

Alternatively, if one is an intentional realist (and I don't include either Dennett or Clark in this category even though they might agree with some of what follows), then one

might be inclined to be skeptical about any explanation of mental processes that leaves intentionality aside. The upshot of this is that the conclusion that one draws with respect to arguments about connectionism and representation may depend on your views about intentional realism. Being an intentional realist myself, it is therefore essential that it be shown that connectionist systems are representational. No middle ground between classical architectures and networks is otherwise possible.

Alternatively, how one feels about mental representations may result from whether or not such entities have any explanatory and/or predictive utility in describing the operation of either biological intelligences or the operation of cognitive models. It is nevertheless necessary to show that connectionist systems are representational. The reason is that if two models of cognition (or models of anything else seeking to explain the same thing) postulate different kinds of entities, in particular, if one entails the existence of representations and the other entails the non-existence of representations, then a contradiction results from any attempts to reconcile the two. This follows because of what constitutes a valid entity in science. The two generally observed criteria are that the entity be necessary for either explanation or prediction, or both. Historical examples are easy to find, for instance the controversy between the phlogiston theory of combustion, and Lavoisier's oxygen theory of combustion. Lavoisier's theory entailed the non-existence of phlogiston, because postulating phlogiston was of no explanatory or predictive utility. The outcome was of course the abandonment of the phlogiston theory, and not a reconciliation with the competing theory.

Now we turn to how the claim that connectionist networks aren't representational

in character can be used as a response to Fodor and Pylyshyn. One of the horns of the dilemma posed by Fodor et al. is that connectionist networks are inadequate as models of cognition. This is because these networks, it is claimed, cannot token or use syntactically structured representations, which are in turn necessary for systematicity. If you take the position, however, that connectionist networks engage in cognitive activity *without representing at all*, then it is moot to argue that they cannot do so in a certain way. It can then be reasonably claimed that it is an open question whether or not connectionist networks can engage in systematic cognitive processes in some other fashion.

Furthermore, as networks *do* seem to be a good way of modeling certain kinds of cognitive activity, such as pattern recognition, then it may be that cognitive scientists need not be as committed to the representationalist thesis as they currently are. Therefore the dilemma is really being attacked in two ways i) It is argued that Smolensky is wrong about connectionism having representations, therefore Fodor and Pylyshyn have arbitrarily restricted the target of the dilemma to a false characterization of connectionist networks (i.e. its a straw man argument) and/or ii) Fodor and Pylyshyn are wrong about the essential characteristics of human cognition, specifically about its representational nature. Now to the arguments in detail, and why they are wrong.

There are two sets of arguments against treating connectionist systems as representational at all. The older arguments come from a variety of sources, but the basic argumentative structure is generally the same. The first step is to take some kind of cognitive activity, and describe that activity according to the standard representational model. The second phase of the argument is to then build a network that will perform the

activity without any recourse to, or need for representational resources. Thus, an alternative explanation is offered that does not require any reference to representation, belief, desires and so on. Brooks and Beers are noted for such work, generally performed with insectile robots, and based on the biological neural networks of cockroaches. The conclusion then immediately follows i.e. that representations are explanatorily unnecessary for at least an important class of behaviors, therefore connectionism can get along just fine without them. Clark and Toribio (1994) have critiqued a number of such claims. As one might imagine, the trouble is with the cases chosen. In each case, there is sufficient ambient information in the environment to drive the network appropriately. This is problematic because explanations that require representations are generally those in which environmental information is inadequate for determining behavior. That is, precisely the same kinds of circumstances that defeated behaviourist explanations of behavior.

If we assume that Smolensky is correct about the nature of connectionism, then for the project at hand, there are other difficulties as well. Both the biological and artificial networks described in these kinds of examples aren't anything like the neural networks that most connectionists are interested in. For one thing, information and processing isn't distributed (recall that Smolensky and others are emphatic about this as a defining characteristic of connectionist systems as distinct from classical architectures). All of the networks in question can be completely described as cybernetic feedback control systems coupled with 1) sensory systems 2) some sort of effector or actuator mechanism (joints and muscles usually) and 3) a central pattern generator for cases of complex periodic motions such as insect walking. There is no need for any level of analysis above the level

of the neurobiological or the electrical, whether it is described as connectionist or otherwise. Thus, talk about the evolution of vector trajectories through a phase space etc. is as equally superfluous as talk about representations are. Given then, that the networks in questions aren't the right kind of network by the lights of many connectionists such as Smolensky and Hinton, it is difficult to see how they could be used to support the conclusion that connectionists either don't have, or can do without, representations. Rather, we are restricted to the statement that *some* networks don't have representations, and as we have seen, these networks only function in very restrictive cases.

There is a second argument against regarding connectionist systems as representational, formulated by William Ramsey. Ramsey sets up three diagnostic criteria for asking whether or not connectionist networks have representations: 1) Are the representations similar in structure and function to those in classical architectures? 2) Is the notion of representation adequately strong? And 3) Would dropping the term "representation" from the connectionist vocabulary result in any explanatory gaps? Ramsey then applies this diagnostic strategy in turn to both the weights and nodes of neural networks.

In the case of the first criterion Ramsey states that the answer is probably "no" by definition, but then concedes that this isn't much of a problem for the connectionist. First, the standard definition of what a representation is was formulated within the GOFAI paradigm, and nobody wants to raise artificial distinctions on the basis of definitions alone, unless the connectionist definition just isn't any good. Second, the connectionist may simply reply that their notion of representation is not different in any important respect,

and indeed some might reply that in fact their notion of representation is better. Smolensky and Hinton, for example, are both quite clear in stating that a distributed representation is one of things that gives connectionist networks their novel and useful properties. Hence, Ramsay turns to his remaining diagnostic criteria to investigate these claims. He looks for representations in two places. First, in the weights of the network and, second, in the individual units of the network. He concludes that the notion of representation is doing no good for the connectionist, and ought to be dropped.

In the case of connection weights, Ramsey describes the notion of representation being used — specifically that the weights are encoding information relevant to performing tasks (and are therefore representations of that information) — and concludes that the notion of representation as applied to the weights of the connections is too weak. His concern here is pan-representationalism. As he says:

[...]it would be impossible to come up with an explanatory model that does not use *some* sort of structure to convert input to output. Any system that carries out a cognitive task — indeed, any system that converts input to output through the use of internal mechanisms — employs structures of this sort. [...] When ‘representation’ is interpreted in this way, saying that a connectionist model carries out a given cognitive task by using internal representations amounts to nothing more than saying that it carries out the task by using structures causally relevant to carrying out the task!

Note however, that there is an alternative conclusion to this argument. Specifically, one might choose to conclude that it is metaphysically necessary for information processing systems of any sort to employ representations. The pan-representationalism here isn’t vicious, after all. It would not, for example, apply to most things that have inputs and outputs (car engines, hearts, turbines, etc.) because the causally relevant internal structures

in these objects don't encode information²².

Ramsey then turns to examining the nodes of a network as potential bearers of representations. He concludes here that the notion simply isn't doing any work. All the explaining that needs to be done can be done without reference to representations. Here is the relevant contrast: In classical models, representations are invoked either in a heuristic role in interpreting a system when performing task decomposition analysis, or as a causal mechanism in teleo-informational accounts such as Millikan's and Dretske's. The former state of affairs certainly does not apply to connectionist architectures, as nearly everyone agrees. After all, Ramsey argues, if task decomposition analysis of a connectionist system was possible, it would *be* a classical architecture — it would necessarily have discrete and identifiable rules as well as representations — and this is precisely what most connectionist explicitly deny. Ramsay then turns to the possibility that nodes may be like Dretske style indicators. He concludes here that the analogy isn't good. The reason is that Dretske insists that 1) an indicator is designed or evolved to represent certain pieces of information and 2) the indicator is recruited into a cognitive architecture because of what

22

In the sense that most cognitive scientists mean. From a physical perspective, however, this is false. All of the above objects do in fact encode information, as does any object with more than one state (i.e. everything), and in precisely the same sense as a computer. This is because information, contrary to popular belief, is a physical property and not a cognitive property. Entropy, for example, is an information theoretic entity as well as a thermodynamic concept. And this has consequences for computation. For example, thermodynamically reversible computations (as a physicist means it) are faster and more efficient (information theoretically) than non-reversible computations, like those in currently available computers. See Frank and Knight (1997) for details and a proposed architecture. The philosophical consequence of this is that problems like pan-representationalism might not be problems at all, they might be basic facts of existence.

it represents. Nodes, especially hidden units (which are usually the category of node invoked as representation bearing), in comparison, simply don't have these properties. So far so good. But if representations are neither heuristically or causally necessary, then what are they doing? Ramsey concludes that they aren't doing anything at all for the connectionist.

Does this go wrong? The answer, I suspect, is yes, and in two places. First, it is probably a mistake to look for representations either in the hidden units or in the connections, as separate entities. Normally when connectionists talk about representations they refer to activation vectors, and while activation vectors are records of unit activity, unit activity results from both the activation functions of the units and the summed strength of the input connections. Given that unit activity and connection strength are so closely intertwined, it seems to be a conceptual error to try to peer into either one or the other to find representations. This follows because, when analyzing a record of unit activity, one is *already* looking at both the activation function inside the node and the weights of the connections automatically. And unit activity, so described, can be causally efficacious in producing subsequent unit activity both in general (Aizawa, 1994), and more transparently, for example in RAAM style networks, where the output of the hidden units in particular, is recursively used as input in later cycles of processing (Pollock, 1990).

Second, while it may be the case that we cannot meaningfully invoke the notion of representation to explain the activities of nodes or connections separately, we certainly can invoke the notion of vector analysis to explain the functioning of the network as a total entity. Moreover, vector analysis isn't just for fun, it is required to explain certain kinds of

network properties and behaviours. The relevant question then, is not “Do connections or hidden units have representations?” rather, it is “Is talk about vectors the same as talk about representations?”. The answer seems to be both yes and no. If you insist that vectors *qua* symbols must have the same kind of constituent structure as symbols in GOFAI, then the answer is very possibly no. However, if one is willing to deny constituent structure as a necessary feature of representations, then there does not seem to be any reason why a vector cannot be a representation. This brings us to the next section, where this possibility is discussed.

2.34 Reply 2: Representations are Distributed:

A second possibility is that connectionist networks do indeed have representations, but these representations are different from the syntactically structured representations of a physical symbol processor. This is the favored position of many, maybe even most connectionists. These people include the original PDP working group (1986), Smolensky (1987 and 1990), van Gelder (1989 and elsewhere) and Haugeland (1989 and elsewhere), just to name a few. The first step here is to explain what is different about connectionist representations, and then why that matters.

The central claim of this position is that connectionist representations are *activation vectors*. An activation vector is the sum of the products of the units by the strength of their activation. Thus, an activation vector defines a point in a phase space of dimension equal to the number of units in the network. What is special about a vectorial representation is its mathematical character. Vectors are a class of mathematical objects

for which certain operations are defined (e.g. vector addition and vector multiplication)²³, and to which certain other rules apply. Importantly, vectors and their operations are only defined on a field, which is a class of algebraic structures, in particular, activation vectors and tensor products are defined for the field $\mathfrak{R}$. In contrast, the mathematical objects in classical computational mathematics (formal logic, automata theory and recursive function theory) are the sets of well-formed formulae of particular languages, over which other, different operations are defined, and to which the syntax of the formula is relevant. Specifically, they are discrete operations on discrete structures. The functions that govern vector transformations in neural networks (systems of differential equations) and transformation of wffs are not of the same character, so its hard to see how a machine employing the one type of representation could do the same kind of symbol processing as the other. Worse, its not at all obvious how one could define an activation vector or tensor product as a wff in a discrete language, so in a real sense its hard to see how a connectionist symbol structure could even exist in a classical computational architecture (or alternatively how to define a wff as a vector in $\mathfrak{R}$).

Another difference with vectorial, distributed representations, is that the kind of information that they encode may be of a different genera than the kind of information encoded in a standard representation. Haugeland, for example, suggests a taxonomy of three representational genera. The first is sentence like (standard representations). It has constituent structure, syntax, and moreover, the elements that constitute the structure are semantically discrete. One can interpret the contents of any one constituent part without

²³See Taylor and Mann (1973) p. 309ff for more detailed discussion

knowing what any of the others are. The second type is imagistic. This genera includes maps, pictures and so forth. Pictures and maps encode information about continuous dependant variables, such as light intensity, color etc., along continuous, independent dimensions (length, width, depth, time). The third genera are the distributed representations. Their semantics are supposed to be something in between images and sentences. Haugeland and van Gelder suggest that distributed representations have two features of note. The first is that they are spatially spread out, but more importantly, information is superimposed over more than one unit or node in the network. In the former case, a piece of information is represented over many active units, in the latter case, any given unit can participate in the representation of more than one thing. Thus a single unit may be part of a representation of 'angelfish' at some time, and part of a representation of 'shark' the next.

When you couple the fact that the operations defined on these mathematical objects are fundamentally different, with Haugeland's notion that the content of the representation might also be fundamentally different, the advocate of the distributed, vector style representation runs into Fodor-style complaints head on, because they acknowledge from the first that these vector representations *prima facie*, cannot do the kinds of structure sensitive operations that Fodor wants to see and may not even encode the right kind of content. There are generally two kinds of replies: First, it is argued that human cognition isn't systematic, and Fodor and Pylyshyn are wrong to insist that this must therefore be a fundamental feature of any viable cognitive architecture. Second, Smolensky argues that systematicity, or more precisely an approximation of it, is an

emergent property of connectionist networks with distributed representations. Hence the motivation for describing networks as sub-symbolic. Let us take these in turn.

Waskan and Bechtel provide a recent example of the kind of argument that is typically employed against the need for syntactically structured representations. The question that they ask is whether or not networks trained to fulfill a variety of cognitive tasks “develop transitions between internal states such that we will identify these states as encoding syntactic structures?” (Waskan and Bechtel, 1997, p. 51). They conclude, one sentence later, that “The answer would seem rather clearly to be ‘no’”. Thus the first premise of their argument is that the number of cognitive functions that require representations with a recursive, compositional syntax is in fact quite small. They initially suggest that it is only language and perhaps logic that need this kind of representation. While arguments are thin on the ground in this section of the paper, Waskan and Bechtel do hint at one earlier in the paper, and it is of little moment to find the required arguments fleshed out in other papers on connectionism.

Here’s a frequently used type of argument. First, the opponent of syntactically structured representations makes the following claim: The combination of syntactically structured representations with explicit rules (and one may as well not bother with the one without the other) makes for a very rigid cognitive architecture. What follows from this rigidity is that such an architecture is unlikely to be able to provide the observed flexibility of human thought. Moreover, the kinds of things that syntactically structured representations explain, such as how logical operations are implemented, are things that are cognitively difficult and usually slow. Humans are error prone in these cognitive areas,

and if they truly possessed the rigid and inflexible rules required to implement the classical architecture, we wouldn't find logic difficult, and wouldn't make the kinds of logical mistakes that are prevalent in human thinking. Connectionist networks can not only be used to explain such errors, but can replicate the pattern of error as well. Thus, an explanatory virtue of a connectionist network is the fact that networks have difficulty with the very same kinds of tasks that people do. Nevertheless, there remain conspicuous examples, such as language learning, that seem to require a compositional syntax in order to explain the systematic nature of linguistic ability.

The next step then, is to suggest that even in the case of language, there is no need for *internal* compositional resources. Waskan and Bechtel refer back to a number of examples of networks, described by Horgan and Tienson (among others) that reconstruct syntactic relations, such as a RAAM network, by identifying or copying the pattern of an external medium. The claim here is that no internal grammar or other innate endowment on the part of the network is required for reconstructing this pattern, as long as there is a syntactically structured external medium available. Language, it is argued, is such a medium, and by implication, must then be completely external to the brain (no innate endowments that specify grammatical structures, and so on). Here I must side firmly with Fodor however, for the trouble with this argument is that from an empirical standpoint, its pretty obviously wrong.

In the first place the flexibility of human cognition is somewhat overstated. While it is true that cognition is effortlessly flexible in some areas, there are many examples where in fact human cognition is every bit as fragile, rigid and unforgiving as a classical

architecture would suggest. Some famous examples can be had from sensory psychology, such as the Muller Lyer illusion, where our perception of arrow length is cognitively impenetrable: even after we learn that the two arrows in the Muller Lyer image are in fact the same length, our visual system persists in interpreting the arrows as being different in length.

Second, the complaint that humans have difficulty with rules and logic is misstated. Alternatively, the complaint ought to be phrased as: Humans don't employ *logically valid* rule-based schemes for reasoning for all tasks. Remember there is a problem for Fodor and Pylyshyn if and only if there are no rules at all, and the modified statement of complaint leaves two ways out of this. First, Fodor and Pylyshyn may reply that there are a great many cases where rule-based logical inference *is* employed, with success, in a variety of everyday cognitive tasks (see Smith, Langston and Nisbett, 1992; and Rips, 1994 for reviews), so any account of cognition, including connectionism, had better be able to account for these, and bringing us firmly back to one horn of the dilemma, i.e. empirical adequacy. There is also the possibility that the cognitive activity is rules based, but that the rules are different from those of logic, for example language grammars. Again, an empirically adequate account of cognitive will need to resort to rules. Finally there are also serious questions about the validity of many of these studies that purport to show that cognition is neither i) rules-based, (which would be a problem for the classical cognitive scientist) nor ii) logical in character (about which the classical cognitive scientist might care little, as long as they get to use rules).

The first problem with these studies is methodological. What a psychologist must

do when testing a subject is to show that the task that is being performed is evolutionarily normal for the subjects in question. Otherwise how the subjects interpret or encode the task may confound the results. This is called ecological validity. Second, ecological contexts can relativize what counts as a rational behaviour, but it does not relativize what counts as logical. So the psychologist must be careful not to conflate mathematical rules of logic with rational thought, or with rules-based problem solving.

Stephen Stich (1993) does cite empirical evidence in support of the claim that logically valid rules-based reasoning makes for a poor description of human cognitive capacities. These are precisely the kinds of results that are illegitimately appropriated by the Waskan, Bechtel and others in support of the conclusion that cognition isn't rules based. I shall take up one example that Stich discusses (Tversky and Kahneman, 1983). The problems with this study apply generally to the papers that Stich cites, and exemplify the problems I have described above. The authors ask subjects to rank the probability of a series of statements, including the following two statements: vi) "Linda is a bank teller" and viii) "Linda is a bank teller and is active in the feminist movement". The task, the authors suggest, involves the law of compound probability, a well-established result in probability and statistics. In an attempt to ensure that subjects understood this, vi) was replaced with vi)* "Linda is a bank teller, whether or not she is active in the feminist movement"²⁴. Nevertheless, subjects frequently ranked statement viii as *more* likely than

24

However, the only way to guarantee that subjects classify the problem correctly is to *tell* them "this is a problem that requires you to use the law of compound probability" Unless it is clear *how* the subjects are classifying the problem, there will invariably be the methodological fear that their interpretation of what they are supposed to do will

vi) or vi)*. Of course, if the law of compound probability is used to assess the validity of this statement, it is false, as the probability of a subset may never be higher than the probability of the set of which it is a member.

Let us recast this conclusion in an ecologically valid context. The context is that we are performing an ecologically useful task, like *looking for Linda*. In this case, it would make no sense for the subjects to use the law of compound probability — as it is the *relative* probability of vi and viii that is useful to formulating search strategies. This is because search strategies are classical examples of game theoretic reasoning — which follows from the fact that search strategies have payoff matrixes — that is, the most worthwhile strategy is a function of the effort invested (the cost, *c*) times the probability, *p*, of success, *p* being assigned relative to other alternatives. Thus, if the relative probability of viii is higher than vi, it may be worthwhile to ask around for Linda at the local Women's center (an easy task) prior to searching every bank in the area (a harder task), which is what the subject will have to do if the relative probability of vi) is low. Surely, even minimal rationality entails that the best way for an agent to perform a task is the way that is most likely to work, all other things being equal (i.e. the subject does not face multiple and conflicting task constraints). As the use of the law of compound probability cannot be usefully employed in a search strategy, it is *not* the best available method, and therefore, *not* the most rational way to proceed. The lesson here is that i) what counts as logical or not may depend on the method of description and ii) again, that we have to be careful about conflating logic and rationality, because the latter may be

confound the results.

context relative while the former most certainly is not. Under one description, the subjects behave illogically, under another, their conduct is logically rational. Stich (1993, p.9) notes that such concerns i.e. 1) ecological validity and 2) what counts as rational²⁵, have been frequently raised by psychologists against the empirical results that he employs. He and his colleagues are frankly dismissive of these concerns, suggesting that the former concerns are easily handled (no argument provided), and that the latter concerns merely lead to a clash of intuitions about what is rational. However, no such clash of intuitions arises here.

What's really happening here is a clash about data interpretation and methodological principles. The widely accepted principle of ecological validity (to which Stich refers) means that the onus is on the psychologist to show that in this specific case the subjects ought to have used the law of compound probability in a *real world* task. In general, what must first be established is what rules are being (potentially) employed by the cognitive system, and what constitutes valid outcomes for a particular collection of rules and tasks. Tversky and Kahneman simply didn't do this. Therefore we cannot conclude that humans have trouble applying the law of compound probability, because it is not clear that in an ecologically valid context that is what they ought to have done for this particular task. Nor can we conclude that *no* rules based scheme was used to solve the given problem, as the authors never attempted to ascertain which system of rules might be cognitively relevant (for example we could test for rules based on relative probability),

25

If the ecological validity is accepted as a methodological concern, then what is rational depends on ecological context. If this is the case, then it follows that 1) and 2) cannot be discussed in isolation.

rather, they arbitrarily imposed an optional (and quite poor) interpretation on their subject's performance.

Waskan and Bechtel's final point, that the required syntactic resources of cognitive systematicity are externally available through language, is also empirically unlikely. The reasons for this are simply that language really isn't an external medium in the way that Waskan and Bechtel require. This has been argued at length by Chomsky in various places, by Pinker (1994), and explicitly against the connectionist in Pinker and Price (1988), and I won't repeat the arguments here, save to describe one compelling case, described in Pinker (1994). The example is a striking case of creoleization. In the 1970's in Nicaragua, linguists observed that the children at the national school for the deaf had successfully grammaticalized sign language. Thus, they created the syntactic structure of the external medium *de novo*. Given that these children couldn't yet read or write, and couldn't understand existing natural languages, there is no external resource off which they could have patterned their cognitive activities, yet not only were these children cognitively normal, but they manufactured a rich, syntactically structured system of representation. If Waskan and Bechtel's hypothesis about the source of cognitive systematicity were correct, then not only would it have been impossible for this particular creoleization to occur — and it did happen — we would also have to predict that the children in question would have had noticeable cognitive impairments in the areas of logical reasoning at least, and they didn't. Arguments of this form (see also Bechtel and Abrahamsen, 1991, p. 248ff, for yet another example) are only intuitively compelling on an instrumentalist conception of cognitive levels, which as I have argued, is probably

false²⁶.

Smolensky (1988) takes a very different approach to responding to Fodor and Pylyshyn's critique. Essentially, Smolensky takes issue with Marr's notion of levels. Fodor and Pylyshyn's complaint, recall, is that either connectionism is empirically inadequate, because it lacks the capacity to model systematic cognitive processes, or it is merely an implementation of a classical architecture. Smolensky wishes to position connectionism as a description of something in between the algorithmic level (the level of the classical Physical symbol system) and the implementational level, the level of the neuron or the integrated circuit. Thus, Smolensky calls connectionism a *sub-symbolic* paradigm. Representations are likewise sub-symbolic, being composed of microfeatures distributed over several nodes, rather than being composed of atomic constituents. This approach has several advantages, but also several bizarre features. The advantages? Well, first, Smolensky need not give up any of the capabilities of the classical approach, because he claims that something like the behavior of classical systems is an emergent property of sub-symbolic systems. Thus, if his proposal is workable, he won't have to worry about how, for example, to implement symbol processing with a compositional and recursive syntax. This gets him out of Fodor and Pylyshyn's dilemma. Second, he does not have to claim that connectionist networks are physiological analogues of biological neural networks.

26

As an aside, an instrumentalist conception of cognitive levels may, in any event, be in more global trouble, if Hacking-style arguments about ontology and experimental intervention can be applied to cognitive science (see Hacking, 1994, ch. 16).

Some connectionists are forced to take the position that artificial networks are like biological neural networks because they also share a commitment to eliminative materialism. This is relevant because one way of escaping Fodor and Pylyshyn's dilemma is simply to accept one of the horns. The eliminative materialist can simply say that yes, they are merely describing an implementation, but *that's all there is to describe anyway*. This is unfortunate for them as it pertains to connectionism, because the physiology of a neuron bears little resemblance to a standard connectionist unit. This is underappreciated in the literature, and on my view constitutes a major advantage for Smolensky's position. To give one example of the differences between a neuron and a unit — one that is more significant than those listed by Smolensky — a connectionist unit processes every input in the same way, and has but a single *type* of output. Not so for the neuron. Each separate input is processed differently by the neuron, depending on where the synapse connects with the dendritic tree, what the post-synaptic receptor is and so on. In addition, neurons have multiple kinds of outputs, some of which are qualitatively different, and therefore are able to modify the activity of post synaptic cells in a variety of ways. One example of this occurs in certain thalamic neurons where a weak inhibitory post-synaptic potential simply modifies the rate at which a continuously active post synaptic cell fires. A stronger inhibitory post-synaptic potential, however, can cause the cell to fire in bursts instead.

To be sure, it is possible to make connectionist networks more resemble their biological cousins. Hinton, for example, has developed networks with fast and slow weights to accommodate the observation that post synaptic activity can be modulated over varying periods of time. What is important and relevant about this example is that

networks with fast and slow weights have useful and novel computational properties. Presumably other features of biological neurons, once copied by the connectionist, might do so also. The consequence of this is that an artificial neural network that really *did* implement the various features of biological networks might bear little resemblance to the connectionist systems we have now. This tends to undermine any claims that connectionists and eliminative materialists may have with respect to one another.

This brings us to the odd properties of Smolensky's views. First, Marr's notion of a cognitive level is a metaphor for actual computer hardware and software. And computers as physical devices actually have an algorithmic and implementational level that's *really* physically there inside the box, as it were. This is what makes the metaphor plausible. Smolensky adds to this picture — he adds a new level — but provides us with no reason for finding it plausible. The metaphor breaks down with his usage, and I cannot see any way of fixing it. Nevertheless, I don't think this is a major problem for Smolensky if he can demonstrate that the properties of classical Physical symbol system's are emergent properties of neural networks, because this would *de-facto* establish that an additional cognitive level is required. This would be nice because the additional level would, as Smolensky claims, solve the problem of reconciling the classical and the connectionist approach to cognition.

Importantly, Smolensky's claim is not superficially implausible for another reason. There are other valid schemes of classifying levels that *are* more fine-grained than Marr's or Pylyshyn's, such as Newell's (1982), for example, which specifies a six level architecture for an artificial intelligence implemented on a classical machine. Can this be

made to work in Smolensky's favour? The answer is no. Two points need to be made about Newell's scheme *vis à vis* Smolensky. First point: Newell makes a principled reference to actual devices, while Smolensky does not. Thus the computer metaphor is a good metaphor for Newell, while it is of doubtful relevance for Smolensky. Second point: For a cognitive scientist, Newell's scheme is easily collapsed into something like Marr's. To illustrate, some of these levels, such as the *device level*, and the *circuit level*, are of relevance and utility only to engineers concerned with hardware. Indeed, Newell remarks on the irrelevance of these levels from a programmer's perspective. A programmer, for example, need know nothing of the circuitry of a computer. Similarly, from a cognitive science perspective, these two levels deal exclusively with physical mechanisms, thus they are easily subsumed into Marr's implementation level²⁷. To continue, a computer scientist needs to distinguish between the level of *register transfer*, and the *program* level, because there is a meaningful distinction between an operating system and an application program for someone who is writing code. Both of these levels, however, work with algorithms, and the distinction as applied to cognitive science is dubious because it is not evident that biological intelligences have anything like an operating system, distinct from the

27

Sadly, neuroscience is thus far not very good at showing how individual elements (neurons) are built up into cognitive devices, although there are some limited success stories of note in sensory neuroscience (Heiligenberg, 1991; Hubel, 1988 for example). There is a meaningful distinction between cellular neuroscience and systems neuroscience, however, which might be analogous to the device/circuit distinction. More probably, Patricia Churchland's so called *quaquaversal* strategy will need to be employed to understand the relationship between neurons and assemblages of neurons, where an algorithm is first specified via cognitive psychology, and then a neural implementation of this algorithm is demonstrated.

algorithms that constitute cognitive processes (see Pylyshyn, 1984, for a similar point).

The result is that there is no distinction between these two that is of immediate relevance to cognitive science, therefore, the two levels can again be subsumed under the more general heading of the algorithmic level, with no loss of explanatory power. Moreover, an operating system is a feature of a classical architecture, so adding the register transfer level to Marr's scheme might make the schema of limited applicability to connectionist systems.

In summary, the lesson is that it is certainly possible for there to be more than three cognitive levels, as Smolensky contends, but any new level ought to meet two conditions. First, it shouldn't collapse back into Marr's scheme. Second, the new level should be added in a principled way. The notion of a sub-symbolic level possibly passes the former test, but certainly fails the latter. To pass it would require an account about how higher cognitive levels, such as the symbolic level, can be built on the sub-symbolic, i.e. how sub-symbols can be built up into symbols, and so on — which is what we don't yet have. For purposes of comparison, Newell's six level scheme does pass this test (while it fails the first test) — because we know exactly what each level corresponds to in a computer, and how they interrelate. The next step is to see if there is a way to show that the sub-symbolic can somehow give us the symbolic, by using the idea of emergent properties.

The second unusual claim made by Smolensky is that something approximating classical computational properties are emergent properties of networks. Yet there is no argument, nor any evidence provided to support this statement in any of Smolensky's work on this subject. I propose to see if an argument can be manufactured on Smolensky's behalf. Let us begin by assuming that there really are emergent properties. This ought not

to be unreasonable; the classical cognitivist is, after all, making the same bet. To be clear, the argumentative assumption is that emergent properties are a *general* feature of physical systems, and not just features of cognitive systems. From a strictly reductionistic perspective, this is a controversial assumption (Don Ross, personal communication). A quick review of emergent properties indicates that there seem to be two different types. There are probably more, but there are no others that I know how to usefully characterize.

The first kind of emergent property shows up in thermodynamics. The macrophysical properties of thermodynamic systems are supervenient on the microphysical properties of the molecules that compose them i.e. quantum mechanical properties. Yet quantum mechanics is very unlike thermodynamics. It turns out, however, that if one takes the ensemble average of the behavior of a group of molecules, thermodynamic properties result. Moreover, emergence of thermodynamic properties is smooth and continuous — the more molecules there are, the more the system behaves thermodynamically. This is called the correspondence principle, and is described in most textbooks on statistical thermodynamics²⁸. Call this aggregative emergence, because the property is simply a result of scale. Once you start to have enough of something, the property starts to emerge.

This kind of emergence might apply to neural networks in one of two ways. The first, and less plausible, is suggested by the notion of a distributed representation. If some properties already emerge from the distributed character of representations, then others might do so also. And a network with distributed representations does exhibit emergent properties. Specifically, as a network is changed from one where representations tokened

²⁸P.W. Atkin's *Statistical Thermodynamics*, third edition, for example.

in a single node, to one where representations are tokened across several nodes, the network acquires new capabilities. Perhaps if the representations are adequately distributed, and the network is appropriately structured to allow recursive operations, and/or variable binding, then classical computational features will emerge/appear too.

There is some evidence to suggest that this may be the case. Both RAAM networks (Pollock, 1990; and Chalmers, 1990) and those employing tensor products (Smolensky, 1990) exhibit some classical computational features, such as variable binding, compositionality etc., albeit in a limited way. For example, Pollock's RAAM network shows limited productivity. Not all of the output sentences, however, were well-formed (Pollock, 1990, p. 94). Chalmers has taken a further step, with the same network architecture. By using one RAAM network to encode a distributed representation of a syntactically structured input (natural language sentences in this case), he was able to get a second network to produce syntactically well formed rearrangements of the sentence components, without first decomposing the distributed representation into its component elements. This demonstrates so-called *direct access*, which is a property of classical systems that can implement semantic systematicity. Chalmers confesses, though, that he was forced to use the initial RAAM network in a way that could legitimately be construed as an implementation of a classical architecture. He says: "So far, what has been described amounts merely to a connectionist implementation of classical symbol structures" (Chalmers, 1994, p. 58). The operation of the second network, however, cannot be so described. Unfortunately, further work on RAAM networks has demonstrated that they are unlikely to fulfill the initial hope of fully implementing the kinds of cognitive features

that classical architectures are good at, such as variable binding, and systematicity (Callan and Palmer-Brown, 1997). RAAM networks have proven to be difficult to train, there are no guidelines for how many units are required for particular tasks, and it has proven difficult to establish how generalizable the network solutions actually are. There is ongoing work trying to solve these problems.

Smolensky has also been able to formally demonstrate that he can decompose vector products into their constituents. Solving this problem was a major obstacle to his claim in (Smolensky, 1988) that vectors could be said to have compositional structure in some meaningful or useful way (he addresses this problem in Smolensky 1990, section 3.1). However, the variable unbinding operation for tensor products, which is what is required to show compositional structure, demonstrates some limitations. The main problem is that for complete unbinding to occur, all of the input variables must be linearly independent. The reason for this is that the unbinding method requires the subsequent input of a vector to the network whose terms all possess a zero value, save for the term representing the variable to be unbound, which is given a value of one. As the units of the network display a pattern of activity that represents the tensor product representation, the zero value terms have the effect of “canceling out” all of the other vector components. However, when there is coarse coding of the input, as is often the case, this assumption does not apply, because there is semantic superposition of content across several input units. Smolensky summarizes other limitations in section 4 of the tensor product paper.

The suggestion that features of classical architectures might emerge from sub-symbolic architectures, though, probably won't work. The idea here, remember, is that as

one moves from the smallest possible unit in a structure (for example one molecule in thermodynamics) to an aggregation of many of these units (i.e. a gas), properties emerge. So in a connectionist network this means moving from representations in one node, to representations carried by an aggregation of nodes or even the entire network (and recall that Smolensky stipulates that connectionism means that representations are distributed, and that I follow him in this stipulation). What's wrong with this is that the suggested mechanisms of emergence would imply that 1) classical representational properties can be emergent properties of 2) connectionist networks which, in turn, emerge from classical computational systems. In the first place, this is the *opposite* of what Smolensky wants. Worse, whatever emergence is, if anything at all, it is a unidirectional, not a bi-directional, relation between the super- and sub-venient properties.

A second problem is that the story I am telling about emergent properties suggests that what an emergent property consists in is 1) a deduction from a structure, that 2) happens to be instantiated in the real world. For example, if we have an axiomatized theory like quantum mechanics, and we place this theory in a structure which allows for ensemble averages — which obviously our world *does* seem to allow — then thermodynamics is *logically entailed* by this structure. Now any reasonably complicated theory of course entails the possibility of many things, not all of which are instantiated in the real world, so this story won't precisely define what subset of all the possible emergent properties actually exist, unless we have lots more information (i.e. constraints) about the total structure. However we can certainly use this picture to tell us what sorts of emergent properties are problematic, by asking whether or not a consistent deduction from the

structure exists.

On my story of emergent properties, what Smolensky proposes amounts to this: We start with a kind of structure — a connectionist network²⁹ — that contains certain kinds of objects (vectors and tensors) that behave according to certain kinds of rules. The task is to show that a deduction exists from these objects and rules to the proposed emergent properties, namely classical computational objects and rules. To say the least, it isn't at all obvious (although it may be possible) how to go from the objects and rules of vector and tensor mathematics to those of languages, automata, and logic, which have very different and perhaps even contradictory properties (for example, the former is defined over a continuous structure, and the latter is defined over discontinuous, discrete structures).

To summarize: What was wanted was a way to show how either the sub-symbolic can be built up into the symbolic, or how the symbolic emerges from the sub-symbolic. If this can be done, then 1) adding the sub-symbolic level to Marr's scheme is made plausible and 2) we have a way of putting a connectionist system and a classical system together, i.e. build the latter on top of (or out of) the former. This, in turn, allows us to use the useful properties of each architecture for purposes of modelling cognition, rather than being restricted to one or the other approach. So far, the strategy of using emergent properties has not worked. After a brief note about interpreting the examples I'm using, I'll explore other possibilities for showing how the symbolic might emerge from the sub-symbolic.

²⁹Recall that connectionism is an *algorithmic* level explanation

A note about interpretation: While the empirical examples I have given may be *useful* as an exploratory tool (i.e to explore Smolensky's suggestion that classical representations can emerge from sub-symbolic systems), it's important to state that it isn't really plausible to read the papers that I cite above as work on emergent properties, because the authors in the Hinton volume aren't doing work on emergent properties. What they *are* doing is trying to either build symbolic/sub-symbolic hybrids, or build connectionist systems that can implement features of classical systems while retaining their distinctive connectionist character. For example, both Smolensky and Touretsky explore the possibility of implementing LISP commands, like *car* and *cdr*, using BolzCONS and vector products respectively.

A second possibility to explain how classical representations emerge from sub-symbolic representations is to suggest that classical computational properties emerge as the network gets bigger: more nodes and connections, or perhaps more layers or even dimensions of layers (i.e. a network in three dimensions) and so on. The problem with this suggestion is that, in the case of nodes and connections, it has proven difficult (Hinton, 1990) to make very large networks do anything at all, let alone do more. The reason for this is that many network properties and learning techniques scale up poorly. One particularly nasty case of the scaling problem involves the gradient descent algorithms, including back-propagation (Minsky and Papert, 1988, p. 260-261), which suffers from combinatorial explosion in large networks. In the case of layers, mathematical network analysis tells us that additional layers of units, even those with non-linear activation functions, don't necessarily add any further computational capabilities (McClelland,

Rumelhart, and Hinton, 1986). This apparently results because additional layers of hidden units can sometimes be reduced to a single hidden layer that is mathematically equivalent. In addition, adding layers often results in a trade-off between computational power and flexibility.

Another kind of emergence can be observed with non-linear dynamical systems. Chaotic systems, particularly, exhibit property changes when one of the system parameters exceeds some value³⁰. For instance, fluid flow changes from laminar flow to turbulent flow once the velocity of the flow exceeds some value (depending on the surface over which the fluid flows, its viscosity and other variables). The properties of the system become quite different — fluids develop eddy currents and such — and old properties obey different laws. For example, in turbulent fluids flow resistance changes from being proportional to the velocity squared, to being proportional to the velocity cubed. Emergent properties of this sort develop quite abruptly and suddenly, so I am going to call this transitional emergence. In this kind of emergence, new properties result from both microphysical properties (which are non-relational), and different arrangements of the microphysical constituents (which can be relational), not from ensemble averages of microphysical properties alone. Another example of this kind of emergence would be crystallization (which is also a chaotic process, as it happens), in which both the new state and the new properties partially results from the geometric positioning of the molecules.

The next step is to suggest what elements of a connectionist network might lead to

30

This is fairly standard non-linear dynamics. For a detailed mathematical exposition see Peitgen, Jurgens and Saupe, 1992.

properties that are transitionally emergent. Probably, the extent to which representations are distributed, and network size alone won't be suitable candidates, for the reasons already mentioned. One other candidate comes to mind, however, and that's network complexity.

A *complex* network can indeed be a chaotic system, or better, some properties of chaotic systems, such as strange attractors³¹, have been produced, and can be computationally useful in neural networks (Hertz, 1995). Note that there's nothing vague about what complexity means, in the context of either networks or chaos theory. There are mathematical techniques to quantify complexity, and indeed to determine whether it is the right kind of complexity to generate chaotic effects. How to generate the right kind of complexity in a network is a trickier issue, though. In order for a network to exhibit chaotic properties that are relevantly analogous to turbulence in fluid dynamics, two necessary conditions apply. First, Smolensky must be right about the mathematical character of distributed representations, because only non-linear functions display chaotic behaviour. Second, the differential equations defining vector evolution in a particular network must be shown to have at least one positive Lyapunov exponent (see Peitgen, Jurgens, and Saupe, 1992 for a definition). This means that, given two different starting points or inputs to the function, no matter how small the difference between them, will yield divergent output over time, or over subsequent iterations of the function. There does not seem to be any canonical method, yet, for producing chaotic effects in networks. The

31

A strange attractor is a somewhat discrete region towards which a function tends. Planetary orbits, for example, are in the *basin* of a strange attractor.

paper by Hertz, above, however, suggests that these two conditions can be fulfilled. The crux is, will a network that is also a chaotic system show emergent properties of the right kind, i.e. the sort that we attribute to classical symbolic systems?

The observation that some chaotic systems possess strange attractors, may, from my admittedly naive mathematical viewpoint, prove to be a useful feature. The reason is that strange attractors look, visually, like a discrete structure. The basin of an attractor, for example, has a roughly definable boundary that can be seen on a graph, the way the boundary of a discrete structure is graphically visible, and discrete structures are the kinds of entities that reside in classical physical symbol systems. If a strange attractor can behave computationally like a discrete structure, then Smolensky's idea that classical symbols (or something approximately like them) can come about from a sub-symbolic system, may be plausible. This is however, *extremely* speculative on my part, and the burden of proof very much rests with the connectionist to show that a case can be made for chaos.

In summarizing the conclusions of this section, I return to the issue of 1) cognitive levels and 2) the appropriate description of a connectionist representation. With respect to cognitive levels, given that, first, it is not plausible that connectionist networks are like biological networks, and second, that the case for emergent properties is nothing more than speculation at present, the only remaining possibility is that connectionist systems really are algorithmic level descriptions of cognition, as McClelland, Rumelhart, and Hinton (1986) have always contended. In spite of, or maybe even because of, the unique method in which connectionist systems represent, they are thus far empirically inadequate models of cognition, because connectionist systems have such difficulty representing and

manipulating the complex symbol structures that are believed to underlie a certain class of cognitive features identified by Fodor and Pylyshyn..

2.35 Reply 3: Connectionist Representations are Classical:

Thus far, I have looked at two ways of describing the representations in a connectionist system. Each was proposed because they were thought to give us a way out of Fodor and Pylyshyn's dilemma. The first was that there weren't any representations at all. This proved to be implausible, however, so it won't help out in resolving the dilemma. The second possibility was that the representations are distributed, as Smolensky stipulates. This was plausible, but it does not turn out to be helpful with the dilemma, because it does not look like either 1) distributed representations, by themselves, can give us productivity and systematicity or that 2) they can built up into symbolic representations.

The third variation on the representationalist theme is that connectionist representations, i.e. distributed ones, are just the same as classical representations, and therefore can implement compositional symbol structures with a recursive syntax as a matter of course. This argument entails that Smolensky (and therefore me as well) is wrong to stipulate that connectionist networks refer to networks that have distributed representations. It is wrong because the stipulation is vacuous. Here again there is a stereotypical kind of argument to support the claim. First, like some of the connectionist arguments against the need for representations at all, frequent use is made of example networks. The examples usually have the following properties in common: 1) the input is usually data with a syntactically structured form; often natural language sentences and the like 2) the output often is syntactically structured as well. Second, some kind of

justification is provided for *interpreting* the system as having syntactically structured symbols. Cummins (1991) for example, asks in virtue of what properties do we call and adding machine, an adding machine? The answer for him, is that the adding machine has an appropriate description at the cognitive level (the top level of Marr's schema), thus it is required to have appropriate input/output relations. Therefore, if we can accurately describe the input output relations of a network — he uses the Simplifier network — as a logic interpreter, the network can be said to be using syntactical information in some kind of a systematic way. Similarly, Clark (1989) suggests that systematicity falls out of holistic content ascription, that is, that we treat the thoughts of other people as systematic because we impose rationality as a constraint on thought ascription. In both cases the computational system, in the former a machine, in the latter a human mind, is treated as a 'black box'. The question of whether or not the system has a certain property is said to depend on how we interpret its activity. If the interpretation applies, then it must have the appropriate internal resources, whether those resources actually include syntactically complex symbols at the algorithmic level of description.

First of all, let me state that there ought to be no surprise or novelty in finding that Simplifier, for example, is able to do what it does (Simplifier performs simplification of logical statements in polish notation). Other networks have even been constructed that can learn and apply basic rules of inference in a limited way (Bechtel, 1994 for one example). These capabilities aren't surprising because what these networks are doing is recognizing and 'interpreting' patterns, a property for which connectionist networks are noted. Given that there is, very obviously, a pattern to the syntactic structure of logic and natural

languages, it is quite evident that networks *ought* be able to recognize it. What would be surprising is if they could not.

Classical computers, however, can do a great deal more with syntactically structured representations than merely recognize patterns, and perform simple pattern matching and analysis. And these other capabilities (see Hinton, 1991 for details) are the kinds of things that connectionist systems cannot do, or can only do with great difficulty under limited circumstances. Moreover, some connectionist researchers are quite specific about this point. Hinton, for example, in writing about automated learning procedures, such as back propagation, says:

[...] if these techniques can be applied in networks with greater representational abilities, we may see artificial neural networks that can do more than just classify patterns. But for now, the problem is to devise effective ways of representing complex structures in connectionist networks without sacrificing the ability to learn the representations. My own view is that connectionists are still a very long way from solving this problem [...] (Hinton, 1991, p. 3)

Where Cummins and Clark are lead astray, I suspect, is in their insistence that the way to decide whether or not a connectionist system has syntactically structured representations is to describe it at the cognitive level; once again using Marr's taxonomy. The trouble is that the issue about representations *isn't about what happens at the cognitive level*.

Representations are entities that exist at the algorithmic level, if they exist at all, because if there is no syntax, then there aren't any representations. Given that we can only describe, control and predict the behavior of computational devices — whether they be networks or classical Von Neumann machines — by appealing to and using representations in a certain way, we have no choice but to postulate their existence. The

instrumentalistic style of explanation and argument preferred by Dennett and Clark, of which the above is one example, just isn't equal to the requirements of the theoretical task at hand.

2.36 Conclusions:

Of the three types of arguments presented about representations in connectionist networks, none was successful in defending the connectionist against Fodor and Pylyshyn's dilemma. It turned out that those who argued that connectionist systems i) either don't have or don't need representations at all or ii) stated that those representations are no different from the classical variety, were erroneous in assigning these properties to connectionist networks. The only correct description of a network is to describe it as containing distributed representations. Distributed representations, however, are poor at representing symbols such as semantic trees, that is, symbols that have the recursive syntactic structure that is required to support systematicity, for the following reason: connectionist representations aren't discrete structures. Indeed, Smolensky's tensor product work suggests that the limitations in using tensor products to represent complex symbol structures, mathematically follow from the distributed nature of connectionist representations. The only solution is then to either i) deny systematicity as a fundamental feature of cognitive architecture, or ii) take Smolensky's route, and claim that connectionist systems are sub-symbolic. Neither seemed plausible.

Now recall that not only were there objections to Fodor and Pylyshyn that exploited the nature of representations in networks, but also objections that exploited the nature of rules in a connectionist network. These responses are now irrelevant. The reason

is that we have established that Smolensky was correct to stipulate that connectionism means that only distributed vector representations are employed by networks with useful and novel properties. Having established this, we have already automatically fixed the nature of the rules. Classical rules are defined for the well formed formulas of particular formal languages. They are *not defined* for vectors and vector spaces. It immediately follows that the rules that operate on connectionist networks are not classical rules, unless it happens, for example, that we can either 1) construct vectors and vector spaces in $\mathfrak{R}$ as wffs in a formal language, or 2) do the converse. This is an important point at this juncture because it also tells us why connectionism and classical approaches may be fundamentally incompatible, which could be a major problem for both models of cognition.

In the former case, to appreciate the difficulty of the mathematical problem we should consider that constructing a vector or vector space in $\mathfrak{R}$ with classical computational structures means using a discrete, finite structure (the lexicon or alphabet of a formal language, for example) to construct a space that constitutes a non-denumerable infinity. The operations or rules that are available to us for this task, say productions in a formal language grammar, can so far at best produce a denumerable infinity. In the latter case, there are similar problems. We have to specify a syntax for objects — vectors and vector spaces — that are everywhere dense when defined on $\mathfrak{R}$, and therefore have either an infinite number of parts, or no proper parts at all.

It is also false to say that connectionist networks have no rules, that is that no rules apply to vector representations. In fact, vectors as a general class are a set of mathematical

objects that are defined by the specific and well characterized set of rules that they obey³²!

It seems at first glance then, that Fodor and Pylyshyn win, for one of the horns of the dilemma is correct. Connectionism is indeed empirically inadequate, because in principle, connectionist networks that are not merely implementations of classical architectures do not have syntactically structured representations and are therefore incapable of doing certain kinds of things. But it is only a limited victory, and it might be a Pyrrhic victory at that. While Fodor is correct to state that as an algorithmic description of cognition, connectionism is empirically inadequate, equally, Smolensky is correct in stating that connectionism is genuinely different from a classical architecture. But classical AI is also empirically inadequate, and the genuine difference between it and connectionism might well ensure that it remains inadequate. The reason for this is that the unique properties of connectionism, such as certain kinds of pattern recognition, result from a vectorial and distributed method of representation. This has bite because Fodor and Pylyshyn's argument cuts both ways. A classical system that does not have vectorial representation won't be able to do the kinds of things that a connectionist network can do. And it does not look like a classical architecture can have vectorial representation, because the mathematical characteristics of a vector representation are not obviously implementable with discrete structures, i.e. what's inside a classical machine. This may mean that there are several features of cognition that classical architectures cannot implement, leaving the classical paradigm permanently deficient.

The common reply to this is to trot out the old result from Turing, stating that a

³²See Taylor and Mann (1973), chapter 10.

digital system (a classical machine) can implement any analog system (like a connectionist machine), if perhaps inefficiently. Indeed, when network emulator software is run on a classical computer, that's exactly what's happening. Nevertheless, it is a misinterpretation of what Turing actually said. The exact result was that any digital machine can simulate any analog process, to any required degree of accuracy, provided that it is *not* a finite state machine, that is, the Turing machine has infinite resources (Cutland, 1992, chapter 3).

What is problematic is that when an analog process is a non-linear dynamic process, it may be, and often is, chaotic. In this case, the accuracy of the required emulation must be 100%. Any error whatsoever in such a system will accumulate rapidly and drastically (Peitgen, Jurgens and Saupe, 1992), and the emulation of the vector state will therefore exhibit large deviations from an actual trajectory through the phase space.

Most of the connectionist networks that computer scientists work with are quite tiny and individual episodes have a short run time, so error isn't a problem just yet. The brain, however, is neither greatly circumscribed (there are an estimated 10^{15} synapses), nor are the 'run times' as it were, short. And incidentally, even individual neurons display chaotic behavior. Thus, the error in modelling any significant chunk of cognitive activity — what cognitive scientists want to end up being able to do — is likely to be large and empirically significant. If these errors are to be avoided then the digital/classical machine cannot be a finite state machine. The brain, and actual computers, however, *are* finite state machines. Therefore if we wish to model cognitive activities that require vectorial representation, and probably have chaotic features, then the artificial brain that we design to do this cannot just make use of a classical architecture exclusively.

Fodor's contention that classical AI is the only "president" we've got must meet two conditions in order to be correct. First it must be shown that, by itself, connectionism is inadequate as a theory of cognition. This seems to be true. Second, it must be the case that either i) by itself, the classical picture is adequate (unlikely), or, if physical symbol systems are inadequate, we must have grounds for choosing a physical symbol system over a connectionist network. The latter entails that there must be a rigid dichotomy between networks and classical AI. If we can make hybrids between connectionism and physical symbol systems — systems that can process both syntactically structure symbols and activation vectors — then there can be no grounds to choose between the theories, because we will use both. Fodor attempts to give us the grounds for choice between the two by claiming that systematicity and productivity are central to cognition. Fine. He must go further, however, and show that systematicity and productivity are *more* central, than say, the capacity to handle, for example, multiple soft constraints. Given that Fodor and Pylyshyn have no grounds for claiming that it is more important for a model of cognition to be systematic (like in language processing and production) than it is for it, say, to solve problems with deformed inputs (as in vision), there isn't any way to support his choice of theory for cognition over the connectionist alternative. What I want to do now is to suggest that Fodor and Pylyshyn aren't forced into this line of argument because, despite the arguments in Chapter 2, there isn't a rigid dichotomy between classical AI and connectionism. If that is so, then it opens up the prospects for a cognitive model that includes both connectionism and classical symbol processing, and incorporating features of both. This has the effect of making Fodor and Pylyshyn's dilemma for the connectionist

much less important, because it permits the connectionist (and the advocate of classical AI) to have a theory that is incomplete without being irrelevant.

Chapter 3: Reconciling Connectionism and Classical A.I.

3.1 Hybrid Systems:

What first suggests the possibility of a middle path between the connectionist network and the classical computer is the nature of a distributed representation. Consider the XOR network described in McClelland, Rumelhart, and Hinton (1986, vol. 1, p. 64). This network belongs to that class of networks that implement a classical architecture. The representations aren't distributed, and it is immediately apparent how to hand-wire the network to perform the XOR task. This network could be a module in a larger network, with similar features, that is designed to solve very simple logic problems. Thus there would be an XOR module, a module for modus ponens and so on, and a module to decide what kind of problem was being posed to the network. The semantics of the individual nodes in this network would be entirely interpretable and the rules by which the network operates on inputs is relatively clear. Now consider a network that develops and employs distributed representations, and that has been *trained* to do the same logic problems as the network in which the XOR network is a module. Such a network will qualify as implementing a connectionist architecture. Note that the answers that we obtain from the second network will be probabilistic in nature, and the computational mechanism that has been used to obtain them will be different. The first point to be made here is that the difference between a computational system and a classical system can, in some cases, consist in the geographical size of the representation alone, that is, whether it exists across one unit or many.

There are other differences that will serve equally well, however. Take a connectionist network that employs distributed representations to perform logical operations, like SIMPLIFIER. If this network is appropriately structured — the connections and units clustered and wired together in the right way — the network can be turned back into a classical architecture. The way to do this is to modularize the network, so that the network is using a task decomposition method with local representations in order to solve the problem. This suggests that the difference between the two networks can also be a difference in the network structure, a difference in how the network is organized to do its job.

The general question is this: When does a network count as having a distributed representational system? Does the content have to be spread out over two nodes, ten nodes...? How much superimposition of semantic information does there have to be? At what point will an appropriate structuring of the network make task decomposition possible? At what point will the distributed representations turn into (if this is even the right word to describe it) localist representations, or *vice versa*? There aren't any good answers currently available to this question, but there seem to be at least three alternatives. The first is that each category of architecture composes a crisp set. The second alternative is that the difference between the two architectures is a smoothly graded continuum, and there is no identifiable point at which one becomes the other. Third, there may be multiple coding schemes (Sun, 1995, Bechtel and Abrahamsen, 1991), where networks simultaneously possess different kinds of representations, i.e. both distributed and localist representations. Some of these coding schemes suggest that what constitutes a distributed

representation, or a localist representation may be a matter of family resemblance (c.f. Wittgenstein, 1953). If Smolensky is correct that the central difference between a classical and connectionist architecture is the style of representation employed, then this entails that architectures themselves may be grouped into family resemblances. The sheer diversity of computational architectures tends to support this conclusion also.

If the second possibility is the correct one, it suggests a way to reconcile the two architectures, a way to have something in between. It is not clear however, that this will work. The example that I have chosen, with the XOR network, involves a task that either architecture can perform, and that isn't an appropriate test. The acid test is to take a task that requires the kind of resources that only both a classical architecture and a connectionist architecture could supply. Thus we want a task that requires both the systematic capabilities of a classical architecture and the unique pattern recognition capabilities of a connectionist network. Furthermore, we must stipulate that a single network should be able to do both. The hope is that a network that is situated in the middle of a continuum of architectures will retain the capabilities of both its more traditional brethren.

There are reasons for both skepticism and hope. The reason to be skeptical is that the mathematical character of the two different kind of representations appears to be fundamentally different. One is discrete and the other is continuous. Furthermore, it is the mathematical character of the representations that lends them their respective computational properties. There are two questions that result from this. The first is: What

kind of mathematical entity sits in between the continuous and the discrete³³? The second question is: Will such a mathematical entity retain the computational properties of both a continuous and a discrete structure (if 'retain' is really the appropriate word)? The reason for hope is that the answer to both questions may be yes.

One way of reading the history of mathematics of the twentieth century is as successive attempts to reconcile the continuous with the discrete. Such efforts include, for example, attempts to construct infinite sets, and at the other end of the scale, to find set-theoretic ways of dealing with the infinitesimal. Both of these tasks are things that continuous mathematics, calculus for example, can now handle effortlessly. There are good mathematical reasons for wanting to do these kinds of things with logic and set theory, and no mathematical result to date that suggests that it is impossible³⁴.

As for the second question, it is quite clear that there can be no definite answer until someone actually goes out and constructs a mathematical object, like a non-denumerable set³⁵, that has properties of both the discrete and the continuous. Nevertheless there are mathematical constructions like fuzzy sets that do seem to sometimes behave a bit like both the discrete and the continuous, and retain interesting

33

An alternative would be to fully define, say, vector mathematics using discrete mathematics or vice versa. I suspect however, that the problems would be the same either way.

34

In fact there is a rigorous treatment of some concepts in calculus, such as limits, using logic and set theory. It presumes, however, the real numbers as a starting point — and a non-standard treatment of $\Re$ at that (see Enderton, 1972, p. 164-73).

³⁵Without, for example, running into Russell's paradox.

computational properties as a result. For instance, a fuzzy set can deliver probabilistic answers — it can say ‘maybe’, where the logic of crisp sets, for example can only return a ‘yes’ or ‘no’. Connectionist networks also have this property, this ability to deal with data and produce answers with either vagueness or uncertainty in them. I am not aware, however, even after extensive checking, that a classical computer implementing a fuzzy algorithm can perform all the same kinds of tasks that a connectionist architecture is able to do.

From the empirical perspective, there are also reasons for optimism from the connectionist literature. Hinton (1990) contains a group of papers that seek to find methods for producing true hybrid networks, in which processing of complex symbol structures such as strings and trees can be implemented with distributed representations and automated learning procedures. I have already talked about some of these findings, explicitly Pollock’s work with RAAM networks and Smolensky’s tensor products. One of the goals of this research, for example, is to try to implement LISP commands in a system with fully distributed representations. While Hinton does say that success is a long way off, it’s important to remember that none of the computer scientists actually working on the problem think that there is any *principled* reason to be skeptical about some sort of a hybrid architecture. There has been progress toward the stated goal, even if the going has been very tough. Provided that such a hybrid architecture can be produced, the consequence for Fodor and Pylyshyn’s argument is that it is just plain wrong-headed, because the effect will be to produce a model that can do everything that we want it to do. The model however, won’t have *either* a strict classical or a connectionist architecture.

Some of the features of the hybrid will clearly be implementations of classical features (e.g. LISP commands), for example, while other features will be characteristically connectionist. Chalmers (1990) even suggests that true hybrid networks may even show novel properties not possessed by a pure architecture of either type, such as his own two-stage RAAM network.

There is however, a second alternative for reconciling the two different architectures, and that is simply to use both rather than insist on the one or the other. Even if the mechanisms inside each architecture are incompatible, the inputs and outputs of each are not. A connectionist system is perfectly able to use sentence-like objects as inputs and to return sentence-like objects as outputs, provided that the input is well-formed. These are of course the kinds of inputs and outputs preferred by classical machines. Therefore one way of using both architectures is to connect them in series — use the outputs of a neural network as inputs for a classical machine and *vice versa*. There are probably other ways of joining the two as well. For instance, in a machine that uses a task decompositional architecture, any one of the given tasks could be performed by a neural network. Thus one could nest one architecture inside of the other. Most of the currently available hybrid architectures are of this type (see Sun, 1993 and Sun, 1996).

One can even imagine the kinds of advantages that would accrue from such arrangements. Connecting a connectionist machine and a classical machine in series would in effect permit the classical machine to naturally handle deformed inputs, and indeed this proves to be the case in practice (for an example, see Greska, Frank, and Geiger, 1997). This connectionist network is used for the pattern classification of sheet metal parts and

can produce a definite answer (which would have a certain probability of being incorrect of course) which is usable by the classical machine, for things like industrial control processes, which are better implemented in a classical rule based system. Now it may turn out that grafting the two architectures leaves computational holes — things that neither can do together, that a ‘pure’ architecture could have done alone, but I am neither aware of any examples of such, nor can I imagine what they might be. If there are such instances however, the search for a true hybrid architecture becomes pressing.

Ron Sun has described and tested another variation on the hybrid network theme. The CONSYDERR network, which is designed to implement deductive reasoning processes (Sun, 1995) does not use either a modular or a nested architecture, as above, rather it runs two networks in parallel. The “upper” network is an implementation of a classical system, with localist representations. The lower network is a connectionist network, with distributed representations. The two networks are connected so that they can interact. Importantly, the task of making deductive inferences is *not* divided between the two networks. The activities of both networks are, rather, fully integrated for purposes of producing solutions. CONSYDERR was shown to produce more robust results than either a classical program, or a connectionist system could produce alone. His formal analysis of the network suggests that this results because both a rule based approach (in the upper layer) and case based reasoning (in the lower layer) were simultaneously employed. CONSYDERR does, however, inherit some of the disadvantages of each approach. It was quite a bit harder to train than a standard connectionist network using a gradient descent training algorithm, for example.

3.12 One Consequence of a Hybrid System:

This kind of compatibility between the two types of machines suggests a different way in which the dilemma posed by Fodor and Pylyshyn is in error. The kind of computational structures that I have described and proposed suggest a third alternative to the dilemma: That the two architectures are distinct alternatives, and the only empirically adequate account uses both, either in different places for different tasks (e.g. a heterogeneous architecture), or uses both architectures simultaneously throughout (e.g. a homogeneous architecture like CONSYDERR). The cost of this third way is that aspects of the modularity thesis have to be given up.

The modularity thesis describes the cognitive system as containing two types of parts. The first type of component is modular. Modular components of cognition are fast, hardwired, automatic, and informationally encapsulated. Their activities aren't cognitively penetrable. Classical examples are sensory systems, like the visual and auditory system. Modules are said to have a classical computational architecture. The second component part of the modularity thesis is the central system. The central system is holistic, and isotropic. It receives information and sends output to all of the various modules. The activities of the central system are cognitively penetrable. While Fodor does not specify the architecture of the central system, a natural candidate would be a connectionist architecture, because it has the holistic and isotropic features that are desired.

A network can be said to be holistic and isotropic (as defined in Fodor, 1980) for the following reasons. First, a vectorial representation does not contain proper atomic elements. Recall that the components of a tensor product, for example are generally

approximate. As a result, the semantics of a vector are only precisely and completely describable as a whole entity. This is the lesson of Smolensky's "coffee cup" example (Smolensky, 1988). Second, semantic information cannot be identified with any particular unit of the network. This is because semantic content is superimposed in a distributed representation. Thus, unit activities may be associated with more than one semantic item. The only way to identify semantic content is to correlate it with respect to global patterns of activity. Third, the connectionist systems of interest to Smolensky generally use gradient descent algorithms, such as back propagation, for training. These algorithms work by detecting the minima of an *error landscape* (see Anderson, 1994, p. 243, for an example). The error landscape is a product of the concurrent values of *all* the weights in the network. As the error landscape is defined in the weight space of the network, one cannot consider the effect of individual weights in isolation because the topology of the error landscape is a property of the arrangement of network weights, *relative* to one another. A slope or valley or other feature in the weight space, for example, is not just a function of one point. It is also the product of the values of a cluster of nearby points, and thus has a topological context that determines whether that point is a maxima (when nearby points have smaller relative values), a minima (when nearby points have higher relative values), or a saddle point, and so on. Similarly, when networks that can be characterized either as Boltzmann machines, or those employing harmony theory, solve a problem requiring multiple constraint satisfaction, they are said to "settle in" to a local minima of the problem space (See Smolensky, 1986, or Hinton and Sejnowski, 1986 for details). Given that a problem space is also a topological construct, similar considerations

apply to these sorts of networks. Of course, it is *not* the case that all networks are holistic and/or isotropic.

What has to be given up with a mix-and-match architecture is the idea that there are two separate types of cognitive system, located in different parts of the mind. The form of hybrid architecture that I have in mind would result in some of the features of a central system appearing in input modules, and some of the features of a module appearing in cognitive capacities that are normally described as part of a central system. Empirically, this isn't a hard sell, as some of these features do appear in places where the modularity thesis neither predicts them, nor can it explain them. For example, the visual system can handle deformed input with ease. Yet this is something that a classical input module should have trouble with, because the input modules employ classical computational architectures, which aren't very good at handling deformed input. Studies of the retina further suggest that there is coarse coding in retinal processing, which is a kind of distributed representation, and therefore a feature of a connectionist system.

If Fodor is truly concerned about maintaining his architectural commitments, that modules are classically computational, atomistic processors, etc. then there are certain anomalous features about input modules that have to be addressed. Specifically, input modules display cognitive capacities that connectionist systems are better at implementing, such as their ability to handle deformed input. And Fodor *is* serious about his architectural commitments. The way a module is defined is practically a definition of a classical cognitive architecture. What is problematic for the modularity thesis is that these anomalous capabilities cannot be addressed with a classical architecture. As a

consequence, the modularity thesis is in trouble.

Similarly, there are features about central systems that suggest that they are not as holistic and isotropic in character as Fodor would suggest. Higher symbolic reasoning, for instance, is best implemented in a classical architecture. The atomistic nature of the process involved — where representations enter into the computational process as independent atoms — is contrary to the holistic nature of a central system, and is also contrary to the nature of the candidate architecture that I have suggested for such a system (i.e. connectionist). I can see no better architecture, given the description of a central system, so again, it appears the modularity thesis has a problem.

Are there any consequences of abandoning the modularity thesis for cognitive science? Of course there are. However, abandoning the modularity thesis does not mean that the holistic cognitive processes of a central system suddenly jeopardize faculty psychology, nor does it jeopardize the idea that there can be discrete cognitive modules with holistic elements. The fear I am trying to allay here is the fear that the problems that are entailed by full-blown holism in semantics, will also be entailed if the holism of the modularity thesis “escapes” the confines of the central system, and starts to look like it might be a widespread feature of cognition. For one thing, abandoning the modularity thesis might only mean keeping the modules, and *getting rid* of the holistic central systems altogether. This does not seem to be necessary. Furthermore, using a holistic element in a cognitive architecture does not mean that the entire cognitive architecture suddenly becomes holistic, or even that the whole module becomes holistic, because that holistic element can be placed inside a module, or sub-modular element. There is no reason why

holistic networks in a faculty or module cannot be informationally encapsulated (c.f. Fodor, 1980), which prevents the holism from being so pervasive that it is problematic. Holistic/connectionist elements in modules only entail that some of the processes inside the module will be holistic. Encapsulating a module limits the holism of the connectionist element of the module to the cognitive processes that the holistic connectionist network implements in the module itself.

Nor does abandoning the modularity thesis make a global, unified, central system any less plausible, because abandoning the modularity thesis might mean mixing and matching the features of modules and central systems, rather than abandoning either central systems, or modules. That is, proposing a cognitive architecture where the construction of cognitive systems, central or otherwise, may include both holistic/connectionist elements, as well as atomistic/classic elements, when a mix of such capabilities seems empirically required. This is the position that I want to adopt. In any event, the only way to *conceptually* bring a global system under attack (i.e. independent of empirical evidence to suggest that there isn't any such thing) is by showing that the idea of a central system is in logical contradiction with the proposed features of the cognitive architecture. One way of demonstrating such a logical contradiction would be by arguing for a pure cognitive architecture of the classical sort. This is because such an architecture accomplishes cognitive tasks by functionally decomposing problems into smaller and easier tasks that are independently handled by particular processing elements. This kind of *homuncular* functionalism, which is characteristic of classical architectures, and is entailed by them, is not compatible with the unified character of a holistic global system, where

cooperative computation and functional integration are the order of the day. Indeed, task decomposition is contradictory to a properly holistic system. Provided that there are connectionist elements in the mind/brain, implementing some sort of a global representation ought not be conceptually problematic, as in principle, the holistic computational resources are available to do the job. What does have to be given up is the idea that modules and central systems have a certain kind of exclusive computational character. Provided that we are not committed to the idea that psychological faculties and/or cognitive modules must be a of a particular and pure architectural type, i.e. that they are all wholly atomistic/classically computational, or completely holistic/connectionist, then it is possible to argue for cognitive faculties that can contain both kinds of architectures.

3.3 Cognitive Architecture and Semantics:

The purpose of this last chapter is to describe what I think the debate between connectionists and defenders of the classical approach to cognition means for semantics. The foundational claim for what follows is that semantics is supervenient on cognitive architecture. That is, that semantics is a description of what takes place at the cognitive level. Cognitive architectures are of course descriptions at the algorithmic level. The first consequence of this is that any complete account of cognitive architecture must explain how semantic systems are implemented. The second consequence is that the structure of cognitive architecture limits theory choice for semantics, because cognitive architecture is the subvenient property.

It is often stated in cognitive science that the multiple realizability argument means

that Marr's levels of description show a high degree of independence from one another. Multiple realizability, for example, suggests that talk about an implementational level feature does not necessarily have any entailments for an algorithmic feature. This isn't necessarily true. The notion of a cognitive level avoids being reductionistic while hanging on to a physicalist description of the mind by appealing to emergent properties/supervenience. Multiple realizability or no, supervenience does not mean that 'anything goes'. What it does mean is that a *class* of implementational or algorithmic strategies can be used to explain/implement algorithmic or cognitive level explanations, respectively. So what is true is that many characteristics of the subvenience base can be changed without altering the supervenient properties, not all of the characteristics can be. Membership in a class is defined by the possession of certain properties and characteristics. If these definitional properties and characteristics of the class are changed, then the supervenient properties implemented by the class change also.

We have seen in the discussion of representationalism and cognitive architecture that commitments at the algorithmic level have consequences for the kinds of cognitive properties that can be implemented. Cognitive features such as pattern recognition require a certain style of representation, in this example a distributed representation. And representations are features of the algorithmic level of explanation. What I wish to argue next is that algorithmic explanations have consequences for semantic styles of explanation.

The central claim of the argument is that holistic representations, like those found in connectionist systems, entail a holistic semantics. Similarly, the atomistic representations (of classical architectures) result in an atomistic semantics. The reasons are

evident. First, as we have seen³⁶, there is no real sense in which a distributed representation can be said to have precisely defined syntactic parts. Vectors have components, but as Smolensky shows, most of the time tensor products can only be incompletely decomposed into their components. Save for these components, the vector as a whole can be thought of as having either an infinite number of parts, or no parts at all. This entails that the semantics won't have precisely identifiable parts either, because there is no structure in the symbol to delineate the parts. In both cases, the parts are only derived, via some kind of context principle, with respect to the whole. This was one of the defining features of holism, from chapter I. Similarly, the syntax of a classical machine is atomistic, systematic, recursive, and combinatorial. We would then expect the semantics to express these features as well. Again, the symbol structure has a causal effect on the supervenient semantic level. Second, the capabilities of a cognitive architecture, at least in part, result from the style of representation that is employed. If a pure classical architecture could implement a holistic style of representation, or a pure connectionist architecture could implement atomistic representations, then there would *be no differences* in the observed cognitive capabilities that they could implement or model. In short, there would be no issue between connectionists and classical cognitivists. But there *are* differences between the kinds of things that the two architectures can do, and more, those differences have turned out to be quite resistant to efforts to overcome them.

36

For example, from trouble with isolating vector components that represent 'coffee' in 'cup of coffee' and also because of the possibility of one piece of syntax meaning more than one thing (semantic superposition)

The next step is to argue that connectionist representations are indeed holistic and that classical representations are likewise atomistic. Let us begin with distributed representations. The fundamental character of a holistic semantics, I have argued, is that it is difficult in principle to separate off bits of semantics from the language as a unit. Thus it is difficult (and if Quine is right, impossible) to describe the meaning of one word without simultaneously describing the meaning of others. The syntax of a distributed representation has precisely this character. This is argued by both van Gelder (1990) and by Smolensky (1988) at some length. In Smolensky's "cup of coffee" example, it isn't possible to state what part of the vector is representing "cup". The approximate best that can be done is some kind of a subtractive contrast with vectors that represent "cup of hot liquid". Another way of stating this is to state that the only unit of meaning in a language is the language as a whole. An activation vector also has this character. Connectionist architectures possess a single activation vector — a single representation — and while this representation can be carved up in the manner Smolensky suggests it does not have a combinatorial recursive semantics.

An atomistic semantics has the character that symbols enter into semantic relations as independent bits. The bits are then combined according to a recursive syntax that produces more complex symbols. The atoms and their arrangement determine the meaning of larger semantic units in a combinatorial fashion. This is precisely the same arrangement that is seen in a classical architecture. The mathematics are the same too. Both systems, semantic and cognitive, are employing the combinatorial mathematics of discrete structures and set theory.

What I have attempted to do in the latter section of this chapter is to show that there are strategies for peaceful coexistence between defenders of classical AI, and defenders of connectionism. Specifically I suggest that, even at this stage of our knowledge, it is possible to link connectionist architectures and classical architectures in at least two useful ways. The consequence that this has for semantics is that it means that a cognitive architecture can implement semantic systems that are holistic (connectionist) on the one hand, and atomistic (classical) on the other, at the same time. It means that architectures can be designed that have both.

So far I have spent some time demonstrating that there is a real dichotomy between two styles of semantic explanation, and then I have used the connectionist/classical debate to argue that this isn't a problem. The first discussion takes place at the cognitive level of description. The second shows that at the algorithmic level of description. Who wins? The answer is that the argument at the algorithmic level wins. Because semantic properties are supervenient properties (they have to be implemented somehow!), any claim that the styles of semantic explanation are incompatible must ultimately demonstrate why compatible implementations are impossible. Furthermore, as a general rule when dealing with supervenient properties, evidence and arguments at the subvenient levels must take priority over evidence and arguments at the supervenient level. If someone were to conclusively show that there is no possible way, for example, to implement a classical architecture on a biological neural network, that would mean that the classical model would be wrong. And it really does not matter if classical architectures can be used to explain every known cognitive level description there is. If it cannot be

implemented in the brain, then it isn't what we use. To hold otherwise would be to make a complete mockery of the notion of supervenient properties, where the point is (in cognitive science) to preserve physicalist commitments without being a full-blown reductionist. The only way to preserve that physicalist commitment in the imaginary example that I have given is to throw the classical architecture out as a model.

One way to get around this would be to first claim that there are different kinds of classical architectures and different kinds of connectionist architectures, and that only some of these kinds are compatible. Second: if there are some incompatible sets of architectures then those architectures possessed by humans may be among them. We can then conclude that it is perfectly possible for incompatibilities to exist at the cognitive level even where there is no principled disagreement at the algorithmic level. That is, appeal to multiple realizability. This will not do. The reason it will not do is because there is an exception to this argument, and the exception applies.

What makes classical architectures and connectionist systems compatible is that they can use each others' inputs and outputs. That is to say that representational information from one system can be used by the other kind of system. Thus, there is no fundamental incompatibility with the information contained in the two styles of representation. The style of representation, however, is a definitional feature of the class — all of the members of the set possess it. The consequence is that there can be no system, classical or connectionist, that isn't compatible.

One could attempt to argue that even where the representational styles are compatible, the content isn't. That is to say, the content of one kind of representation

couldn't be carried by the other kind of representation. This however, is a crazy argument. One of the definitional properties of classical architectures is that they are formal systems. It does not matter what the content is, the syntactic engine will not care. As there are already numerous examples of connectionist networks that can use symbol structures as inputs (and even though they cannot use the syntactic information in processing, they can preserve it), so the objection just does not apply.

There is also a second series of objections that cannot be ruled out on the basis of this argument. The consequences of these, however, are so unpalatable, particularly to people like Fodor and Lepore, that very few would be willing to swallow them. The first of these is to object to the notion of cognitive levels. This is problematic for the following reasons. Marr's taxonomy is little more than the claim that cognitive science deals with supervenient properties. Now the details of Marr's scheme could be wrong, but not in any way that would be pleasing to most cognitive scientists. There could be for example, one less level to Marr's scheme. The level that would have to go is the algorithmic level, because it's the only one that's optional. If we say that there is no cognitive level, that's tantamount to saying that there is no such thing as semantic properties, which would make the semantic debates a moot point. The implementational level is of course not optional unless someone is seriously proposing substance dualism as an alternative. The consequence of eliminating the algorithmic level, however, is to state that there is no such thing as cognitive science. Indeed, Fodor and Pylyshyn *use* this in their argument against Smolensky: either his proposal is an inadequate algorithm, or it is merely an implementation, and therefore irrelevant to cognitive science.

Alternatively, one could add levels to Marr's scheme. Smolensky attempts to do this in fact. The trouble is that the attempt does not work without an explanation of how higher level properties emerge from the extra level, and there is no explanation, nor any evidence, in the offing. In Marr's scheme it is the computer as a physical metaphor that is doing this work. The last alternative is to *change* the levels. Given that there is no doubt about the bottom, implementational level — the brain — and we wish to preserve cognitive level explanations in order to preserve semantics (for purposes of this discussion alone, if nothing else), the only level that can be changed is again the algorithmic level. This is equivalent to saying that both connectionism and classical AI are implausible. Perhaps they are, but given that there is at present no satisfactory theoretical alternative, most cognitive scientists, and certainly Fodor and Lepore, will be understandably reluctant to endorse this option.

Outside of meddling with Marr's schema, there is one other alternative to be raised. What one could do is argue for full-blown reductionism. That is, take the position of the eliminative materialist. The trouble with this of course, is that the only explanations likely to survive are the neuroscientific ones. At best, semantics will be relegated to the category of a convenient shorthand, rather than an autonomous discipline. If both semantics and the contemporary practice of cognitive science are to be preserved, then, there can be no incompatibility between atomism and holism in semantics.

Bibliography:

- Adams, F., and Aizawa, K. (1994) "Fodorian Semantics" in Stich and Warfield (1994), pp. 23-242.
- Aizawa, K. (1994) "Representation without Rules, Connectionism and the Syntactic Argument" *Synthese* 101: 465-492.
- Anderson, J.A. (1995) *An Introduction to Neural Networks*, MIT Press.
- Arbib, M. (ed.)(1995) *The Handbook of Brain Theory and Neural Networks*, MIT Press.
- Atkins, P.W. (1986) *Statistical Thermodynamics, Third Edition*, Freeman Press.
- Bechtel, W. (1994) "Natural Deduction in Connectionist Systems" *Synthese* 101: 433-63.
- Bechtel, W. and Abrahamsen, A. (1991) *Connectionism and the Mind*, Blackwell Press.
- Beer, R.D. (1990) *Intelligence as Adaptive Behaviour*, Academic Press.
- Berkeley, I.S.N. et al. (1995) "Density Plots of Hidden Value Unit Activations Reveal Interpretable Bands" *Connection Science* 7: 167-186
- Block, N (1986) "Advertisement for a Semantics for Psychology" in Stich and Warfield (1994).
- Block (1993) "Holism, Analyticity and Inferential Role" *Mind and Language* 8: 1-26.
- Block (1995) "An Argument for Holism" *Proc. Arist. Soc.* pp. 151-69.
- Boghossian, P. (1994) "Inferential Role Semantics and the Analytic/Synthetic Distinction" *Philosophical Studies* 73: 109-122.
- Brooks, R.A. (1991) "Small Autonomous Robots: Sensing and Action" *Proceedings: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 452.
- Callan, R. and Palmer-Brown, D. (1997) "(S)RAAM: An Analytical Technique for Fast and Reliable Derivation of Connectionist Symbol Structure Representations" *Connection Science* 9: 139-59.

- Carnap, R. (1947) *Meaning and Necessity*, University of Chicago Press.
- Chalmers, D. (1990) "Syntactic Transformations on Distributed Representations" *Connection Science* 2: 53-62.
- Churchland, P.S. and Sejnowski, T.J. (1992) *The Computational Brain*, MIT Press.
- Clark, A. (1991) "Systematicity, Structure Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn" in Horgan and Tienson (1991), p.198-218.
- Clark, A. (1993) *Microcognition*, MIT Press.
- Clark, A. and Toribio, J. (1994) "Doing Without Representing?" *Synthese* 101: 401-431.
- Cutland, N.J. (1992) *Computability: An Introduction to Recursive Function Theory*, Cambridge University Press.
- Davidson, D. (1967) "Truth and Meaning" in Martinich (1990) pp. 79-90.
- Davidson, D. (1976) "Reply to Foster" in Evans and McDowell (1976).
- Davidson, D. (1980) *Essays on Action and Events*, Oxford University Press.
- Davidson, D. (1984) "Radical Interpretation" in *Inquiries into Truth and Interpretation*, Clarendon Press.
- Devitt, M. (1996) *Coming to our Senses: A Naturalistic Program for Semantic Localism*, Cambridge University Press.
- Donnellan, K. (1966) "Reference and Definite Descriptions" in Martinich (1990).
- Dretske, F. (1981) *Knowledge and the Flow of Information*, MIT Press.
- Dretske, F. (1995) *Naturalizing the Mind*, MIT Press.
- Enderton, H.B. (1972) *A Mathematical Introduction to Logic*, Academic Press.
- Evans, G. (1982) *The Varieties of Reference*, Oxford University Press.
- Evans, G. and McDowell, J. (eds.) (1976) *Truth and Meaning: Essays in Semantics*, Clarendon press.

- Field, H. (1977) "Logic, Meaning and Conceptual Role" *Journal of Philosophy* 74: 379-408.
- Fodor, J.A. (1975) *The Language of Thought*, Harvard University Press.
- Fodor, J.A. (1983) *Modularity of Mind*, MIT Press.
- Fodor, J.A. (1987) *Psychosemantics*, MIT Press.
- Fodor, J.A. (1994) *The Elm and the Expert*, MIT Press.
- Fodor, J.A. and Lepore, E. (1992) *Holism: A Shopper's Guide*, Blackwell Press.
- Fodor, J.A. and Lepore, E. (eds.) (1993) "Holism: A Consumer Update", *Grazer Philosophische Studien* 46
- Fodor, J.A. and Lepore, E. (1993b) "Reply to Critics" *Philosophy and Phenomenological Research* 53: 673-682.
- Fodor, J.A. and Pylyshyn, Z. (1990) "Connectionism and Cognitive Architecture: A Critical Analysis" in MacDonald and MacDonald (1995): Ch 2.
- Frank, M. and Knight, T (1997) "Ultimate Theoretical Models of Nanocomputers" presented at the Fifth Foresight Conference on Nanotechnology, available online at <http://www.ai.mit.edu/~mpf/Nano97/abstract.html>
- Frege, G. (1892) "On Sense and Nominatum" in Martinich (1990).
- Frege, G. (1918) "The Thought: A Logical Inquiry" in Harnish (1994), pp. 517-535.
- Geirsson, H. and Losonsky, M. (1996) *Readings in Language and Mind*, Blackwell Press.
- Gibson, R.F. (1982) *The Philosophy of W.V. Quine: An Expository Essay*, University Press of Florida.
- Greska, W., Franke, V. and Geiger, M. (1997) "Classification Problems in Manufacturing Sheet Metal Parts" *Computers in Industry* 33: 17-30.
- Hanson, S.J. and Burr, D.J. (1990) "What Connectionist Models Learn: Learning and Representation in Connectionist Networks" *Behavioral and Brain Sciences* 13: 471-518.

- Harnish, R.M. (ed.)(1994) ***Basic Topics in the Philosophy of Language***, Prentice Hall.
- Harrell, M. (1996) "Confirmation Holism and Semantic Holism" ***Synthese* 109**: 63-101.
- Haugeland, J. (1991) "Representational Genera" in Ramsey, Stich and Rumelhart (1991), p. 61-90.
- Heal, J. (1994) "Semantic Holism: Still a Good Buy", ***Proc. Arist. Soc.***, pp. 325-339.
- Heiligenburg, W. (1991) ***The Neural Networks of Electric Fish***, MIT Press.
- Hempel, C. (1950) "Empiricist Criteria of Cognitive Significance: Problems and Changes" in Martinich (1990), pp. 13-25.
- Hertz, J. (1995) "Computing with Attractors" in Arbib (1995) p. 230-34.
- Hinton, G. (ed.)(1990) "Connectionist Symbol Processing" ***Artificial Intelligence*** vol. 46, MIT Press.
- Hinton, G. and Sejnowski, T. (1986) "Learning and Re-Learning in Boltzmann Machines" in McClelland, Rumelhart and Hinton (1986): ch. 6.
- Horgan, T. (1997) "Connectionism and the Philosophical Foundations of Cognitive Science" ***Metaphilosophy* 28**: 1-30.
- Horgan, T. and Tienson, J. (eds.)(1991) ***Connectionism and the Philosophy of Mind***, Kluwer Academic Publishers.
- Hubel, D.H. (1988) ***Eye, Brain and Vision***, New York: Scientific American Library.
- Kim, J. (1993) ***Supervenience and Mind***, Cambridge University Press
- Kripke, S. (1972) ***Naming and Necessity***, Harvard University Press.
- MacDonald, C. and MacDonald, G. (1995) ***Connectionism: Debates on Psychological Explanation***, Blackwell Press.
- Marr, D. (1982) ***Vision***, Freeman Press.
- Martinich, A.P. (ed.) (1990) ***The Philosophy of Language***, second edition., Oxford University press.
- McClelland, J.; Rumelhart, D. and Hinton, G. (1986) ***Parallel Distributed Processing***,

Vol. I and II, MIT Press.

Millikan, R. (1984) *Language, Thought, and Other Biological Categories*, MIT Press.

Millikan, R. (1994) "Biosemantics" in Stich and Warfield (1994).

Minsky, M. and Papert, S. (1988) *Perceptrons*, MIT Press.

Neale, S. (1990) *Descriptions*, MIT Press.

Newell, A. (1982) "The Knowledge Level" *Artificial Intelligence* 18: 87-127.

Pagin, P. (1997) "Is Compositionality Compatible with Holism?" *Mind and Language* 12: 11-33.

Peitgen, H.O., Jurgens, H., and Saupe, D. (1992) *Chaos and Fractals*, Springer Verlag.

Pinker, S. (1994) *The Language Instinct*, Morrow Press.

Pinker, S. and Prince, A. (1988) "On Language and Connectionism: Analysis of a PDP Model of Language Acquisition" *Cognition* 28: 73-193.

Pollock, J.B. (1990) "Recursive Distributed Representations" *Artificial Intelligence* 46: 77-105.

Pylyshyn, Z. (1984) *Computation and Cognition: Towards a Foundation for Cognitive Science*, MIT Press.

Putnam, H. (1976) "The Meaning of 'Meaning'" in Geirsson and Losonsky (1996), pp. 157-198.

Quine, W.V. (1951) "Two Dogmas of Empiricism" in Martinich (1990), pp. 26-39.

Quine, W.V. (1960) *Word and Object*, MIT Press.

Ramsay, W. (1997) "Do Connectionist Representations Earn Their Explanatory Keep?" *Mind and Language* 12: 34-66.

Ramsey, W., Stich, S. and Rumelhart, D. (1991) *Philosophy and Connectionist Theory*, Laurence Erlbaum and Associates.

Rips, L. (1983) "Cognitive Processes in Propositional Reasoning" *Psychological Review* 90: 38-71

- Rips, L. (1994) *The Psychology of Proof*, MIT Press
- Russell, B. (1908) "On Denoting", in Martinich (1990), pp. 203-211.
- Russell, B. (1919) "Descriptions", in Martinich (1990), pp. 212-218.
- Russell, B. (1985) *The Philosophy of Logical Atomism*, Open Court Press.
- Smith, E.; Langston, C. and Nisbett, R. (1992) "The Case for Rules in Reasoning" *Cognitive Science* 16: 1-40.
- Smolensky, P. (1986) "Information Processing in Dynamical Systems: Foundations of Harmony Theory" in McClelland, Rumelhart, and Hinton (1986): ch. 5.
- Smolensky, P. (1988) "The Proper Treatment of Connectionism" in MacDonald and MacDonald, (1995), p. 28-89.
- Smolensky, P. (1990) "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems" *Artificial Intelligence* 46: 159-216.
- Smolensky, P. (1995) "A Reply: Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Architecture" in Macdonald and MacDonald (1995): Ch 6.
- Stainton, R. (1996) *Philosophical Perspectives on Language*, Broadview Press.
- Stich, S. (1993) *The Fragmentation of Reason*, MIT Press.
- Stich, S. and Warfield, T. (1994) *Mental Representation: A Reader*, Blackwell Press.
- Sun, R. (1993) "Integrating Neural and Symbolic Processes" *Connection Science* 5: 203-204.
- Sun, R. (1995) "Robust Reasoning: Integrating Rule-Based and Similarity based Reasoning" *Artificial Intelligence* 75: 241-95.
- Taylor, A.E., and Mann, R.W. (1983) *Advanced Calculus*, 3rd ed., John Wiley and Sons
- Touretsky, D.S. (1990) "BoltzCONS: Dynamic Symbol Structures in a Connectionist Network" *Artificial Intelligence* 46: 5-46.

- Touretsky, D.S. (1995) "Connectionist and Symbolic Representations" in Arbib (1995), p. 243-46.
- Tversky, A. and Kahneman, D. (1983) "Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement" *Psychological Review* **90**
- van Gelder, T. (1991) "What is the 'D' in PDP?" in Ramsey, Stich and Rumelhart (eds.) (1991), p. 33-60.
- van Gelder, T. (1990) "Defining 'Distributed Representation'" *Connection Science* **4**: 175-179
- Waskan, J. and Bechtel, W. (1997) "Directions in Connectionist Research: Tractable Computations Without Syntactically Structured Representations" *Metaphilosophy* **28**: 31-62.