CLASSICAL AND CONNECTIONIST MODELS OF COGNITION

by

ERIC PAUL LORMAND

B.A., Phil./Cog. Sci., B.S., Math.
Tulane University
(1986)

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of
the requirements of the degree of
Doctor of Philosophy in Philosophy

at the

Massachusetts Institute of Technology

June 1990

Signature of Author_____
Department of Philosophy
April 27, 1990

Certified by_____
Ned Block
Professor, Philosophy
Thesis Supervisor

Accepted by_____
George Boolos
Chairman, Committee on Graduate Students

CLASSICAL AND CONNECTIONIST MODELS

OF COGNITION

by

ERIC PAUL LORMAND

Submitted to the Department of Linguistics and Philosophy
on April 27, 1990 in partial fulfillment of
the requirements of the degree of
Doctor of Philosophy in Philosophy

ABSTRACT

My thesis is about the hypothesis that human cognitive processes employ a language of thought--a system of mental representation which supports syntactically complex mental symbols, physically realized in brains. First, I offer a formulation of this hypothesis which eliminates various obscurities in standard formulations. With this in hand, I suggest directions for empirical comparisons of classical language-of-thought models (including many connectionist models) and those connectionist models which do not implement a language of thought. I argue that nonclassical connectionist models are unlikely to succeed as general accounts of cognition, but that they have promise as part of an account of the (alleged) inferential processes guiding skillful activity, which are unconscious, rapid, and holistically sensitive to a vast range of potentially relevant conditions. I show how representations in nonclassical connectionist models, despite having no syntactic or semantic structure, can realize genuinely propositional attitudes (and can therefore undergo genuinely inferential processes). Finally, I argue that classical models can themselves be applied to holistically sensitive inference, in the face of various objections which philosophers have advanced under the name of the "frame problem".

Thesis Supervisor: Ned Block

Title: Professor of Philosophy

PHILOSOPHY

(With apologies to Lewis Carroll)

'Twas brilliant, and the flighty tomes
  Did guile and quibble in the libe;
All whimsy were the conundromes,
  And the morass did gibe.

"Beware the Philosoph, my son!
  The jaws that jaw, the claims that grip!
Beware the Journal bird, and shun
  The Voluminous Manuscript!"

He took his verbal sword in hand:
  Long time the grue-some foe he sought--
So rested he by the Theorem tree,
  And stood awhile in thought.

And, as in published thought he stood,
  The Philosoph, with eyes of flame,
Came scoffing through the cogent word,
  And babbled as it came!

One, two! One, two! And through and through
  The verbal blade went snicker-snack!
He left it dead, and with its head
  He went commencing back.

"And hast though slain the Philosoph?
  Come to my arms, my thesis boy!
O factious day! Callooh! Callay!"
  He chortled in his joy.

'Twas brilliant, and the flighty tomes
  Did guile and quibble in the libe;
All whimsy were the conundromes,
  And the morass did gibe.

CONTENTS

THANKS!

. . . to my thesis advisors, Ned Block, Jim Higginbotham, and Bob Stalnaker, for providing substantive comments in good cheer, while under serious time constraints.

Chapter 0

CLASSICAL AND CONNECTIONIST MODELS

Much of the philosophical interest of cognitive science stems from its potential relevance to the mind/body problem. The mind/body problem concerns whether both mental and physical phenomena exist, and if so, whether they are distinct. In this chapter I want to portray the classical and connectionist frameworks in cognitive science as potential sources of evidence for or against a particular strategy for solving the mind/body problem. It is not my aim to offer a full assessment of these two frameworks in this capacity. Instead, in this thesis I will deal with three philosophical issues which are (at best) preliminaries to such an assessment: issues about the syntax, the semantics, and the processing of the mental representations countenanced by classical and connectionist models. I will characterize these three issues in more detail at the end of the chapter.

0.1 The Classical Framework

0.1.1 Functionalism and Token Physicalism

From a highly abstract but useful perspective, cognitive science is a kind of test for a particular philosophical theory of mental phenomena, namely, functionalism (see the related articles in Block, 1980). While functionalism can be cast as a general claim about all mental phenomena, it is most usefully introduced as an account of particular types of mental states or events. What is it to be a functionalist about a type of mental occurrence[1] *M*, such as thinking that sugar is white, wishing that one's teeth were as white as sugar, planning to eat more sugar, having a toothache, or being in a bad mood? A functionalist identifies *M* with a particular functional state or event type. Functional types are those which are individuated solely by considerations of causal relations. A state or event token *f* belongs to a functional type *F* if and only if *f* participates in causal relations of the sort which define *F*; in this case we may say

---

[1]I will usually use the term "occurrence" as a general term for anything that occurs or happens, including events and states (and also facts, situations, conditions, etc.). From the standpoint of functionalism, the interesting feature common to all occurrences is that they may enter into causal relations. I mean to abstract away from more detailed claims about the individuation of these phenomena, e.g., that a token event of stabbing Caesar can be identical to a token event of killing Caesar, or that a token state of stabbing Caesar cannot be identical to a token state of killing Caesar (for discussion, see Davidson, 1969).

that *f* has an "*F*-role". This means that a functionalist about a mental state or event *M* must specify the causal relations--the *M*-role--taken to be necessary and sufficient for a state or event token *m* to be of type *M*. This is standardly done by specifying causal generalizations or laws which relate tokens of type *M* (under certain conditions) to external occurrences (e.g., through sensory or motor processes), and to other internal occurrences (e.g., through inferential or computational processes). Functionalists typically hope that as cognitive science develops, they will be better able to specify these generalizations and laws, and so will be able to specify in greater detail which functional types are identical with which mental types.

How can this be worked into a solution to the mind/body problem (at least for those mental occurrences to which functionalism is applied)? The standard hypothesis is that there are physical state or event tokens which satisfy the functionalist's causal generalizations and laws, and are therefore tokens of mental types.[2] Given this hypothesis, then, it is possible to adopt the position known as "token physicalism" (about *M*): each token mental state or event *m* of

[2]Unless otherwise mentioned, I follow the standard practice of construing "physical" broadly, to cover not only the entities mentioned in the theories of physics, but also any natural (as opposed to supernatural), nonmental, spatiotemporal phenomena.

type *M* is identical to a token physical state or event *b*.[3] The token physicalist answer to the mind/body problem, then, is that both mental and physical phenomena exist, but that (at least in certain cases) they are not distinct.

The functionalist and token physicalist approach to the mind/body problem can only work if token physical states or events can enter into the causal relations specified by functionalism. This raises what might be called (not the mind/body problem but) the "mental-body problem": how is it possible for a physical system to implement the requisite processes of sensation, motor control, inference, memory, and so on?[4] To answer this question, we need to formulate specific claims about the nature of the token physical occurrences located by functionalism (e.g., *b* in the previous paragraph).

---

[3]This position is distinguishable from (though compatible with) "type physicalism", which identifies *M* with some physical state or event type *B* to which all tokens of *M* belong. For discussion, see the related articles in Block, 1980.

[4]This should not be confused with the problem of "interaction": how is it possible for physical states or events to enter into causal relations with mental states or events? Interaction is only a problem for theories which, unlike token physicalism, treat token mental occurrences as nonphysical. In short, the question is not about how it is possible for a body to interact with the mental, but is instead about how it is possible for a body to *be* mental. Hence the name "mental-body problem" (with a dash) rather than "mind/body problem" (with a slash). It bears emphasis that the mental-body problem must be solved in order for functionalism and token-physicalism to work as a solution to the mind/body problem.

A useful way to highlight issues about mental representations is to focus only on functionalist theories of "propositional attitudes"--certain mental states, such as thoughts, wishes, and plans, which have propositional content.[5] It seems likely from a functionalist standpoint that some inferential relations are components of propositional-attitude-roles (although it is notoriously difficult to specify these relations). To explain how physical states can implement these inferential processes, cognitive scientists have traditionally appealed to an analogy with the ways computers implement algorithms. Several philosophers have tried to extract from this analogy specific, substantive claims about the physical embodiments of propositional attitude types. The two most important claims for my purposes are representationalism and the language-of-thought (LOT) hypothesis, as formulated in a number of works by Jerry Fodor. These claims characterize what is common to classical models in cognitive science.

0.1.2 Representationalism

It is an intuitively obvious but theoretically striking fact that a person can have many different attitudes with the

[5] I do not make the assumption that all propositional attitudes are familiar to common sense, since scientific psychology may well discover new sorts.

same propositional content. We can perceive that toothpaste is white, believe (to various degrees) that it is, and hypothesize, imagine, or desire (to various degrees) that it is. Indeed, it seems that we never find people who are able to have one of these attitudes toward a given proposition, but are unable to have another one of these attitudes toward the same content. Furthermore, there seem to be widespread, regular, and important causal relations between different attitudes to the same proposition: for example, perceiving *that p* tends to cause believing *that p*, while doubting *that p* tends to interact with wishing *that p* to generate action. Whatever the ultimate details, these causal and presuppositional relations are likely factors in any functionalist account of propositional attitudes. We can therefore formulate a special case of the mental-body problem: how is it possible for these special relations to be implemented in a physical system?[6] This question is raised, but not answered, by functionalism and token physicalism about propositional attitudes. Representationalism seeks to supply an answer to this question (among others--see Fodor, 1981).

Understandably, representationalism is sometimes simply put as the claim that there are mental representations--entities

---

[6]It matters that the attitude-types listed above are such that we can have a token of any one type without also having a token of another type. By comparison, presumably there is no mystery as to how a physical system which can know *that p* can also manage to believe *that p*.

with content mediating between sensation and motor control--and that these representations are (by token physicalism) physical. However, this formulation does not reveal any claim stronger than token physicalism about propositional attitudes. Any such token physicalism postulates physical states which, as token mental states, mediate between sensation and motor control, and which, as token attitudes with propositional content, are representations. If representationalism is to yield an explanation of how physical systems can implement special relations among attitudes toward the same content, we need a construal which specifies in more detail the nature of token propositional attitudes.

The standard strategy is to treat propositional attitudes as *computational relations* between thinkers and physical representations. It will help to formulate this idea if we consider an arbitrarily chosen propositional-attitude type, say, the attitude *A* with the content *that p*. Also, let *t* be a variable ranging over the thinkers (i.e., potential *A*-ers *that p*) to which representationalism is applied (i.e., representationalism about *A*-ing *that p*). The claim (to be modified shortly) is as follows:

> (At least typically) *t*'s *A*-ing *that p* is identical with *t*'s having a certain computational relation to a physical representation *r* with the content *that p*.

Since (as explained in the previous paragraph) token physicalism, even without representationalism, "already"

postulates physical representations *that p*, any extra force of this claim must stem from its appeal to "a certain computational relation".

Unfortunately, the notion of a computational relation is typically left unclear, and the specification of which relations are appropriate for which attitudes is typically left to the development of cognitive science. This leads back to the worry that the standard formulation of representationalism fails to add any substance to functionalism and token physicalism (about *A*-ing *that p*). By functionalism, any token occurrence of *A*-ing *that p* has a particular inferential role. It follows from this that any such occurrence stands in *some* computational relation to its thinker *t*.[7] Without specific constraints, then, *t*'s *A*-ing *that p* satisfies the requirements for *r*. As a result, the current formulation of representationalism fails to require

---

[7]This inference depends on certain assumptions about what it is for a thinker to stand in a computational relation to a representation, but these assumptions are natural and (as far as I know) never called into question by representationalists. To a near-enough approximation, we may understand computation as a process of using representations as premises or conclusions of inferential processes. In this way, it appears, we can at least make sense of a representation's bearing a computational relation to *other representations*--namely, those for which it is used as a (co)premise or (co)conclusion. By a slight extension, we may imagine that representations also stand in computational relations to the inferential *processes* which act upon them. But what is it for a representation to stand in a computational relation to a *thinker*? The only natural account I can think of is that a thinker has a computational relation to a representation if and only if that representation has a role in the thinker's inferences. Given this, the inference in the text is correct.

the posculation of any representations other than those "already" postulated by functionalism and token physicalism.

To strengthen representationalism, we might simply add to the formulation the clause that *r* is distinct from *t*'s *A*-ing *that p*. However, this is compatible with *r*'s being identical to some *other* propositional attitude, and so this revision would not insure that representationalism is stronger than functionalism and token physicalism about propositional attitudes in general. For this reason, I am inclined to take representationalism as claiming that *r* is not a propositional attitude at all. This may be puzzling at first, since *r* is supposed to be a mental representation *that p*. How can something be a mental representation *that p* without being a propositional attitude *that p*? So far as I know, there is only one way for this to happen: *r* must be a propositional *idea*--an idea with the content *that p*.[8] So

---

[8]Strictly speaking, *r* can be the having-of-an-idea *that p*, i.e., a certain kind of representational *state* or *event* with the content *that p*. I will use "idea" in a broad sense, namely, as being neutral with respect to the object/occurrence distinction. It is sometimes suggested that representationalism requires representations to be ordinary objects (things that exist but do not *happen*) rather than occurrences (things that do happen). For reasons given below, I want to reject this proposed requirement. Although "idea" seems to be the best word for expressing representationalism, it may misleadingly bias one towards the requirement. While I admit that it seems unnatural to say that ideas *happen*, I am inclined to treat this as having minuscule ontological import. To take an analogous case, it seems equally unnatural to say that *representations* happen, although it *is* proper to say that representational occurrences (e.g., an assertion that sugar is white) are representations.

The proposed requirement of representational objects has other, nongrammatical sources of support. First, it would be sufficient to distinguish representationalism from token physicalism, since token

representationalism comes to the claim that propositional attitudes *that p* are typically computational relations to propositional ideas *that p*. It is incumbent upon the representationalist, then, to give an account of the difference between propositional ideas and propositional attitudes.

Having an idea *that p* is akin (i.e., presumably identical) to what philosophers sometimes call "grasping" the proposition *that p*. Since representationalism is a species of functionalism about attitudes, it seems natural for the representationalist also to adopt a functionalist stance about ideas. Given this, the task is to display the respects in which the roles of ideas and attitudes differ. We can begin

---

physicalism about propositional attitudes is committed only to the existence of representational physical states or events, and not to the existence of representational physical objects. The requirement also derives some plausibility from the claim that propositional attitudes are realized as relations between thinkers and representations; it is most natural to think of relations as holding between ordinary objects. An analogy with natural language also lends support to the requirement, since languages include representational objects such as sentences.

Nevertheless, the requirement is too strong. It is perfectly possible for there to be computational relations between thinkers and occurrences. Just as a speaker can draw a conclusion from a sentence (a kind of ordinary object), so can he draw a conclusion from an utterance (a kind of occurrence). For analogous reasons, little of computational interest seems to hang on whether mental representations are realized as, say, neurons (objects) or as, say, the firings of neurons (occurrences). Indeed, any apparent importance of the object/occurrence distinction vanishes when we compare treating an object as a representation with treating the *existence* of the object as a representational state. Finally, we can insure the distinction between representationalism and token physicalism by simply denying that ideas are propositional attitude occurrences, without denying that they are occurrences of some other sort.

with a few natural claims which, though strictly speaking circular in this context, may bring out what is meant by "idea". A token idea *that p* (i.e., a particular thinker's idea *that p*) typically persists through changes in token attitudes *that p*. One's idea *that p* (i.e., one's grasp of the proposition *that p*) doesn't vanish when one's belief or desire *that p* vanishes, and indeed one can have an idea *that p* (and not merely in a dispositional sense) without having *any* attitudes *that p*. Second, a token idea *that p* is typically involved in many different token attitudes *that p*.[9]

What can be said by way of giving a more principled distinction between propositional attitudes and propositional ideas? My suggestion is that propositional attitudes are those mental representations which standardly function as *units of reasoning*. Such representations have *rationality values*, i.e., degrees of rationality or irrationality, which can influence the rationality values of other representations or actions, or at least be influenced by other representations or perceptions. A belief that Paris is pretty--or a wish that it were--has a rationality value. By contrast, a mere idea of Paris' being pretty is neither rational nor irrational.

---

[9]As I will explain in a moment, it is this claim, coupled with token physicalism about ideas, which advances us toward the goal of solving the special case of the mental-body problem with which we began this section, namely, the problem of explaining how a physical system can implement systematic causal relations among different attitudes *that p*.

Beyond drawing this connection between propositional attitudes and rationality values, I have very little to say about the proper conception of rationality values. I imagine that, at a minimum, having rationality values is corequisite with having a role as a (potential) premise or conclusion of inference.[18]

However the distinction between ideas and attitudes is ultimately to be spelled out, another way to see the theoretical bite of representationalism is to display its promise as an explanation of how physical systems can implement special causal relations among different attitudes *that p*. The details of such an explanation vary according to the nature of the physical realization of ideas. A possibility at one extreme is that one's idea *that p* is realized as a token physical structure which is part of the physical structures realizing *all* of one's (specially related) token attitudes *that p*. In such a case, I will say that all of these token attitudes are relations to the same token "symbol". The special relations between these attitudes might then be explained by the fact that they literally share an ingredient, a token symbol *that p*. For example, if we

---

[18]What is less clear is whether there is any way to distinguish inferential relations from non-inferential relations among representations (e.g., association of propositional ideas), short of appealing to the rationality values of the representations. I will return to this point in section 2.2.1, where I will also criticize alternative accounts of the difference between attitudes and ideas.

postulate that a thinker has the capacity to have beliefs and desires at all, this explains why his ability to have beliefs *that p* presupposes his ability to have desires *that p*. A possibility at the opposite extreme is that one's idea *that p* is realized as a mechanism which can *reproduce* physical structures each of which is part of only *one* token attitude *that p*. In this case, I will say that each token attitude is a relation to a distinct token symbol of a single physical kind (defined by reference to the reproduction mechanism). Although the token symbols are distinct, if we postulate processes which can *match* symbols of the given physical kind, then we can also begin to understand how to implement the special relations among attitudes with the same content.[11]

---

[11]Since I will often employ the notion of symbols illustrated in this paragraph, I would like to call attention to a few features of my use of the word. Although there is a perfectly useful sense in which anything with content is a symbol, I will usually subject this term to a number of restrictions. First, unless otherwise clear from the context I will usually reserve the word "symbol" for *mental* symbols--i.e., symbols which help to realize propositional attitudes--rather than natural-linguistic or other nonmental symbols. Furthermore, when there is a danger of confusion between token attitudes and token ideas (and there usually is), I will reserve the word "symbol" for physical structures related to the latter (either by identity or by reproduction, as illustrated in the text). Given this usage, although functionalism and token physicalism about propositional attitudes are committed to the existence of mental representations, they are weaker than representationalism in not being committed to the existence of mental symbols. (While this distinction can be expressed in terms of *ideas*, "symbol" emphasizes the physical nature of the structures involved, and also allows me to ignore differences between the two sorts of physical realizations of ideas mentioned in the text.) Finally, "symbol" (like "idea" and "representation") is neutral with respect to the object/occurrence distinction (see footnote 8).

0.1.3 The Language-of-Thought Hypothesis

Just as there appear to be special causal relations among different attitudes with the same content, so there appear to be special causal relations among attitudes of the same kind with different contents. For a wide variety of things we can think about (e.g., sugar, toothpaste, and teeth) thoughts that such a thing is white typically bear special causal relations not only to desires that it be white, but also to thoughts that it is not yellow, that it is white or blue, that something is white, and so on. Again, whatever the ultimate details, such causal relations are widespread, regular, and important enough to be likely factors in any functionalist account of propositional attitudes. Fodor and Pylyshyn have provided a useful characterization of similar relations in terms of what they call "systematicity" (Fodor and Pylyshyn, 1988). I will focus on the details of their treatment in chapter 1. For now, as before, we can appeal to the special causal relations to formulate another aspect of the mental-body problem: how is it possible for these systematic relations to be implemented in a physical system? The language-of-thought (LOT) hypothesis, unlike token physicalism or representationalism, is intended to serve as an answer to this question (among other questions--see Fodor, 1975; Fodor, 1987a).

The LOT hypothesis goes one step beyond representationalism, just as representationalism goes one step beyond token physicalism. According to the LOT hypothesis, the physical symbols postulated by representationalism admit of *syntactic complexity*. What is it for a symbol to be complex?[12] Although this is a very difficult question, we can operate with an intuitive idea, leaving technicalities aside. The prototypical complex symbols are written sentences and phrases in natural language. Each complex sentence and phrase has two or more symbols--e.g., words--as spatiotemporally proper parts, where parthood is taken quite literally (that is, as the phenomenon studied in mereology). Accordingly, syntactically complex mental symbols are thought to have other mental symbols as literal parts.[13] The parthood may be

---

[12]To avoid repeated use of the modifier "syntactic", I will often speak of "complexity" and "simplicity" intending the modifier to be understood.

[13]Although there may be viable but weaker conceptions of syntactic complexity according to which syntactic constituents do not need to be *parts* of complex symbols, Fodor is emphatic that parthood is required for a language of thought. He insists repeatedly that the LOT hypothesis claims that "(some) mental formulas have mental formulas as parts" (Fodor, 1987a, p. 137), and that this notion of parthood is literal:

> Real constituency does have to do with parts and wholes; the symbol 'Mary' is literally a part of the symbol 'John loves Mary'. It is because their symbols enter into real-constituency relations that natural languages have both atomic symbols and complex ones. (Fodor and Pylyshyn, 1988, p. 22)

spatial, as with written sentences, or temporal, as with spoken sentences, or a mixture of the two.[14]

---

[14]If (some) mental symbols are physical occurrences rather than ordinary physical objects, then the LOT hypothesis demands a notion of spatiotemporal parthood for (token) occurrences as well as for (token) individuals. There is some leeway in the construction of such a notion. I will illustrate the tactics available for the case of states, conceived of as instantiations of properties by objects. (I will have to leave open the question of whether a similar account can be given for other sorts of occurrences, such as events.) One intuitively plausible idea is that a state *P* is a part of a state *W* iff *P* is a *logically necessary condition* of *W*. This might serve to capture the intuitively important fact that parts, taken together, *constitute* wholes, so that the existence of a whole depends upon that of its parts. For example, let *S* be the state of neuron *n*'s firing at *t*, and let *S'* be the state of *n*'s firing *and* having a high threshold at *t*. On this account of parthood for states, *S* is a part of *S'*.

While this may seem roughly right as an account of state-parthood in general, the notion of syntactic complexity which it generates is, I think, too weak. As I will explain in section 1.2.4, given this notion of a syntactic state-part, one can literally *prove* the existence of syntactic complexity (and so, nearly enough, the truth of the LOT hypothesis) from the nearly indisputable premise that there is *some explanation or other* which is common to one's ability to engage in a range of semantically related inferences (see Davies, 1990). This would make the LOT hypothesis virtually impossible to reject (without also rejecting, say, token physicalism about propositional attitudes).

To avoid this result, we need to motivate a stronger constraint on syntactic parthood than that suggested by the current account of state-parthood. I think that this account leaves out the *spatiotemporal* aspects of syntactic parthood which are intuitively important for the LOT hypothesis. Suppose, as seems natural, that a state is where the individuals which "participate" in the state are. (Where did Mary hug John? Wherever Mary and John were, of course. But see section 1.2.4 for criticism of this as a general account of state-locations.) Then state *S* is not a spatiotemporally proper part of state *S'*, since their spatiotemporal locations *coincide*. If, as I am suggesting, a reasonably strong LOT hypothesis postulates that some mental symbols are spatiotemporally proper parts of other mental symbols, then I don't think we should count such conjuncts as candidate *syntactic* parts. Rather, a state *P* is a spatiotemporally proper part (and so a candidate syntactic part) of a state *W* only if (i) *P* is a necessary condition for *W and* (ii) *P*'s location is properly within *W*'s location. For example, state *S* is a spatiotemporally proper part of the following two states: (a) neuron *n* and neuron *m*'s firing at time *t*, and (b) *n*'s firing at *t* and *t'*.

I will return to these points in section 1.2.4, where I attempt to show that fully-fledged spatiotemporal complexity is necessary to explain the range of phenomena typically taken to be explained by the LOT hypothesis.

In addition to the requirement of symbolic parts, a semantic requirement is standardly placed on syntactic complexity. Not only do sentences and phrases have other phrases and words as parts, but they also bear some sort of close semantic relation to these parts. Fodor and Pylyshyn express this relation by saying that "the semantic content of a [complex] representation is a function of the semantic contents of its syntactic parts, together with its constituent structure" (Fodor and Pylyshyn, 1988, p. 12). In other words, without delving into too many technicalities, the content of the complex symbol must *depend* on the contents of its parts, as the content of "Mary loves John" depends on the content of "loves", but not, intuitively, on the content of "neighbor" or "weigh".

The LOT hypothesis helps to explain how a physical system can implement systematic relations among attitudes, in much the same way that representationalism helps to explain the relations among different attitudes with the same content. For example, on the assumption that symbols have syntactic parts, it might be that two systematically related token attitudes physically overlap, i.e., share some of the same

token parts.[15] Alternatively, such attitudes might contain tokens of a physical kind such that they can easily be reproduced from or matched against one another. This would allow the implementation of inferential processes such as variable-introduction and variable-binding which are sensitive to the syntactic structure of symbols, and are thereby sensitive to some of the semantic dependencies of the attitudes the symbols help to realize.[16]

---

[15]This is the case with certain semantic (or propositional) networks of the sort often contained in traditional cognitive-scientific models (for a review, see Johnson-Laird, et al., 1984). In such a network, nodes are symbols for objects and properties (among other things), and pieces of the network (i.e., groups of nodes along with their connections) are symbols which help to realize attitudes. Groups of attitudes (thought to be) about the same thing (e.g., toothpaste) typically share a node representing that thing. This allows mechanisms easily to implement inferential relations among these attitudes.

[16]This sort of matching occurs, for example, in "production system" models (see Anderson, 1983; Holland, et al., 1986). In these models, a thinker's thoughts and goals are realized by storing syntactically complex symbols in various buffers, including long-term and working memory, where they may be acted upon by inferential processes. Some of these inferential processes are realized by special kinds of "IF...THEN..." rule-symbols called "productions". Although details vary from theory to theory, a production may be thought of as a rule-symbol with a (tiny) processor. The processor's task is to watch out for the presence of a symbol matching its "IF" part (modulo differences between variables and constants), and to perform some simple action corresponding to its "THEN" part, such as forming a copy of the "THEN" part in working memory (perhaps with variables bound or introduced). It is as if one could write a conditional sentence in a book, give it tiny eyes and arms, and give it one reflex: when you see a copy of your "IF" part written somewhere, write down a (possibly modified) copy of your "THEN" part (or do something comparably simple). With the use of variables, a single production (e.g., "IF *x* is white, THEN *x* is not yellow") can implement systematic causal relations among a wide range of pairs of token attitudes which have parts of a single physical type (e.g., the beliefs *that toothpaste is white* and *that toothpaste is not yellow*, the desires *that teeth are white* and *that teeth are not yellow*, etc.).

In the mid-seventies, Fodor focused philosophical attention on the fact that nearly all existing models in cognitive science satisfy the language-of-thought hypothesis (and so, by implication, representationalism). Models of this sort have since become known as "classical" models. Although such models continue to dominate the field, from the earliest days of cognitive science various philosophers and scientists have found themselves dissatisfied with the classical framework as a whole. We can understand some of the phenomena which have seemed to cast doubt on classical models by focusing on a few striking facts about a certain sort of skill acquisition.

Novices at many activities (driving, playing basketball, delivering speeches, etc.) usually operate by deliberating--by applying memorized rules and recipes (or, at least, rough, partial ones). Often, the novice's rules are of limited reliability and scope, and are applied slowly and haltingly, consciously and self-consciously. With routine practice, however, performance often improves dramatically. One can recognize more of the relevant aspects of what's happening, settle on better decisions in a wider range of situations, and execute those decisions more fluently even under duress. It therefore appears that the expert's inferential processes are somehow sensitive to vastly more conditions than are the novice's inferential processes. Paradoxically, these increases in what might be called "holistic sensitivity" are

also accompanied by great increases in speed--the expert can do far more than the novice, but can do it far more quickly. This paradox has seemed to many to lead to a difficulty for classical models.

I have been portraying the classical framework as being motivated by a specific aspect of the mental-body problem: how can inferential processes be realized physically? The analogy with computers--as enshrined in the LOT hypothesis, for example--does appear to be a promising answer to this question. However, as the inferential processes are considered to be sensitive to more and more conditions, and to operate more and more quickly, critics of the classical framework have felt less and less sure of the computer analogy as an account of how such processes are to be implemented.[17] The worry needs to be formulated more carefully. Certainly present-day computers can store large numbers of symbols, and quickly access and transform them all. Similarly, classical theorists can appeal to analogies with ever-larger and ever-faster computers.

---

[17]These objections to the classical model should not be confused with considerably weaker objections which appeal to unconscious processing or processing without rule-symbols. Notoriously, the expert's improvements over the novice with respect to holistic sensitivity and speed coincide with--and seem to require--a reduction in the conscious application of rule-symbols. But it is not a necessary feature of classical models that processes should be conscious, nor even that they should be realized by rule-symbols.

Since computers are physical, this strategy is relevant to the mental-body problem. Seen in a larger philosophical context, however, such a solution to the mental-body problem would be unsatisfying. We don't merely want to know how *some* physical system can implement mental processes (such as holistically sensitive inference). We want to know how *we* can do so, i.e., how it is possible for these processes to be implemented in *brains*. Call this the "mental-brain problem". We are especially interested in this problem not simply because we have a parochial interest in human cognition, as we might be especially interested in news from the old hometown. Rather, the special interest derives from the philosophical interest in assessing the fate of functionalism and token physicalism as answers to the mind/body problem. The mind/body problem is a problem about *existing* mental phenomena, including at the very least human mental phenomena. If functionalism and token physicalism cannot furnish answers to the mental-brain problem, then they cannot be general solutions to the mind/body problem.

Why should the mental-brain problem seem any more difficult than the mental-body problem? The worry only begins to take shape when we notice relevant differences between brains and supercomputers. Without entering into needless detail, the important point is that neurons are extremely *slow* in relation to the symbol-manipulating processes of familiar computers

(even setting aside imaginary improved computers).[16] As a result, the aspect of the mental-brain problem having to do with the implementation of rapid, holistically sensitive inferential processes seems especially challenging. This is one of the central challenges the connectionist framework is meant to address. We will be better able to understand the potential difficulty for the classical framework if we introduce an alternative connectionist approach.

0.2 The Connectionist Framework

0.2.1 What Is Connectionism?

Before entering into a direct description of connectionist approaches to this issue, it is best separately to introduce a few key notions, and to consider standard sorts of

---

[16]Here are the numbers as reported by a trio of highly influential connectionists:

> Neurons operate in the time scale of milliseconds whereas computer components operate in the time scale of nanoseconds--a factor of 10^6 faster. This means that human processes that take on the order of a second or less can involve only a hundred or so time steps. (Rumelhart, et al., 1986, p. 75)

The appeal to neurons is necessary to motivate the worry, since a classical theorist can point out that some physical processes in brains--e.g., quantum-mechanical ones--are fast even in relation to the symbol-manipulating processes in present-day computers. It is an interesting question why physicalists are loathe to postulate that token mental occurrences are much "smaller" than neurons. Robert Cummins offers the plausible argument that doing so would deprive us of an explanation of the biological inheritance of mental properties, since (as far as we know) it is only roughly cell-sized features which are determined by genes (Cummins, 1983, pp. 63-65).

connectionist models which are not centrally concerned with explaining holistically sensitive inference. Fortunately, the issues of present concern can be described without focusing on the fine details of connectionist networks. The most important idea is that of a *node*. Nodes are simple energy-transmitting devices which, in the simplest case, are characterized at a given time by their degree of activation, or propensity to affect other nodes. Nodes are connected to one another by stable energy conduits, by means of which active nodes tend to alter the activation of other nodes. (Finer details about these connections will turn out to be of no concern.) Although some nodes may have direct connections only to other nodes, others may interact with sensory or motor mechanisms, or (perhaps) with nonconnectionist cognitive mechanisms.

This specification of connectionist models makes no mention of mental phenomena, and so is silent on the question of functionalism and token physicalism about propositional attitudes. Indeed, it is possible to be a connectionist and deny that any propositional attitudes (whether of a sort familiar to common sense or of a sort to be discovered by scientific psychology) are realized in the models one adopts. Presumably, such a position would not respond to our latest outbreak of the mental-body problem--the question of how (rapid, holistically sensitive) inferential processes can

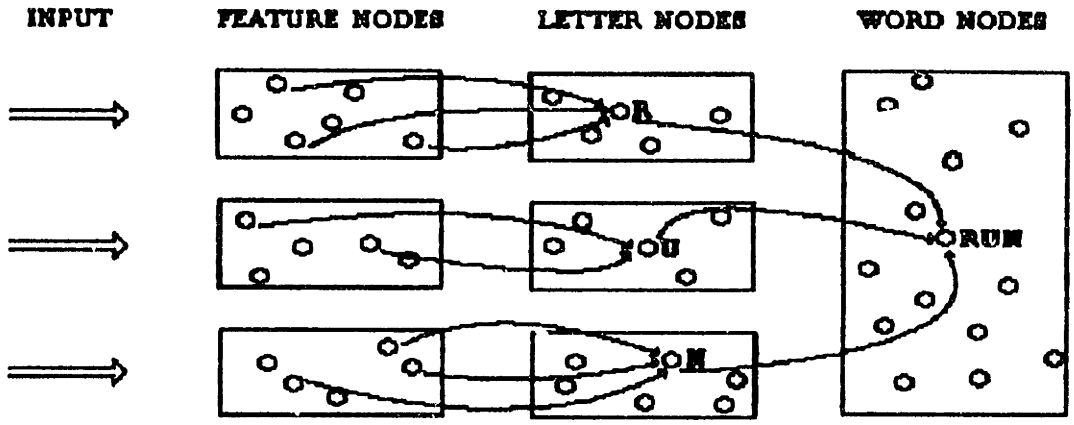
possibly be implemented in brains. For this reason, there is special interest in connectionist theories which are coupled with a functionalist and token physicalist approach to propositional attitudes; I will limit my discussion to such versions of connectionism.

Where are the representations in connectionist models? On one conception, individual nodes are representations.[19] Perhaps the most famous example is the "interactive activation model" of reading (Rumelhart and McClelland, 1982). It may be pictured in part as in Figure 1. The network contains "word nodes" each of which standardly becomes activated as a result of the presentation of a particular word, and each of which represents the presence of that word. These word nodes are activated by "letter nodes" each of which represents and standardly responds to the presence of a particular letter at a particular position in the word. Finally, each letter node is activated by "feature nodes", each of which represents and standardly responds to the presence of a certain feature of the shape presented at a particular position: a horizontal bar, a curved top, etc.

---

[19]Connectionists often fail to be explicit about whether the representations are nodes *themselves*, or rather *states* of nodes (such as activation levels). I will discuss this distinction in the next section, while discussing the relation between connectionism and representationalism. For now, when I speak of a node as a representation, this can be taken as a simplified way of speaking of a node *or* one of its states as a representation. The same simplification applies to talk of *groups* of nodes as representations.

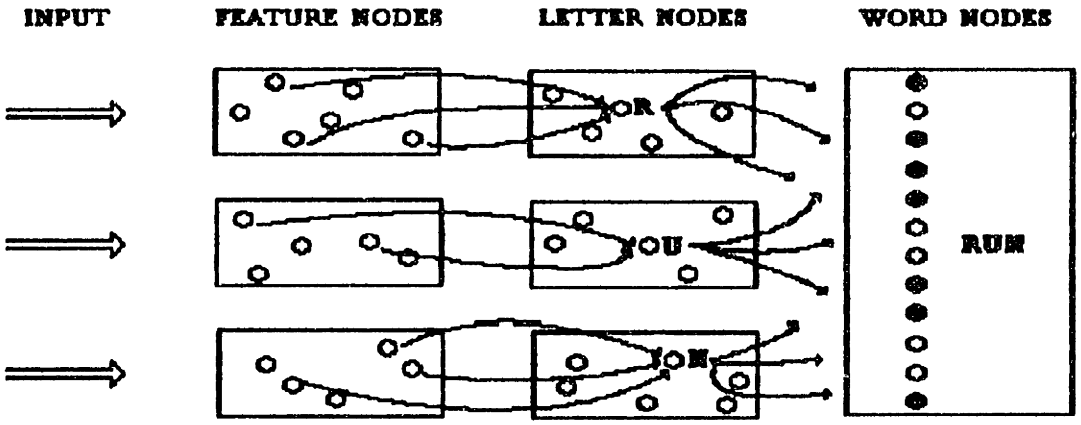
Figure 1: *Rumelhart and McClelland's (1982) interactive-activation model of reading.*



Individual nodes which serve as representations are called "local" representations, in contrast with "distributed" representations, which are patterns of activity of many nodes. For example, we can imagine a modification of the interactive activation model of reading, in which the presence of each word is represented not by an individual node, but by several nodes, as in Figure 2. Here the connections between the letter nodes and the nodes in the word patterns are arranged so that, for example, the letter nodes "R", "U", and "N" tend to activate all and only the nodes in the pattern which represents the presence of the word "RUN" (those darkened in Figure 2). On most distributed schemes of representation, representations overlap, that is, share nodes. In the present example, the word "ANT" might be represented by a pattern of activation which includes some of the same nodes as those in

the pattern which represents "RUN". Similarly, we might imagine further modifications of the model in which the letter nodes are replaced by letter patterns, the feature nodes are replaced by feature patterns, and perhaps even the extreme case in which all the representations--for features, letters, and words--are patterns of activation defined over all the nodes in the model.

Figure 2: *The interactive-activation model of reading modified to include distributed representations for words.*



0.2.2 Connectionism and Representationalism

While it is clear that the most common and interesting sorts of connectionist networks are supposed to contain token

propositional attitudes,[28] it is a tricky question whether any of these representations qualify as *symbols* rather than token attitudes, in the sense required by representationalism (see section 0.1.2). While the situation is bound to vary from model to model, reflection on certain fairly standard features of the (real and imagined) reading models seems to suggest that most familiar connectionist models are supposed to adhere to representationalism.

As a preliminary consideration, connectionists normally attribute content to (groups of) nodes in such models. Since (groups of) nodes are ordinary objects rather than occurrences, they cannot be token attitudes, and so (by a process of admittedly nondemonstrative elimination) must be ideas or symbols. Of course, it is possible to object that this talk is merely sloppy, and that the only representations are states of *activation* of the nodes, which do stand a chance of being identical to token attitude states. Intuitively, the models "cognize" *that the word "RUN" is present* only when the appropriate nodes are activated, so there is support for the view that only activation states are representations. Nevertheless, I think there is a point to considering the

---

[28]This is only "clear" if we do not limit the class of propositional attitudes to those which are familiar from commonsense, and if we ignore claims of the sort that genuine propositions can only be represented by (say) syntactically complex representations. The latter sort of claim receives extended discussion in chapter 2.

nodes themselves to be representations. The point is that these nodes appear to play the functional and explanatory role of ideas, and so (by functionalism) they must be ideas.

Consider first that the models are capable of entering into a variety of attitudes with the same content, e.g., *that the word "RUN" is present*. Some states of the "RUN" nodes play a functional role akin to that of *forming a hypothesis*, while other states of the same nodes play a role akin to that of *concluding*. There are a wide range of intermediate attitudes as well. The nodes themselves are common parts of these various attitudes, just as a token idea is typically involved in many different token attitudes with the same content. The nodes therefore afford a physical explanation of the special causal relations among these different attitudes, in precisely the sense in which ideas are supposed to explain these relations. Furthermore, like token ideas, the nodes persist through changes in the corresponding attitudes, and may exist without the presence of any corresponding attitude. I suggest that these considerations make plausible the claim that standard connectionist models adhere to representationalism, and that typically (groups of) nodes are token ideas and symbols. If this is right, the relation of connectionist models to the classical framework appears to turn on whether they adhere to the language-of-thought hypothesis, and in

particular on whether they support syntactically complex symbols.[21] This will be the main focus of chapter 1.

0.2.3 Connectionism and Holistically Sensitive Inference

We are now in a position to consider how the connectionist framework affords strategies for solving our specific version of the mental-brain problem: how can rapid, holistically sensitive inferential processes be realized by physical relations between slow (at least by current technological standards) neuron-like elements? Recall that an inferential process is holistically sensitive to the extent that it depends on a wide range of conditions. Experts at various skills--auto racing, basketball, conversation, etc.--appear to engage in processes which are sensitive to an indefinitely wide range of conditions, and do so in "real time", even in fractions of a second. A cognitive system with an unimaginably fast processor might duplicate this performance by considering each condition in turn, assessing the relevance of each to the inferential task at hand (possibly making several "passes"), and then drawing the appropriate inferences. However, this strategy is unavailable for us, on the assumption that our processors are (relatively slow) neurons.

---

[21]At any rate, most connectionist critics of the classical framework set up the issue specifically in terms of syntactic complexity.

The connectionist framework makes available two strategies for minimizing the time necessary to engage in holistic processes within the limits of neuron-like processors. The first strategy is massively parallel processing. If many conditions relevant to an inference can be considered at the same time, by separate processors, the total time needed for the inference may be decreased substantially. The second strategy is the avoidance of syntactic complexity. It appears likely that syntactically complex representations place a heavier computational burden on processors than syntactically simple representations do. For this reason, if conditions relevant to an inferential process can be adequately represented by syntactically simple representations, the time it takes to consider each condition is reduced.

These ideas are illustrated in a partial connectionist theory of skill due to Philip Agre (Agre, 1989). On Agre's theory, well-practiced, routine activity is guided by an enormous "dependency network", each node of which corresponds to a potential belief, desire, or other propositional attitude of the agent. The agent has the attitude when the node is activated, and doesn't have it when the node is inactivated. The network is connected in such a way that the nodes activate and deactivate each other in "mimicry" of inferential arguments which were previously generated (presumably) by the

relatively long, drawn out calculations of a more flexible mechanism (e.g., the novice's conscious trial-and-error search). The circuitry allows hundreds or thousands of these "arguments" to take place in massive parallel and at blinding speed. As Agre describes the system, it "is continually redeciding what to do insofar as the circuitry produces a fresh set of decisions about action many times a second" (Agre, 1989, p. 139).[22]

0.3 An Overview of the Thesis

Like many others, I find models within both the connectionist framework (such as dependency networks) and the classical framework (such as production systems--see section 0.2.3) attractive in accounting for inferential processes. One underlying aim of my thesis is to get us closer to the goal of deciding between these models, or between their relatives (although I regret to say that my conclusions leave us very far from this goal). The thesis addresses three specific issues raised by philosophers interested in the emerging debate between the classical and connectionist frameworks.

---

[22]Although Agre's dependency networks rely almost exclusively on local schemes of representation (one representation per node), analogous models are possible which make more use of distributed symbols. Such alternatives can decrease the number of nodes required of a network, and can increase its reconfigurability (ability to change which symbols affect which other symbols), but at the cost of decreased speed.

0.3.1 A Syntactic Issue

Fodor's contention in The Language of Thought (Fodor, 1975) was that, whatever the fate of particular existing models of cognition, the language-of-thought hypothesis must be satisfied by any adequate model. Those who wished to resist this conclusion were especially hampered by the fact, repeatedly stressed by Fodor, that no serious alternatives to LOT models had ever been proposed. Today, however, with the advent of connectionist models, many people at last see promise of a plausible way to avoid commitment to the LOT hypothesis. This is where the fun starts. Fodor stands his ground; along with Zenon Pylyshyn, he argues that connectionist models which fail to implement a language of thought also fail to account for certain fundamental aspects of cognition, and in particular "systematic" relations among propositional attitudes akin to those with which I began section 0.1.3. Some critics, most notably Paul Smolensky (1989), have tried to provide counterexamples to this conclusion.

In chapter 1, I will display the failings of Fodor and Pylyshyn's argument which make these counterexamples initially plausible, but I will also argue that Fodor and Pylyshyn's conclusion withstands the alleged counterexamples. This issue

is relevant to whether models of skill such as Agre's dependency networks can be modified so as to account for systematic relations among attitudes *without* adopting the LOT hypothesis and falling within the classical framework.

0.3.2 A Semantic Issue

My concern in chapter 2 is to address a puzzle about the content of representations in certain connectionist models, such as the interactive activation model of reading and Agre's dependency networks. The puzzle arises within particular philosophical views of content, versions of what I will call the "fine-grained theory". According to these views, contents admit of degrees of *complexity*. Even the simplest propositions (e.g., the proposition *that a is F*) are thought to be complexes of constituent concepts (e.g., the concepts *a* and *F*). What is required for a representation *r* to have a complex content, say, the proposition *that a is F*? On the fine-grained theory, as I will construe it, *r* must display "semantic dependence" on *other* representations of the concepts *a* and *F*, i.e., the constituent concepts of the proposition. What sort of relation between representations counts as semantic dependence? The most familiar examples are syntactic relations: the content of the English sentence "sugar is white" depends on the content of its syntactic parts "sugar" and "white". Another example of semantic dependence might

loosely be termed "abbreviation": a syntactically simple representation "p" in a logical formalism may have a complex, propositional content in virtue of being treated (by convention or otherwise) as "standing in place of" a sentence such as "sugar is white", and so depending semantically on the parts of that sentence.

Although virtually all nonconnectionist models in cognitive science, as well as many connectionist models, postulate relations such as syntactic complexity and abbreviation, many connectionist models, including dependency networks and the reading model, appear *not* to support these relations. Nevertheless, for reasons I will describe there is at least an appearance that representations in these models *do* have propositional contents. This generates a philosophical puzzle, at least for those sympathetic to the fine-grained theory of content: how it is possible for a representation to have propositional content without displaying semantic dependence on other representations (e.g., without being either syntactically complex or an abbreviation)? My goal in chapter 2 is to explain how.

0.3.3 A Processing Issue

In my discussion of holistically sensitive inference, I have considered a kind of brute-force approach: that such

inferential processes do access a large number of representations, and that the main problem is to show how a system can maximize the number of representations which can be accessed in a short amount of time. While massive parallelism and avoidance of syntactic complexity are effective techniques for increasing speed, they (especially the latter) suffer from serious computational deficiencies (of the sort described in chapter 1). It therefore appears that we need an alternative to the brute-force approach; in particular, we need to develop processing strategies which minimize the number of representations which need to be accessed in implementing a given inferential process.

We can begin by noticing two senses in which an inferential process can be sensitive to a condition, and so two senses of "holistically sensitive inference". An inferential process can be sensitive to a condition in the sense that it always operates by accessing a representation of that condition, or it can be sensitive in the sense that it has potential access, when necessary, to the representation. Although it is clear that expert activity is sensitive to a wide range of conditions, it is unclear in which sense this is so.

If the expert's increased holistic sensitivity consists of accessing far more representations than novices access, then the increased speed associated with expertise is paradoxical,

and a brute force approach involving connectionist technology may appear inevitable. If, instead, the increased holistic sensitivity consists in an increased range of representations which the expert can access only when necessary, then we can even begin to *explain* the increased speed. Perhaps the expert can avoid continually accessing representations which the novice continually has to access, because the expert has a better ability to "tell" when they are irrelevant to the activity. While cognitive scientists, particularly those studying artificial intelligence, have tried to develop algorithms for modeling such an ability, several philosophers have tried to portray one or another version of the "frame problem" as principled objections to these strategies.

The frame problem is widely reputed among philosophers to be one of the deepest and most difficult problems of cognitive science. Chapter 3 discusses three recent attempts to display this problem: Dennett's problem of ignoring obviously irrelevant knowledge, Haugeland's problem of efficiently keeping track of salient side effects of occurrences, and Fodor's problem of avoiding the use of "kooky" concepts. In a negative vein, I argue that these problems bear nothing but a superficial similarity to the frame problem of AI, so that they do not provide reasons to disparage standard attempts to solve it. More positively, I argue that these problems are easily solved by slight variations on familiar AI themes.

Finally, I devote some discussion to more difficult problems confronting AI. If the arguments I provide are correct, then we may not need to abandon classical models (or abandon classical features in connectionist models) in explaining how rapid, holistically sensitive inference can be implemented with processors as slow as neurons.

0.3.4 Summary

In this chapter I have tried to do two things: (1) display classical and connectionist models as responses to questions raised by a functionalist and token physicalist approach to the mind/body problem, and (2) display a substantive dispute between classical and (certain) connectionist models of inferential processes in skills. When I have completed my discussions of the language-of-thought hypothesis, fine-grained theories of content, and the frame problem, how will the models appear to stand with respect to skills and the mind/body problem? I don't know, to be honest. I will not be presenting an argument to the effect that one of the models is better than the other with respect to skills. At best, I will be trying to remove certain difficulties with interpreting these models, and trying to locate the theoretically interesting differences between them, so that we have a better chance of testing them against facts about skills. Nor will I be presenting an argument to the effect

that either framework is likely to solve the mental-brain problem, and so I will not be taking a stand on the likelihood that functionalism and token physicalism solve the mind/body problem. At best, I will have cleared away some apparent difficulties for these models, and will have brought other difficulties into sharper focus.

Chapter 1

CONNECTIONIST LANGUAGES OF THOUGHT

Fodor and Pylyshyn (1988) have presented an influential argument to the effect that any viable connectionist account of human cognition must implement a language of thought. Their basic strategy is to argue that connectionist models which do not implement a language of thought fail to account for the systematic relations among propositional attitudes. Several critics of the LOT hypothesis have tried to pinpoint flaws in Fodor and Pylyshyn's argument (Smolensky 1989; Clark, 1989; Chalmers, 1990). One thing I will try to show is that tne argument can be rescued from these criticisms. (Score: LOT 1, Visitors 0.) However, I agree that the argument fails, and I will provide a new account of how it goes wrong. (The score becomes tied.) Of course, the failure of Fodor and Pylyshyn's argument does not mean that their conclusion is false. Consequently, some connectionist criticisms of Fodor and Pylyshyn's article take the form of direct counterexamples to their conclusion (Smolensky 1989; van Gelder, 1989; Chalmers, 1990). I will argue, however, that Fodor and Pylyshyn's conclusion survives confrontation with the alleged counterexamples. I then argue more generally that no genuine

counterexamples are likely to be forthcoming. (Final Score: LOT 2, Visitors 1.)

The chapter is divided into two major sections. In section 1.1 I consider Fodor and Pylyshyn's argument as directed against "ultralocal" connectionist models which contain only individual nodes as representations (see section 0.2.1). Then in section 1.2 I will try to locate the basic flaw in Fodor and Pylyshyn's argument as directed against distributed connectionist models (ones which posit groups of nodes as representations), and analyze the two major counterexamples to Fodor and Pylyshyn's conclusions: Paul Smolensky's "coffee case", and Tim van Gelder's account of "nonconcatenative compositionality" as exemplified in representational schemes such as Smolensky's "tensor product" framework. My conclusion will be that, contrary to initial appearances, these connectionist models do implement a language of thought.

1.1 Ultralocal Connectionism and the LOT Hypothesis

The point of this section is to set out the dispute between the LOT hypothesis and certain forms of connectionism, particularly "ultralocal" connectionist models which contain only local representations (individual nodes). First I will describe the language-of-thought hypothesis and assess the importance of the issue (sections 1.1.1 and 1.1.2). Then I

will present and defend Fodor and Pylyshyn's systematicity argument as applied to ultralocal models (section 1.1.3).

1.1.1 What is a Language of Thought?

As I explained in section 0.1.3, the language-of-thought hypothesis is a specific version of representationalism, which is the claim that propositional attitudes are (at least typically) physically realized as computational relations between thinkers and propositional ideas or "symbols". The LOT hypothesis goes one step beyond representationalism in asserting that some mental symbols are syntactically complex. I will proceed with the assumption that for a symbol to be syntactically complex, it must have (multiple) symbols as parts, and its content must depend on the content of these symbols (see section 0.1.3). Conveniently, we can come to understand the current debate without considering whether this is ultimately the right account of syntactic complexity. This is because no critics of Fodor and Pylyshyn--at least none of which I am aware--have taken issue with the present account of what complexity is. While the dispute has been operating at this intuitive level, however, perhaps some of the consequences of this account will motivate devoting more attention to the notion of syntactic complexity itself.

It might be worth pointing out a few things that the LOT hypothesis does not require. It does not say that *all* mental symbols are syntactically complex. Nor does it say that there are syntactically complex symbols in every natural "faculty" of the mind: it would be consistent with the LOT hypothesis for there to be no complex symbols in (say) the olfactory and motor systems, so long as there are complex symbols elsewhere in the mind. This raises the question: how *many* syntactically complex symbols are needed to establish the truth of the LOT hypothesis? We should not consider the LOT hypothesis true of a mental life which contains only *one* complex symbol, buried under a sea of noncomplex symbols. On the other hand, there is no precedent for any requirement to the effect that a certain minimum *percentage* of mental symbols must be complex.

To make the LOT hypothesis into an interesting psychological hypothesis, I will construe it as going one step beyond the requirement of syntactic complexity. A common understanding of the LOT hypothesis can be captured intuitively by the claim that the *expressive power* of a language of thought is at least as extensive as the expressive

power of natural languages.[23] To a first approximation, this means that any content possessed by complex natural-linguistic symbols can also be possessed by complex mental symbols.[24] Without going into too many technicalities, this claim needs to be restricted. The language-of-thought hypothesis doesn't (by itself) require that *everyone's* language of thought must *at each stage of their development* contain symbols corresponding to *all possible* natural-linguistic symbols. While something like this extremely strong claim has been defended by Fodor, he has never treated it as part of the LOT hypothesis itself. Instead, I will suppose that for the LOT hypothesis to be true, any content possessed by complex natural-linguistic symbols *comprehensible to a thinker at a time* can also be possessed by complex mental symbols available to that thinker at that time.

Despite the fact that the language-of-thought hypothesis goes beyond the syntactic complexity requirement, it is this

---

[23]A weaker construal of the LOT hypothesis is required for nonlinguistic thinkers, such as monkeys and babies, and for nonlinguistic faculties of the adult mind. However, I don't know how to specify, without bare stipulation, how much syntactic complexity should be required in these cases, if any.

[24]There is no requirement to the effect that *simple* symbols in natural language must correspond to simple symbols in a language of thought. On many traditional LOT models, individual words in natural language are analyzed into conglomerations of mental "feature" symbols which, taken individually, are unexpressed in natural languages. This is why Smolensky's distinction between "conceptual-level" and "subconceptual-level" semantics is irrelevant to the language-of-thought hypothesis (Smolensky, 1988).

requirement which has stimulated the most influential connectionist attacks. Accordingly, my focus in the rest of this chapter will be on syntactic complexity. I will only address criticisms of Fodor and Pylyshyn which at least implicitly begin with the assumption of representationalism (see section 0.2.2), and I will ignore issues of expressive power.

1.1.2 The LOT Hypothesis and Particular LOT Models

It is important to distinguish the general language-of-thought hypothesis from hypotheses about particular implementations of languages of thought. There are many different cognitive models in existence which implement languages of thought, including several dozen versions of production systems (see section 0.1.3), and several dozen versions of "logic-based systems".[28] The language-of-thought hypothesis does not require that any particular one of these traditional models applies to human cognition. It doesn't even require *any* of these models to apply, since it may be possible to invent new LOT models which differ in cognitively

---

[28]For readers unfamiliar with these architectures, the details will not be relevant. For the sake of concreteness, such readers can think of any high-level programming language--such as Basic, Pascal, or Fortran--when I mention production systems or logic-based systems. Each of these programming languages supports syntactically complex symbols, and yet they all differ in the primitive symbol manipulations they make possible. This suits them well as analogies for cognitive architectures in the language-of-thought mold.

interesting ways from traditional ones. Indeed, if the arguments I will give below are correct, some of the most interesting connectionist models qualify as "new" LOT models.[26] To be clear, then, we must distinguish two questions which may be asked about a given connectionist model:

(1) Does the model implement a language of thought?
(2) Does it implement some (or any) particular preexisting *implementation* of a language of thought (such as production system *X*)?

I think that the dispute between connectionists and defenders of the LOT hypothesis has been marked by a confusion between these two questions.

Fodor and Pylyshyn argue that the answer to question (1) is likely to be "yes" for any potentially plausible connectionist models. However, many of their supporters and critics have supposed that they have argued for a "yes" answer to question (2).[27] For this reason, it is sometimes supposed that Fodor

---

[26]Furthermore, these are *important* LOT models because they may advance us toward an explanation of how the classical framework can be implemented in a brain-like system, and so may provide a strategy for solving the mental-brain problem (see section 0.1.3).

[27]The temptation to this construal is understandable. Fodor and Pylyshyn do sometimes express their conclusion by denying that connectionism provides a (plausible) "new cognitive architecture". However, all they appear to mean by this is that any plausible connectionist cognitive architecture will implement a language of thought. They don't appear to mean that no plausible connectionist architectures will differ in any interesting ways from preexisting LOT models. Suppose tomorrow someone designs a cognitive architecture which differs from existing LOT systems--production systems, logic-based systems, etc.--as Pascal differs from Basic: it introduces a number of new primitive operations, while maintaining a clear commitment to syntactically complex symbols in a language of thought. Strictly speaking, of course, this would be a "new" architecture. In the context of discussing the fate of

and Pylyshyn have provided an argument "against" connectionism, or at least an argument to the effect that connectionism is "uninteresting" from a philosophical or cognitive-scientific perspective. Smolensky, for example, says that Fodor and Pylyshyn are "after" the conclusion that

> . . . since the Classical account provides a complete, precise, algorithmic account of the cognitive system, there is nothing to be gained by going to the [connectionist] account, as long as the phenomena of interest can be seen at the [Classical] level (Smolensky, 1989, p. 4).

This construal of Fodor and Pylyshyn's claims is misleading.

First of all, because of the variety of existing (and possible) models in the language-of-thought mold, there is no such thing as "*the* Classical account" or "*the* Classical level". Fodor and Pylyshyn use the term "Classical" for any system which implements a language of thought. It would seem, then, that "the" Classical account of cognition is simply the LOT hypothesis, which as Fodor and Pylyshyn are well aware certainly does not provide "a complete, precise, algorithmic account" of cognition. For all of Fodor and Pylyshyn's arguments, connectionism may be the revolutionary breakthrough cognitive science has been awaiting for decades, philosophy for centuries. At best, what they have done is to interpret

---

LOT, however, it would be reasonable for someone like Fodor and Pylyshyn to say that it *isn't* really "new", because it is of the LOT variety. Since this is precisely the gloss they give to their denials that connectionism promises anything "new", it would clearly be wrong to construe them as answering "yes" to question (2).

the likely nature of any connectionist breakthrough, namely, that it would not overturn the LOT hypothesis, the hypothesis that there are syntactically complex symbols.

While Fodor and Pylyshyn have question (1) in mind, Smolensky's arguments appear to be directed at question (2). He objects to Fodor and Pylyshyn's use of the word "implementation" on the grounds that it is too weak:

> There is a sense of "implementation" that cognitive science has inherited from computer science, and I propose that we use it. If there is an account of a compositional system at one level and an account at a lower-level, then the lower one is an *implementation* of the higher one if and only if the higher description is a complete, precise, algorithmic account of the behavior of that system. It is *not* sufficient that the higher-level account provide some sort of rough summary of the interactions at the lower level. (Smolensky, 1989, p. 4)

Smolensky's proposal is very surprising. He requires for implementation that the higher-level account be "complete", but does not explain what this requirement comes to. In the familiar logical sense, a complete description of something is one which provides every truth about the thing. But this cannot be the relation between high-level descriptions and models. Surely leaving out some lower-level truths--being an *in*complete or "abstract" description of the lower-level--is precisely what makes a description "higher-level" in the first place. We therefore cannot require that a higher-level description provide the "whole truth" about the lower level. This requirement would simply be impossible to fulfill.

Instead, for a model to implement a description what is required is that the description must be *sound*--must contain no falsehoods about the behavior of the model. Rather than requiring an account to contain the "whole truth", as Smolensky appears to do, all that can be required for implementation is that it contain "nothing but the truth".[28]

---

[28]It is *this* sense of implementation which is operative in computer science. For instance, consider this crude, high-level algorithm:

Step 1. Read a word.

Step 2. Reverse the letters.

Step 3. Print the result.

Now, suppose that there is a program--in an IBM language, a Macintosh language, or whatever--which, when run, makes it true that the system reads a word, reverses the letters, and prints the result (in that order). This would be sufficient for the program to be an implementation of these three steps. It wouldn't matter if there were all sorts of side-effects or novel features of the program, unmentioned in the algorithm (such as printing the result in *italics*); in fact, there are bound to be such extra effects. On the standard logical construal of "complete", Smolensky's claims about implementation would make it literally impossible for steps 1-3 to be implemented in a program. Although this account of implementation would show that connectionist systems do not implement a language of thought, it would do so only by making it impossible for *any* theory (except, perhaps, a completely specified fundamental physics) to be implemented.

I imagine that other, more plausible, construals of "complete" are possible. One possibility is that Smolensky means "complete" as "taking one all the way from input to output", to allow for the implementation of algorithms such as steps 1-3 above. Even this seems too strong, since algorithms which are "partial" in this sense seem implementable. (Doesn't a program which implements the algorithm consisting of steps 1-3 "also" implement the algorithm consisting of steps 1-2?)

Another possibility is that Smolensky's "completeness" and "precision" requirements should simply be downplayed in favor of his "algorithm" requirement. Perhaps his concern is to distinguish the "implementation" of algorithms from the "implementation" of representational schemes, such as syntactic complexity. I have found that some people find it unnatural to speak of "implementations" of representational schemes, preferring to speak of "realization" or "instantiation". Even so, this does not appear to be of more than terminological significance. Fodor and Pylyshyn need not be understood as claiming that connectionism must implement syntactically complex symbols. If "implementation" is only possible for algorithms, Fodor and Pylyshyn should be understood as claiming that connectionism must implement some algorithm or other which operates on syntactically complex representations. The substance of their claim would

The hypothesis that Smolensky is focusing on question (2) is, so far as I can see, the only plausible explanation of his temptation to strengthen the notion of implementation in the way he appears to do. The explanation might go as follows. First, unlike a formulation of the language-of-thought hypothesis, a specification of a production system or other cognitive architecture is (ideally, anyway) a specification of *all* the basic symbol manipulating processes available to a given system. Therefore, even given the *weak* conception of implementation for which I have argued--namely, that implementation requires simple lack of falsehoods--no connectionist model would be an implementation of a particular production system (say) unless they shared *all* their basic symbol manipulating processes. But this would be to say that the production system architecture must be a "complete" description of at least the *cognitive* behavior of the connectionist system (ignoring physical breakdowns, etc.). On the assumption that Smolensky has latched on to question

---

be unaffected.

At any rate, Smolensky needs to do more than support the "complaint" that the language-of-thought hypothesis merely "provides a rough, higher-level approximation to the connectionist account" and merely "involves some of the same basic ideas about how information is represented and processed". Similarly rough, approximate relations holds between connectionist theories and neurological (not to mention biochemical and quantum-mechanical) models. What Smolensky needs to show in order to deny implementation (properly construed) is that his connectionist models are incompatible with the language-of-thought hypothesis. Fortunately, he does provide an argument to this effect. I will take it up in section 1.2.2.

(2), then, we can understand his otherwise bizarre construal of "implementation" as the application of the *proper* construal in a very special case, namely, that of implementations of particular, completely specified, cognitive architectures. But since the LOT hypothesis seeks only to constrain mental representations, and not to specify a complete cognitive architecture, it needn't even come close to giving a complete description of the behavior of its implementations.

If I am right, Smolensky argues for a "no" answer to question (2), but concludes that the answer to question (1) is "no". Confusing the two questions in this way is a mistake. Even if connectionist models *do* differ from every preexisting implementation of a language of thought (production system architecture *X*, logic-based architecture *Y*, etc.) in cognitively interesting ways, they might still implement a language of thought, since they might still support syntactically complex symbols.

It is not surprising, I think, that the confusion of questions (1) and (2) has ruled the debate. The first question might seem a more natural focus for philosophers, the second more natural for cognitive scientists. Thus, Fodor and Pylyshyn are far more interested in question (1) than Smolensky needs to be: from all indications, if the questions were adequately distinguished, Smolensky might reasonably

declare himself completely uninterested in question (1). Even if Smolensky's connectionist models are forced, on pain of implausibility, to implement a language of thought, this would not detract from their novelty and scientific interest. On the other hand, Fodor has a special stake in question (1), since he wants to defend the LOT hypothesis from philosophers who have sought to use connectionism against it. It is less clear that Fodor has any special stake in question (2): he is, after all, a notoriously vehement critic of traditional cognitive-scientific models, especially those rooted in artificial-intelligence research (I discuss some of his criticisms in chapter 3). From all indications, Fodor and Pylyshyn would be quite happy if connectionists conceded that their most plausible models satisfy the LOT hypothesis, even if it were insisted that these models differ in some other cognitive details from traditional forms of production systems, or logic-based systems, etc.

If the dispute between the LOT hypothesis and connectionism has been so riddled with confusion, shouldn't it be allowed to die quietly? The problem with doing so is that philosophical suspicion of the language-of-thought hypothesis has been boiling under the surface for too long to be downplayed in what are still early stages of theorizing about connectionism. I think, instead, that the dispute should be split into two disputes, corresponding to the two questions

I have been distinguishing. Both questions are interesting and important, although perhaps to thinkers of different temperaments. Neither issue seems especially easy to settle. I don't share the common feeling that upon cursory examination the differences between connectionist models and traditional models "jump out" at us. What jumps out at us is that there are *some* interesting differences. But it is far from clear what these differences are. In the rest of this chapter, I will be concerned only with syntactic complexity as an alleged difference, and so will be concerned only with question (1), as it arises with respect to various connectionist models.

1.1.3 Ultralocal Connectionist Models

In connectionist networks, individual nodes with content are called "local representations", while contentful groups of nodes are called "distributed representations" (see section 0.2.1). Although it has become common to speak of local and distributed *models* as well as representations, most connectionist models contain both local and distributed representations, in various proportions. For purposes of understanding Fodor and Pylyshyn's argument, however, it is useful to begin with an extreme case, that of "ultralocal" models. An ultralocal connectionist model is one in which there are *only* local representations. The original interactive model of reading (shown in section 0.2.1, Figure

1) approximates this case, as do the dependency-network models developed by Agre (see section 0.2.3). The most relevant thing to notice about such models is that they do not contain any syntactically complex symbols. Since local symbols (individual nodes) do not have symbols as parts, they are syntactically simple. If ultralocal models provide a general account of the mind, therefore, the language-of-thought hypothesis is false. Fodor and Pylyshyn argue, however, that ultralocal connectionist models do not provide a general account of the mind. This is because these models fail to explain certain pervasive facts about human cognition, most importantly what Fodor and Pylyshyn call "systematicity". Although they try to extend these arguments to connectionist models other than ultralocal ones, it is best to begin with the ultralocal case.

The mark of systematicity, for Fodor and Pylyshyn, is the presence of presupposition relations among mental capacities. For example, our ability to have certain thoughts presupposes an ability to enter into certain others. For some reason, we never find people who can entertain the thought *that Mary loves John* without being able to entertain the thought *that John loves Mary.* Nor do we find people who can form an idea of *a cup of black coffee* without being able to form an idea of *a black cup of coffee.* And so on, for most familiar mental states. Inferential capacities are also systematically

related; we don't find people who can infer *P&Q* from *P&Q&R*, for example, but who can't infer *P* from *P&Q*.[39] Fodor and Pylyshyn argue that pervasive presupposition relations can hardly be accidental, and so a theory of cognition should provide a guarantee of systematicity. Given the language-of-thought hypothesis, we can explain the systematic presuppositions by postulating that the symbols which help to realize systematically related attitudes or ideas literally overlap, or share certain token syntactic parts. Alternatively, we can postulate that the syntactic parts of such symbols are tokens of a physical kind such that there are mechanisms which can easily reproduce and match these tokens, and can easily arrange them in various ways. This would also explain why the ability to form certain propositional attitudes systematically presupposes the ability to form certain others. Similarly, by postulating inferential processes which can operate consistently on representations containing a common (token or type) syntactic part (such as a conjunction symbol), we can explain systematic relations among inferential capacities.

---

[39]These examples all consist of *pairwise* presupposition relations between attitudes. Systematicity applies more generally as a relation between groups of attitudes: e.g., the availability of the pair of thoughts *that sugar is white* and *that coffee is brown* presupposes the availability of the pair of thoughts *that coffee is white* and *that sugar is brown.* This is so even though these thoughts do not enter into pairwise presupposition relations to one another.

On the other hand, according to Fodor and Pylyshyn, ultralocal models contain no mechanism for insuring that any given node presupposes the existence of any other node. There is thus no explanation of the systematic relations between, say, the thought *that Mary loves John* and the thought *that John loves Mary*. If these thoughts were realized as individual connectionist nodes, Fodor and Pylyshyn argue, there would be no principled reason why one thought is thinkable by all and only the people able to think the other. Someone who wished to embrace ultralocal connectionist models as plausible alternatives to languages of thought might, perhaps, be willing to settle for a huge number of stipulations about the mental "hardware"--e.g., an *ad hoc* stipulation to the effect that the two nodes which realize these particular thoughts about John and Mary do in fact presuppose one another. But this would seriously weaken the case against the LOT hypothesis. Until some connectionist mechanism is specified which can *insure* that the stipulations hold, it is reasonable for Fodor and Pylyshyn to point out that the only known mechanism for doing so is one which includes syntactically structured symbols. If their criticisms are correct, then while ultralocal models may be appropriate for certain aspects of cognition (such as initial processing in reading), these models may have to share space in the mind with LOT models. Therefore, ultralocal connectionist models would not provide critics of the

language-of-thought hypothesis with a plausible, general alternative.

It appears that many of the most prominent connectionist commentators are willing to concede this result to Fodor and Pylyshyn, with a reminder that ultralocal models are atypical of connectionist models in general. Before looking at the wider class of connectionist models, however, I want to consider a philosophical line of reply which, if correct, would rescue even ultralocal models from Fodor and Pylyshyn's criticisms.

Many philosophers hold that systematicity is a conceptual necessity rather than a fact in need of psychological explanation. They hold that it is simply part of proper practices of thought-ascription that if one can't think *that John loves Mary*, then one can't think *that Mary loves John* (see Evans, 1983). How might this view, if correct, be used on behalf of ultralocal models? It might be argued that ultralocal models *can* guarantee systematicity of thoughts, because if a node in a particular model fails to stand in a systematic presupposition relation to another one, we should simply refuse to ascribe to it the content *that Mary loves John* or *that P&Q*. Far from being a mystery for ultralocal connectionists, systematicity is a conceptually necessary

feature of cognition. Thought must be systematic, because whatever isn't, isn't thought.

Fodor concedes that his argument "takes it for granted that systematicity is *at least sometimes* a contingent feature of thought", for otherwise "you don't need LOT to explain the systematicity of thoughts" (Fodor, 1987a, p. 152). I think this is understating his case, however. Even if the philosophical position described in the previous paragraph is correct, and correct for *every* case of systematicity, a variant of Fodor and Pylyshyn's argument still applies. For if it is a requirement on the ascription of thoughts that they be systematically related, then ultralocal connectionist models simply have no explanation of how we *do* manage to have the thought *that Mary loves John* or *that P&Q*. It is surely not an accident that every person has a great many such thoughts, but it is a mystery why this should be so if thoughts are realized as individual nodes. In other words, by enforcing a strict philosophical requirement on thought ascription, we are left with no account of how people manage to have as many thoughts as they do--assuming that these

thoughts are realized by ultralocal models.[30] To provide such an account, ultralocal models need to provide some mechanism for insuring that systematic relations hold among nodes, and this is precisely what Fodor and Pylyshyn argue has not been accomplished.[31]

1.2 Distributed Connectionism and the LOT Hypothesis

Connectionist opponents of the language-of-thought hypothesis have considered their most powerful weapon to be features of distributed symbols, or contentful patterns of activation of multiple nodes. If Fodor and Pylyshyn's argument were directed solely at ultralocal models, therefore, it would be little cause for connectionist alarm. My first

---

[30]A possible escape route for the defender of connectionist models is to appeal to a behaviorist or instrumentalist theory of mental states which denies representationalism (see Clark, 1989, for such a response to Fodor and Pylyshyn). Consideration of such a view is beyond the scope of this paper, which seeks to operate within the mainstream cognitive-scientific (and mainstream connectionist) assumption of representationalism.

[31]Why is it relevant that we have "a great many" (systematically related) thoughts? Wouldn't a creature with only a few systematically related thoughts (say, an insect-like creature capable of "thinking" only *that warm food is near*, *that cold water is near*, *that cold food is near*, and *that warm water is near*) be a mystery for an ultralocal connectionist theory (say, which holds that these four insect-thoughts are realized by nodes *n*1,...,*n*4, respectively)? Yes, in the sense that the connectionist architecture itself does not explain why every insect has all four nodes. However, the connectionist could appeal to an *alternative* explanation of this fact, perhaps a biological one. The situation is different for humans, who (appear to) have vast numbers of thoughts (which, moreover, vary from person to person). This fact renders a directly biological explanation of the copresence of thoughts implausible.

task will be to discuss their attempt to extend their conclusion from ultralocal to distributed cases (section 1.2.1). My claim will be that they *fail* to show that distributed connectionist models which account for systematicity are likely to implement a language of thought. This will pave the way for a consideration of attempts to provide actual counterexamples to their conclusion, in the form of distributed models which account for systematicity without syntactic structure (sections 1.2.2-1.2.3). I will argue that, on a careful analysis of these models, they do in fact use syntactically structured representations. Thus, my final conclusion will be that Fodor and Pylyshyn's conclusion withstands the fall of their argument.

1.2.1 The Limitations of Fodor and Pylyshyn's Argument

Surprisingly Fodor and Pylyshyn say next to nothing by way of extending their conclusions about the ultralocal case to their conclusions about connectionism in general. While they consider the path between the two conclusions to be short, it is not immediately clear from their discussion what they consider the path to be. This has led critics to misidentify Fodor and Pylyshyn's strategy for extending their conclusions, or even to suggest that they offer no extending argument at

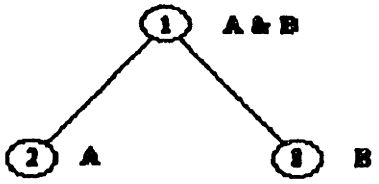
all."[11] What they do say is confined to a single sentence in a footnote, a comment on the particularly crude ultralocal network shown in Figure 3, which they use in mounting their systematicity argument. (The labels "A", "B", and "A&B" are not parts of the model; rather, they specify the contents of the nodes.) Before proceeding, let me recap their argument that such ultralocal models cannot account for systematicity. The key problem they identify for ultralocal models is that there is no plausible connectionist mechanism for insuring the presence of one node given another. Given connectionist architecture, it is perfectly possible for a mental life to contain node 1 in the diagram without containing node 2 or node 3. For that reason, as I have put it, ultralocal connectionists must implicitly make large numbers of independent, *ad hoc*, assumptions about presupposition relations among bits of mental hardware (i.e., among the individual nodes which realize systematically related symbols).

Given this way of looking at the argument about ultralocal models, we can understand their attempt to extend the argument to distributed models:

---

[11]For example, while Fodor and Pylyshyn argue persuasively and at some length that distributed models don't *necessarily* provide systematicity, Smolensky mistakenly takes this to be a (bad) argument to the effect that distributed representation *cannot* provide systematicity (Smolensky, 1989, p. 8). For a similar suggestion see Chalmers, 1990, p. 4.

Figure 3: *Fodor and Pylyshyn's (1988) sample ultralocalist network for drawing inferences from* A&B *to* A *or to* B.



> To simplify the exposition, we assume a 'localist' approach, in which each semantically interpreted [symbol] corresponds to a single Connectionist [node]; but nothing relevant to this discussion is changed if these [symbols] actually consist of patterns over a cluster of [nodes]. (Fodor and Pylyshyn, 1988, p. 15)"

Given the astonishing brevity of this formulation, it is perhaps understandable that critics have either ignored it or responded to it with an equally brief assertion that there is "no argument to be found there" (Chalmers, 1990, p. 4). Nevertheless, the passage contains an argument of some force. Fodor and Pylyshyn's idea is to have us imagine an interpretation of Figure 3 on which the circles signify groups of nodes rather than individual nodes. The point of this seems clear. Just as there is no plausible mechanism for insuring the presence of a particular *individual* node given another individual node, so there is no mechanism for insuring the presence of a particular *group* of nodes given another group. The only way to insure systematicity, for such models, is simply to posit for each bit of hardware (i.e., each

---

"Fodor and Pylyshyn use the term 'units' for what I have been calling 'nodes', and the word 'nodes' for what I have been calling 'symbols'. I have altered their quotation accordingly.

relevant group of nodes) that there are certain corresponding bits of hardware (i.e., other relevant groups of nodes).

The fundamental difficulty with Fodor and Pylyshyn's attempt to extend their argument to distributed symbols is that it assumes that the distributed symbols are realized by *distinct* groups of nodes. It appears likely, as they insist, that given the presence of one group of nodes, there is nothing to insure the presence of another, distinct, group of nodes. However, on many of the most interesting connectionist schemes of representation, distributed symbols are patterns of nodes which *overlap*. How might this help to account for systematicity? Suppose that *s*1 and *s*2 are systematically related symbols, realized in a particular distributed connectionist model as different patterns of activation levels *p*1 and *p*2 over the *same* group of nodes *n*. Given that the model is capable of producing *s*1, it must contain the nodes over which *s*2 is distributed--namely, *n*--since the two symbols are distributed over the very same nodes. To explain why the two symbols are systematically related, then, one needn't make ad hoc assumptions of presuppositions among various bits of hardware. All that is needed is some mechanism for insuring that a model capable of producing pattern *p*1 over a group of nodes is also capable of producing some other pattern *p*2 over the same nodes. This is fundamentally a *programming* problem for connectionism; the hardware problem is nonexistent. Fodor

and Pylyshyn have given no reason to suppose that this programming problem cannot be solved in a general and principled fashion. They have simply failed to address the most formidable connectionist models, those based on overlapping distributed symbols.

Of course, this purely negative point does not mean that distributed models *can* be developed which insure systematicity without appealing to syntactically structured representations. Perhaps the only way to insure systematic relations among overlapping patterns of activation is to identify *subpatterns* as syntactic parts. Or perhaps not.[34] At this stage, only examination of particular representational schemes will help to decide this point. That is what I want to do in the rest

---

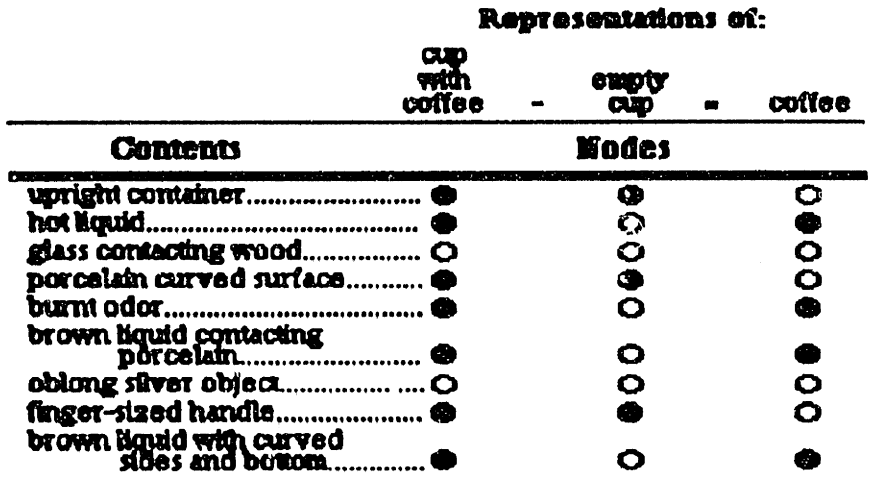
[34]For certain purposes, Fodor and Pylyshyn's failure to consider overlapping distributed representations is a *good* thing. They want to show that no connectionist models can insure systematicity without implementing syntactically complex symbols. They do not and should not want to make the stronger claim that no connectionist models can implement complex symbols, period. This stronger claim appears to be false, since connectionist networks can implement Turing machines (at least those with a finite tape), and Turing machines can implement complex symbols. But it is at first difficult to see how Fodor and Pylyshyn's argument can avoid this strong conclusion. Since they argue that ultralocal models do not implement syntactic complexity, and assert that "nothing is changed" by moving to distributed models, their argument seems to apply to all possible connectionist models. Chalmers (1989) offers a "refutation" of their argument based on this fact. His objection fails because what Fodor and Pylyshyn are actually claiming is that "nothing is changed" by moving to distributed models by substituting *distinct* groups of nodes for *distinct* individual nodes. This leaves open the possibility that *some* connectionist models may implement syntactic complexity. However, it also leaves open the possibility that these models may implement systematicity *without* complexity, so Fodor and Pylyshyn fail to accomplish their argumentative goal.

of this chapter. I will consider models which illustrate two broad strategies connectionist opponents of the LOT hypothesis might pursue. Since there are two requirements on syntactic complexity--a symbolic-parthood requirement and a semantic-dependence requirement (see section 0.1.3)--a connectionist model need only fail one or the other of these requirements in order to avoid a language of thought. One option for a connectionist is to account for systematicity with distributed models in which symbols do have symbolic parts, but do not stand in an appropriate semantic dependence relation to these parts. The best-articulated example of such a strategy is Paul Smolensky's influential *coffee* example, which I discuss next. A second strategy is to seek to account for systematicity with distributed models in which symbols stand in appropriately semantic dependence relations, without standing in part/whole relations. I will consider Tim van Gelder's attempt to analyze Smolensky's "tensor product" networks in this way, and then conclude with more general reasons why no connectionist counterexamples to Fodor and Pylyshyn's claims are likely to be forthcoming.

1.2.2 Smolensky's *Coffee* Case

Smolensky is concerned to highlight various differences between complex LOT symbols and corresponding distributed connectionist symbols. He draws a contrast between a particular LOT symbol of a cup with coffee--the formula "with(cup,coffee)"--and a particular distributed connectionist representation of a cup with coffee. In the typical LOT case, it is possible to start with the complex symbol, and "subtract" the representation of an empty cup--say, "with(cup,NULL)"--to produce a representation of *coffee* itself. Furthermore, this representation of *coffee* is context-independent in that it retains its meaning in other complex symbols, such as "with(can,coffee)", "with(tree,coffee)", and "with(man,coffee)". Smolensky suggests that we perform a corresponding subtraction procedure on a distributed connectionist representation of *cup with coffee*, to see if what results can properly be called "a connectionist representation of *coffee*". His arrangement can be pictured as in Figure 4. (The black dots in the rightmost three columns represent, respectively, the nodes constituting Smolensky's representations of *cup with coffee*, *empty cup*, and *coffee*. The left column represents the content of each individual node.) Smolensky emphasizes two features of this representational scheme.

Figure 4: *Smolensky's (1989) connectionist representation of* coffee.



First, unlike local symbols (see section 1.1.4), the representation of *cup with coffee* has a sort of "compositional structure", since it can be formed by combining the other two groups of nodes. The model therefore provides at least a measure of systematicity among the symbols for *cup with coffee*, *empty cup*, and *coffee*. Second, Smolensky hesitates to call this compositional structure *syntactic* structure, since he insists that the structure is present only in an "approximate" sense. What he means by this is that, unlike in the LOT case, the *coffee* symbol (represented by the black nodes in the rightmost column) is not a context-independent representation of coffee. Instead, Smolensky suggests, it is "really a representation of *coffee* in the particular context of being inside a cup". We can see why he says this by

looking at the contents of the darkened nodes: *hot liquid, burnt odor, contacting porcelain,* etc. Also, as he points out, if we were to start with corresponding assemblies of nodes representing *can with coffee, tree with coffee,* or *man with coffee,* subtraction would result in very *different* representations of *coffee*: as burnt smelling granules stacked in a cylindrical shape (from the subtracted can), as brown beans hanging in mid-air (from the subtracted tree), or as a cup with coffee (from the subtracted man). Therefore, he concludes:

> [The structure is] *not* equivalent to taking a context-independent representation of *coffee* and a context-independent representation of *cup*--and certainly not equivalent to taking a context-independent representation of *in* or *with*--and sticking them all together in a symbolic structure, concatenating them together to form the kind of syntactic compositional structures like "with(cup,coffee)" that [Fodor and Pylyshyn] want connectionist nets to implement. (Smolensky, 1989, p. 11)

In short, according to Smolensky, the context dependence of the representations in the model keep it from implementing a language of thought.

To assess this argument, one thing we need to know is what the relevance of context dependence is for the fate of the language-of-thought hypothesis. Here we must be particularly careful to respect the distinction between languages of thought in general and particular preexisting implementations of languages of thought. The two questions I distinguished

in section 1.1.2 must be kept separate in dealing with Smolensky's distributed *coffee* model:

(1) Does the model implement a language of thought?
(2) Does it implement some (or any) particular preexisting *implementation* of a language of thought?

It is easy to imagine that context dependence is relevant to question (2). Many traditional LOT models do not employ context-dependent symbols, so that Smolensky can use context dependence to distinguish his *coffee* model from those. To say that context dependence is relevant to question (2) is not to say that it is *conclusively* relevant, since there might be some traditional models which do exhibit context dependence.

However this issue is settled, something more is needed to show that context dependence is even partially relevant to question (1), the question which Fodor and Pylyshyn address. In particular, it must be shown that context independence is required for syntactic structure. If it is not required, then Smolensky's argument is simply misdirected. He does not take a clear stand on this point, saying for example that context dependence renders the structure of representations "approximate". Certainly he provides no argument to the effect that syntactic structure requires context independence. Given this, and given other aspects of his discussion such as those discussed in section 1.1.2 above, it appears likely that this is due to his confusion of questions (1) and (2). Setting this aside, however, we need to consider what can be

said in defense of the idea that context independence is a requirement on syntactic complexity.

As I mentioned in section 1.1.1, there is widespread agreement (for better or worse) that syntactically complex symbols not only have symbolic parts, but also depend semantically on these parts, in the sense that the meaning of a syntactically complex symbol must depend on the meanings of its parts. But suppose that a particular symbol *a* (connectionist or not) is context dependent, having different meanings as part of the symbols *Fa*, *Ga*, etc. Then it seems wrong to suppose that the meaning of *Fa* depends on the meaning of *a*. Instead, the dependence relation is reversed: the meaning of *a* seems to depend on the meaning of *Fa*. At best, the meaning of *Fa* depends on some property of *a other* than its meaning. We might imagine, for example, that there is some property of *a* other than its meaning, which is context independent and which, together with context, determines its context-dependent meaning. The meaning of the resulting symbols *Fa*, *Ga*, etc. would depend on this property of *a*, but this would not be a dependence relation between *meanings*, and so would not be a semantic dependence relation of the sort typically required. Consequently, context independence is at least an initially plausible requirement on syntactic

complexity."[11] Whether it is a genuine requirement on syntactic complexity depends, among other things, on whether semantic dependence, or indeed any semantic relation, is genuinely required of syntactic complexity. But for purposes of this chapter I am working along with the consensus that these requirements are genuine.

Even if a defender of the language-of-thought hypothesis must favor (at least some degree of) context independence, it is possible to take a more direct tack in replying to Smolensky. I would like to make a case for denying that his coffee case *does* display context dependence. The illusion of context dependence stems from misassignments of content. Consider again the darkened nodes in the first column of the

---

[11]A more careful account would distinguish between two sorts of context dependence: (1) context dependence of *contents*, in which a single symbol changes meanings depending on what other symbols are present, and (2) context dependence of *symbols*, in which a single content is expressed by different symbols depending on what other symbols are present. While the first case is that treated in the text, it appears that Smolensky's model is at best an example of the latter variety: he suggests that *coffee* is represented by different sets of nodes in different contexts. The conditions for syntactic complexity are violated in some, but not all, such cases. Let *a*1, *a*2, etc. be all the potential *coffee* symbols in a given model. Let *C*(*a*1), *C*(*a*2), etc. be the set of contexts in which *a*1, *a*2, etc., respectively, are used to mean *coffee*. If each *ai*, taken alone, means *coffee*, then despite the multiplicity of *coffee* symbols, the semantic condition for syntactic complexity is met: the content of each member of *C*(*ai*) depends on the content of *ai*. The exceptional case is that in which some *ai*--say, *a*1--fails to mean *coffee* when taken alone. In this case, the content of *a*1 changes depending on its context, and so we have the same situation as that discussed in the text. For all Smolensky indicates, however, his various *coffee* symbols do represent *coffee* when taken alone. (What else might they represent? Tea? Nothing at all?) I don't know whether this is an artifact of his particular example.

previous diagram. Smolensky says that when these nodes are activated, they form a symbol with the content *cup with coffee* (or the proposition *that a cup with coffee is around*, or some such content). An alternative interpretation is that the assembly of nodes has as its content the *conjunction* of the contents of the individual nodes. If when one node is on, it means *upright container*, and when another is on, it means *hot liquid*, why shouldn't the two together mean *upright container and hot liquid*? Following this train of thought, we might conclude that the assembly in the third column is *not* a context-dependent representation of *coffee*, but a straightforward, context-independent representation of *hot brown liquid with burnt odor, curved sides, and bottom contacting porcelain*. Why isn't Smolensky's third column simply a syntactic *concatenation* of a symbol for hot liquid, one for burnt odor, and so on? This looks exactly like an implementation of the following "syntactic compositional structure that [Fodor and Pylyshyn] want connectionist nets to implement":

```
HotLiquid(x) &
BurntOdor(x) &
BrownLiquidContactingPorcelain(x) &
BrownLiquidWithCurvedSidesAndBottom(x)
```

There are several advantages to this interpretation. First, it is more precise, and reflects better the system's treatment of the symbols. For example, imagine what would show that the content of the first column was *cup with coffee* instead of

this alternative. Plausibly, a deciding factor is whether the symbol (even ideally) responds to cups with samples of coffee *other* than hot, brown liquids with . . ., such as cups with *dried* coffee, or cups with coffee *beans*, etc. But this is precisely what Smolensky's alleged *cup with coffee* symbol does not (even ideally) do, by his own hypothesis.[36]

It is worth describing one other manifestation of the fact that Smolensky's content ascription is too generous. After illustrating the variety of *coffee* symbols in different contexts (cup, can, tree, man, etc.), he makes the following provocative claim:

> That means that if you want to talk about the connectionist representation of *coffee* in this distributed scheme, you have to talk about a *family of distributed activity patterns*. What knits together all these particular representations of *coffee* is nothing other than a type of family resemblance. (Smolensky, 1989, p. 12)

On my suggested reinterpretation, we can see that the alleged role of family resemblance is also an illusion. Suppose we interpret the system as having a variety of symbols representing *hot brown liquids . . ., burnt smelling granules . . ., hanging brown beans . . .*, and so on. Even having every member of this "family" of symbols is not enough to represent something as *coffee*. To do this, plausibly, a

---

[36]Similarly, it is not clear that the symbol (even ideally) avoids responding to cups with *fake* coffee--hot, brown liquids with . . ., but which are not coffee.

system must at least represent the hot liquid in the porcelain as being *the same kind of stuff* as the granules in the cylinder or the beans on the tree. Furthermore, it is important to realize that this is not at all explained by Smolensky's "family resemblance" of the assemblies of nodes. For all Smolensky says, there is no (causally) relevant respect in which these nodes resemble each other any more than they do any other random assemblies! Instead, for all Smolensky has shown, what is needed is for the system to contain *another* symbol which it can use to categorize all of these assemblies as being about the same stuff. And why wouldn't *that* symbol be a context-independent representation of *coffee*?"

---

[3]"One way for this to happen is for this symbol to be a shared (token or type) part of the various assemblies Smolensky speaks of as bearing a family resemblance to one another. It may even be the symbol's sole function to label the assemblies as representing the same kind of stuff, and so to label information about one as relevant to hypotheses about the other. If Smolensky wishes to avoid postulating a separate coffee symbol which is used to "knit together" the various coffee-assemblies, and if he wishes to deny that there are any microfeatural representations common to all these assemblies, then he needs some other explanation of how the system treats the various assemblies as representations of the same kind of stuff.

Perhaps Smolensky's idea of "family resemblance" is to be cashed out as follows: the *hot brown liquid . . .* assembly shares some critical number of nodes with the *burnt smelling granules . . .* assembly, which in turn shares the critical number of *different* nodes with the *hanging brown beans . . .* assembly, and so on. If this is so, although the *hot brown liquid . . .* and *hanging brown beans . . .* assemblies (say) would not share the critical number of nodes needed to be classified immediately as representing the same kind of stuff, the system might classify them as doing so, by virtue of detecting a chain of sufficiently overlapping assemblies which "connects" them. If the same-stuffness of *hot brown liquid . . .* and *hanging brown beans . . .* (and all the other relevant entities) is kept implicit in the existence of these chains, then the system could avoid having a *coffee* symbol. In order to determine whether two assemblies represented the same kind of stuff, the system would

I conclude that Smolensky's *coffee* case constitutes no threat at all to the language-of-thought hypothesis. The appearance of a threat stems from Smolensky's mistaken assignments of particular contents to his models. As far as I can tell, my reinterpretation strategy does not depend on the particular contents Smolensky assigns to the individual nodes in the *coffee* model (e.g., *hot liquid*, *burnt odor*, etc.). Although Smolensky suggests that the "microfeatures" represented by individual nodes in more realistic models will be less easily expressed in natural language, my argument will generalize to these cases. The reinterpretation strategy seems to depend only on Smolensky's assertion that symbols for *coffee* are composed of symbols for microfeatures of coffee."[11] Can Smolensky run a version of his argument without microfeatures? I don't think so. Without microfeatures, although Smolensky could stipulate that *coffee* is represented by a number of distinct groups of nodes on different

---

attempt to find a suitable chain between the two assemblies. I don't know whether such a search process can be constrained to the right sort of chains (e.g., to exclude representations of *tea*, *cocoa beans*, etc.), and Smolensky provides no hints as to whether this is even the sort of account he would wish to develop.

[11]Strictly speaking, the strategy also depends on the *conjoinability* of the microfeatural symbols. But all this requires is that the microfeatures be micro*properties* or micro*relations*--conjoinable as "*F*(*x*1,...*x*n) & *G*(*y*1,...,*y*m)"--or else micro*occurrences*--conjoinable as "*p* & *q*"--or else micro*objects*--conjoinable as "*a* & *b*". There is no indication from connectionist theories that microfeatures might be anything else.

occasions, he would have no argument for context sensitivity--that is, no argument that these different groups of nodes have (even slightly) different *contents* (e.g., *coffee in the context of a cup, coffee in the context of a tree,* etc.)."

"It may be worthwhile to consider an independent argument which Smolensky gives to cover distributed models in which individual nodes are not representations at all (and so are not representations of microfeatures). In particular, he considers models which are supposed to meet the following two conditions, presented as he formulates them:

(a) Interpretation can be assigned to large-scale activity patterns [i.e., groups of nodes] but not to individual units [i.e., nodes];

(b) The dynamics governing the interaction of individual units is sufficiently complex that the algorithm defining the interactions of individual units cannot be translated into a tractably-specified algorithm for the interaction of whole patterns. (Smolensky, 1989, p. 5)

He claims that in such models, "the *syntax* or processing algorithm strictly resides at the lower level, while the *semantics* strictly resides at the upper level", implying that "there is no account of the architecture in which the same elements carry both the syntax and the semantics", so that "we have a fundamentally new candidate for the cognitive architecture which is simply *not* an implementation of the Classical one" (Smolensky, 1989, pp. 5-6).

There is reason to doubt the possibility of the models Smolensky has in mind, as well as reason to doubt the conclusions he wishes to draw from them. It is clear enough that a model could satisfy (a), but what about (b)? Although Smolensky doesn't argue that (b) is possible to satisfy, he suggests that the possibility is "brought out nicely" in work by Robert Cummins and Georg Schwarz (Cummins and Schwarz, 1987; Schwarz, 1987; and Cummins, 1989). It is true that in the earlier works Cummins and Schwarz *suggest* the possibility of satisfying (b), but they do not show how it is possible. Furthermore, in the later work Cummins *retracts* the suggestion that it is possible to satisfy (b):

> Given an algorithm for computing the activation levels of individual nodes, it is mathematically trivial to construct an algorithm for computing the activation vector [defined over a group of nodes] whose elements are computed by the original algorithm. It is, of course, a mere change in notation. (Cummins, 1989, p. 151)

He explains that in the earlier works he and Schwarz

> . . . were overly conservative on this point, leaving it open that an algorithm defined at the single-node level might not "aggregate up" into an algorithm defined at the vector level. But it is obvious that the algorithm is always possible--indeed trivial. (Cummins, 1989, p. 171)

Although Cummins doesn't stop to defend these claims, his retraction at least undercuts Smolensky's appeal to his authority! Rather than

1.2.3 Nonconcatenative Complexity

Recently, Tim van Gelder has tried a different strategy in response to Fodor and Pylyshyn. His argument seems best described with the aid of a generic example. Suppose that *s* and *w* are mental symbols--connectionist or not--and that *s* means *sugar* and *w* means *white*. Suppose that we are interested in developing a representational framework involving such symbols which exhibits systematicity. This would involve, at a minimum, the framework's supporting the use of *s*, *w* and other symbols to generate systematically related symbols such as a symbol *P* with the content *that there is a cup with coffee and white sugar*, and a symbol *Q* with the content *that there is a white cup with coffee and sugar*. As van Gelder points

---

attempting to settle the issue of the possibility of satisfying (b), I would like to grant this possibility to Smolensky and argue that his conclusion with respect to the LOT hypothesis is unwarranted.

I agree with Smolensky that the LOT hypothesis requires there to be entities with both syntax and semantics. What is unclear is why in such models the semantically interpreted activity patterns could not have syntactic properties. Smolensky equation of "syntax" with "processing", although initially surprising, has some point: in the classical framework syntactic properties are supposed to play a causal-explanatory role in (e.g.) inferential processes. If (b) is satisfied, Smolensky argues, "at the higher level, . . . complete, precise algorithms for the processing cannot be stated" (Smolensky, 1989, p. 5). Regardless of whether completeness and precision are required for theory-implementation (see section 1.1.2), a process need not be complete and precise (much less completely and precisely *stated*) to be real, and to have causal-explanatory power. If this were a requirement on causal power, only entities at the level of fundamental physics could have causal-explanatory power (and then, perhaps, only if we ignore the statistical--imprecise?--nature of quantum-mechanical laws).

out, such a framework should provide "general, effective and reliable processes" for producing *P* and *Q* given *s*, *w*, and other symbols, and for using *P* or *Q* inversely to produce these symbols (van Gelder, 1989, p. 5). Call these processes "conversion processes". His argument rests upon a distinction between two ways of providing conversion processes.

The first is the familiar method of forming syntactically complex symbols, by "concatenating" *s*, *w*, and other symbolic parts to form systematically related wholes such as *P* and *Q*. Since, on this method, *s* and *w* are physically present within *P* and *Q*, it is fairly easy to design processes which can "parse" *P* or *Q* back into their parts; in effect, *P* and *Q* "wear their logical form on their faces". van Gelder is clear that models employing this first method are implementations of the language-of-thought hypothesis, whether they are connectionist or nonconnectionist. However, he introduces a second sort of conversion method in which the systematically related symbols are "nonconcatenative". He cites several examples of connectionist models (ones with distributed symbols) which have the interesting feature of providing the required conversion processes *without* preserving *s*, *w*, et al. as literal parts of *P* and *Q*. He emphasizes that in a nonconcatenative scheme of representation the systematically related symbols such as *P* and *Q* do not literally contain symbols corresponding to semantic constituents, such as *s* and

w. Therefore, these systematically related symbols have no syntactic structure. It is this feature which van Gelder exploits to argue that connectionist models using nonconcatenative conversion processes can explain systematicity without satisfying the language-of-thought hypothesis.

I think it should be conceded without a fuss that nonconcatenated symbols don't satisfy the LOT hypothesis, since they don't have symbolic parts. Rather, a defense of the LOT hypothesis should begin with a reminder of the obvious: the hypothesis doesn't say that *every* mental symbol is syntactically complex, but says only that *some* are (or, as I suggested in section 1.1.1, that enough are to generate the expressive power of natural languages). Furthermore, of course, from the fact that van Gelder's models[40] use *some* syntactically simple symbols, it doesn't follow that they use *no* syntactically complex ones. Indeed, my strategy will be to argue that these models do, and apparently must, supplement the use of nonconcatenative symbols with straightforwardly concatenative ones. If so, this is enough to show that they implement languages of thought.

---

[40]van Gelder's role in this discussion is that of a philosophical commentator, so that the phrase "van Gelder's models" should be taken as shorthand for "the models which van Gelder considers", a broad class including, most notably, Smolensky's tensor-product representations (Smolensky, 1989).

To begin, I want to look at a conversion method which, while nonconcatenative in van Gelder's sense, is clearly part of a language-of-thought model. Suppose that people have syntactically complex mental symbols--for illustration's sake, we can suppose that these are tiny strings which resemble English formulae, such as "there is a cup with coffee and white sugar", and that homunculi use tiny blackboards for writing and erasing these formulae. Also, suppose that in many cases the homunculi are unable to store the formulae efficiently in memory--perhaps the boards fill up. To alleviate the problem, we might imagine, the homunculi take to storing synonymous, but syntactically simple, substitutes for these formulae. For lack of a better word, call these substitutes "abbreviations". Which abbreviations they choose don't concern us, so long as they can reliably convert between the syntactically complex formulae and the abbreviations.[41] Of course, there is a price for using these abbreviations to save board space: to perform inferences, the homunculi must first convert an abbreviation back into its expanded,

[41]They might adopt a Godel-numbering scheme for the formulae, substituting numbers for formulae in a way which is known to allow for "general, effective, and reliable" conversion processes. On this scheme, rather than storing a cumbersome syntactically complex symbol such as "there is a cup with coffee and white sugar", the homunculi would store some numeral (perhaps in compact, scientific notation). Except by the rare accident, this numeral would not contain any parts corresponding to the parts of the original string--i.e., it would not contain numeric abbreviations for "cup", "white", and so on. This would not, then, count as a concatenative conversion process.

syntactically complex, form. Nevertheless, it is easy to imagine the space savings to be worth the extra effort during inference.

The point of describing this fanciful model is that, although its method of conversion is nonconcatenative, it does support syntactically complex symbols (used in inference), and so is a genuine language-of-thought model. My suggestion is that, at least as far as the fate of the LOT hypothesis is concerned, the connectionist models van Gelder discusses *do not differ* from this model. There are no blackboards and no homunculi in these models, and the tiny English symbols are replaced with distributed symbols--groups of nodes--but the relevant features are the same. The most important shared feature is that, in inference, the nonconcatenated abbreviations in the connectionist models must first be converted into syntactically complex form.

To see this, suppose that a group of nodes *P* is a nonconcatenative but systematically generated abbreviation, with the content *that there is a cup with coffee and white sugar*. What is needed in order for a system to use this stored "belief" in inference, say, to draw conclusions about how to prepare breakfast this morning? Since *P* is nonconcatenative, it is in effect a "blob" whose internal structure may differ radically from that of a symbol with a

semantically "near" content--say, *that there is a can with coffee and white sugar*--and may be nearly identical structurally to a symbol with a "distant" content--say, *Sonny drinks more coffee than Cher*. At a minimum, then, the system would need to be able to regenerate, from *P*, symbols for *cup* (as opposed to some other container, such as a can or a supermarket) and *with coffee and sugar* (as opposed to some other fact about the cup). Once these symbols are regenerated, of course, they must not be allowed simply to "drift apart", effectively losing the information that the coffee and sugar are *in* the cup, but must be used *jointly* in planning breakfast. But to use them in this manner is no different from using them as systematic, symbolic parts of a syntactically complex whole, a *concatenative* symbol with the content *that there is a cup with coffee and sugar*. This complex symbol is the "mereological sum" of the two symbols plus whatever items indicate that their roles are tied to one another. It is syntactically complex because its content is appropriately dependent on that of its symbolic parts (the symbols for *cup* and *with coffee and sugar*).[42] While it is true

---

[42]These points are clearer when we realize that, in the general case, it would not be enough to regenerate only the symbols for *cup* and *with coffee and sugar*. All regenerating these symbols tells the system is that *P* says *something or other* about cups and sugared-coffee containers. This is compatible with *P*'s meaning *that Cher once spilt a cup with coffee and sugar on Sonny*, or something equally irrelevant to the search for this morning's coffee. When more and more symbols must be regenerated, and more and more items are needed to maintain information about their roles with respect to one another, it is increasingly evident that what is regenerated from the stored nonconcatenative symbols are genuinely syntactic formulae. In the next section I support this analysis by

that van Gelder's models present ways of reducing the number of syntactically complex symbols which must be present at any one time, this is also true of the fanciful language-of-thought case. The relevant similarities between the connectionist models and the model with homunculi and blackboards are deep enough, I think, to compel the conclusion that all are alike in making crucial use of a language of thought for inference, if not for long-term storage in memory.

It is possible for van Gelder's models using nonconcatenative storage to be modified so as to avoid their commitment to syntactically complex symbols. Specifically, the models could be redesigned to perform inferences *without* the regeneration (or decomposition) steps. Such "shortcut" inferences are easy to implement in connectionist networks: all that is needed is to build connections between the nonconcatenative symbols themselves, so that *P* (say) might activate some other abbreviation *Q* without the system needing first to decompose *P* or *Q*.

To see the effects of this, consider again the analogy with the homunculi and blackboards. In the original story, given the ability to decompose the abbreviations in inference, the homunculi can engage in "structure-sensitive" processing.

---

considering the possibility that the regenerated symbols might not combine spatiotemporally, unlike genuine syntactic parts.

What this means, at a minimum, is that the homunculi can apply fairly *general* rules--such as, "if I want some *X* and there is some in container *Y*, then find container *Y*"--to a variety of specific representations--such as the belief that there is sugared coffee in the cup, or the belief that there are tools in the hardware store. This is possible because, with syntactic complexity, the homunculi can "match" the rules to the representations in virtue of the form of their parts, binding variables to constants, and the like (see the illustration of production systems in section 0.1.3). If we suppose that the homunculi forego decomposition of the abbreviations into complex counterparts, however, this sort of inference would be precluded.

Instead of being capable of structure-sensitive inference with general rules, the system would need an immense independently stored battery of information about every specific case it represents (e.g., *tea* and *spark plugs* as well as *coffee*, and *bags* and *warehouses* as well as *cups*). Either all of this information would have to be innate, or else the system would have to find out "the hard way" that it should find a cup when some desired sugared coffee is in the cup, and even then it would have no way of applying such information to analogous problems such as what to do when some desired tea is in a bag. As a result, at very best, the system would only perform sensibly in *routine* situations: everyday, highly

practiced activities performed under optimally smooth conditions.[43]

This limited, structure-insensitive sort of inference is known as purely "associative" or "statistical" inference (Hume's "habits of the mind"). As most connectionists agree, it is characteristic of connectionists models which use only local symbols (individual nodes), such as the interactive activation model of reading (see section 0.2.1). A desire to avoid pure associationism is one of the main reasons for the popularity of models with distributed symbols, and in particular for the models which van Gelder discusses. To redesign these models, then, would deprive them of their most attractive feature, namely, their ability efficiently to *combine* structure-sensitive and associative inference. Indeed, if structure-sensitive inference is abandoned, these models lose their advantage over models involving only local symbols. Without inferential processes which are sensitive to the form of nonconcatenative abbreviations, and without processes for converting these abbreviations back into syntactically complex formulae, there is little computational point to using distributed rather than local symbols for the

---

[43]Notice that van Gelder's models would not avoid the LOT hypothesis simply by *adding* direct connections between nonconcatenative symbols, while leaving intact the mechanism for decomposing these symbols to handle nonroutine situations. Rather, the decomposition mechanism must be completely disabled for purposes of inferential processing. Chalmers (1990) appears to miss this point.

abbreviations (in fact, the increased complication carries with it some computational disadvantages such as decreased speed). Cheapened versions of van Gelder's models--ones stripped of their support for syntactic complexity--would have as little hope to be the whole story of cognition as do models with only local symbols. But this is simply to reiterate the dilemma for connectionists posed by Fodor and Pylyshyn--connectionists appear forced to choose *either* to implement a language of thought or else to adopt pure associationism. The point of my argument has been to show that this dilemma remains standing even though Fodor and Pylyshyn's argument falters, and even in the face of the most sophisticated connectionist schemes of representation yet offered.

1.2.4 A More General Perspective

Connectionists have not shown that distributed models can account for systematicity without implementing a language of thought. I suspect, but cannot *quite* show, that my arguments about Smolensky's and van Gelder's models are general enough to be recast into a general argument that no connectionist model can explain systematicity without syntactic complexity. If so, it is natural to ask whether and how it is possible for *any* model to abandon the LOT hypothesis without sacrificing systematicity.

To see how this is possible, consider the following example of a "drinks machine", adapted from Martin Davies (1989).[44] The outputs of the machine are drinks of one of four kinds: coffee or tea with or without milk. The inputs of the machine are tokens ("coins") of four kinds: *A*, *B*, *C*, *D*. The relations between the inputs and outputs are as follows:

A-token in --> coffee with milk out
B-token in --> coffee without milk out
C-token in --> tea with milk out
D-token in --> tea without milk out

Imagine that the tokens, when put in the machine, realize "thoughts" about the drinks they respectively cause the machine to produce (e.g., the thought *that the customer wants coffee with milk*). Given this, we can apply Fodor and Pylyshyn's notion of systematicity, and ask whether there are presupposition relations among these representations.

Suppose that the machine is so constructed that, given that it can think *that the customer wants coffee with milk* and *that the customer wants tea without milk*, it must be able to think *that the customer wants coffee without milk* and *that the customer wants tea with milk*. (Translation: given that the machine can process *A*-tokens and *D*-tokens, it must be able to process *B*-tokens and *C*-tokens.) If (as we are supposing) these presupposition relations are not accidental, then (as

[44]As will become clear, the morals I wish to draw from the drinks machine differ from, and apparently contradict, the morals Davies wishes to draw.

Davies argues) the machine must contain four mechanisms *M1*,...,*M4* of the following sort:

M1 explains coffee delivery given *A*-tokens or *B*-tokens;
M2 explains tea delivery given *C*-tokens or *D*-tokens;
M3 explains milk delivery given *A*-tokens or *C*-tokens;
M4 explains milk nondelivery given *B*-tokens or *D*-tokens.

But if this is the case (as Davies also argues) these pairs of tokens must share distinctive properties to which their respective mechanisms are sensitive. For example, *M1* might respond to roundness, *M2* to squareness, *M3* to redness, and *M4* to blueness. In this case *A*-tokens would have to be round and red (since they trigger both *M1* and *M3*), *B*-tokens would have to be round and blue, *C*-tokens square and red, and *D*-tokens square and blue. Davies argues that the insertion of tokens (i.e., the thoughts) would therefore qualify as syntactically complex states, articulated into shape-related states (representing demand for coffee or tea) and color-related states (representing demand for milk or no milk).

Along these lines, Davies shows how it is possible to argue *a priori* (or, at least, "prior" to psychological research) for the presence of syntactic complexity, given systematicity. If this is right, then it is no surprise that the various distributed connectionist models I have discussed all turn out, on close inspection, to implement a language of thought. However, since the LOT hypothesis is intended to be a substantive psychological explanation of systematicity, there

is reason to resist a construal of the LOT hypothesis which renders it logically (or at least physically) equivalent to the existence of systematicity. The account of syntactic structure offered in section 0.1.3 is stronger than Davies' account in one crucial respect, and can therefore be used to show how it is at least possible to account for systematicity without syntactic structure.

On the account of syntactic structure I have developed, (proper) syntactic parts must be (proper) *spatiotemporal* parts, as (written or spoken) words are spatiotemporal parts of sentences. (In a moment I will illustrate the importance of this requirement.) By this criterion, an object's state of being round is *not* a syntactic part of its state of being round and red, since their spatiotemporal locations *coincide*.[48] It is possible for representational states to have spatiotemporal syntactic parts. The state of activation of a connectionist node *n* at time *t* is, for example, a spatial part of the state of activity of *n* and *n'* at *t*, and is a temporal part of the state of activation of *n* at *t* and *t'*.

---

[48]I suppose that states are instantiations of properties (or relations) by individuals, and I suppose for the moment that a state is where the individuals which "participate" in the state are. This account of state-locations will have to be amended, as I will explain shortly. Furthermore, I ignore the view that being red is a relational state holding between (say) red objects and normal human perceivers, since if this view is true, the drinks machine is presumably responding to some other states of the inserted tokens, the location of which does coincide with the tokens themselves.

This is why the nodes in Smolensky's coffee model (see section 1.2.2) combine to form syntactically complex symbols, for example.

Syntactic complexity appears inevitable if each systematically-related symbol is a combination of states or objects with different spatiotemporal locations. To explain systematicity without a language of thought, then, a representational scheme might be developed in which each systematically-related symbol is a combination of states with the same locations. For short, let us say that such models support only "nonspatiotemporal combination". In Davies' drinks-machine example, this is achieved by treating a systematically-related symbol as a combination of simultaneous states of the same object. I will conclude by considering an analogous strategy for mental symbols: models in which symbols combine only by being simultaneous states of the same object.

Although this strategy may seem easy to implement, in fact it raises serious computational difficulties (which in turn help to explain the appeal of genuine languages of thought, with spatiotemporal combination). Natural languages and genuine languages of thought (including the connectionist models I have considered in the previous two sections) can have an indefinitely large "primitive vocabulary" consisting

of simple symbols (e.g., nodes, patterns of nodes, or strings of letters or sounds). The number of primitive vocabulary elements in English appears to be at least as large as the number of entries in a standard dictionary, and the number of elements in an English speaker's (hypothetical) language of thought appears to be at least as large as the number of English primitive vocabulary elements the person understands. These symbols can enter into fairly large combinations with other symbols by being placed in certain spatiotemporal relations (to which processes are sensitive). For example, easily comprehensible English sentences can run to twenty or more primitive elements (e.g., words), as can sentences in a (hypothetical) language of thought.[46] Furthermore, these combinations are "nonexclusive" in the sense that virtually any pair of primitive elements can enter into some sensible combination or other--as virtually any two words (or ideas) can be combined in some sentence (or thought). It is not at all clear how fairly large, nonexclusive combinations formed from indefinitely large vocabularies can be implemented in a system which supports only nonspatiotemporal combinations.

---

[46]There is a tradeoff between primitive vocabulary size and required combination size. The number of primitive mental elements available to a person might be smaller than the number of primitive linguistic elements he understands, if the person supposes that the linguistic elements are definitionally reducible to a small stock of mental elements. In this case, however, these mental elements would have to be capable of entering into larger combinations than the linguistic elements, to form symbols with the same contents.

In such models, states must "combine" by being simultaneous states of the same object (on pain of not being in the same place). If two such states are mutually inconsistent--e.g., an object's being round and it's being square--then they cannot be combined in this way. If combinations are to be nonexclusive, the primitive vocabulary of such a system must be realized as a stock of mutually consistent states. Furthermore, like all mental symbols, these states must be causally relevant to the behavior of the system. The problem is: how can there be enough such states to go around?

Consider, first, the situation with respect to individual connectionist nodes. Such a node typically has only *two* mutually consistent, causally relevant states at a given moment: its activation level and its threshold (or degree of activation necessary for it to transmit activation to other nodes). If a model's primitive vocabulary elements are realized as particular momentary activation levels or thresholds, and if they are to combine nonspatiotemporally, then at most they can combine two at a time--in mentalistic terms, the most complex thoughts would involve only two ideas. Although a node may change activation level and threshold over an extended time interval, this does not help the situation. A combination of activation levels (or thresholds) of an object, as they exist at different moments, is a syntactic

complex of these states, since it has them as temporal parts."[47] Therefore, local representation can support large, nonexclusive, combinations of primitive vocabulary elements only by supporting (temporal) syntactic complexity.

The analysis of distributed representations is more complicated, as usual. It may appear that groups of nodes are in more causally relevant states at a time, and so may support large combinations of primitive vocabulary elements without supporting syntactic complexity. In particular, it appears that a group of k nodes has $2k$ causally relevant states at a given time, corresponding to the activation and threshold states of each node in the group at that time. For example, consider two nodes $n1$ and $n2$, each with a state of activation and a threshold state. Now consider their mereological sum, the *pair* of nodes. Call this object "Fred". At any time, Fred is in four relevant states: its $n1$-activation, its $n1$-threshold, its $n2$-activation, and its $n2$-threshold. If states are located where their participating objects are

---

[47]Although I have considered only the case in which the vocabulary elements are activation states (and thresholds) of a node at a moment, nothing is changed by generalizing to the case in which the vocabulary elements are patterns of activation states (or patterns of thresholds) of a node over an interval. Since a node can have only one activation level (and one threshold) at a given moment, it can have only one pattern of activation levels (and one pattern of thresholds) over a given interval. If such patterns are to combine nonsyntactically, they also can combine at most two per interval. Finally, a combination of patterns of activation levels (or thresholds) of an object, as they exist over different (possibly overlapping) intervals, is a syntactic complex of these states, since it has them as temporal parts.

located, as I have been supposing, then all of these states are in the same place--namely, where Fred is--and so their combinations do not count as syntactically complex. If we consider larger *n*-tuples of nodes, it might be suggested, we can accommodate large combinations of primitive vocabulary elements without postulating a language of thought.

However, I think we should reject the idea that all states are located precisely where their participating objects are. The central--perhaps, only--explanatory function of the notion of a state's spatial location is to track and explain which other states it can influence in a given time.[46] The states which now exist wholly within my room, for example, cannot influence the states existing next month wholly within a room a light-year away. In many cases we lose this sort of explanation, if we adhere to the principle that states are where their participating objects are. For example, suppose that *n*1 and *n*2 are at such a distance from one another that light travels from one to the other in a positive (but perhaps very tiny) time interval *e*. Then (by special relativity) it is not physically possible for Fred's *n*1-activation at *t* to influence Fred's *n*2-activation at any time before *t*+*e*. If we suppose that these states are located in the same place,

---

[46]Other functions of the notion of a state-location, such as helping to specify where one has to be to witness the state at a certain time, seem to depend on this function.

however, we lose a straightforward explanation of this fact. If, instead, we suppose that Fred's *n*1-activation is where *n*1 is, and Fred's *n*2-activation is where *n*2 is, then we can explain their inability to influence each other in suitably small times. But given any account of state-location which tracks influential potential in this fashion, combinations of Fred's four states would count as syntactically complex--as containing the states as spatial parts. Therefore, connectionist representational schemes restricted to combinations of simultaneous states of an object are unable to support large combinations of primitive vocabulary elements, even if the objects in question are distributed.

Although we can explain systematicity without a language of thought (e.g., by tokens of the drinks-machine sort), the price (in any known connectionist mechanism, at least) seems to be a primitive vocabulary which lends itself only to combinations of extremely small size. In fact, the problem is even more severe. In a genuine language (or language of thought) a symbol can contain multiple "copies" of the same type of primitive vocabulary element (e.g., "John loves Mary more than Sally loves Harry"). There does not seem to be an analogous possibility in systems restricted to nonspatiotemporal combinations: it doesn't make sense for an object to have multiple "copies" of the same state type (that is, at the same time--copies at different times would qualify

as temporal and so syntactic parts of the whole). Without a language of thought, combinations must not only be small, but must also be restricted in combinatorial possibilities (e.g., recursion).

If (as seems likely, although I know of no way to prove it) models with symbols analogous to Davies' drinks-machine tokens are the *only* possible nonsyntactic explanations of systematicity, we would have a fairly general reason to deny that connectionist models are likely to explain systematicity without implementing a language of thought. Importantly, this would be achieved without making it literally impossible to explain systematicity without a language of thought. The argument for the classical framework (and so for models which fall under both the connectionist and classical frameworks) rests on its ability to account for a family of intuitively obvious but fully empirical phenomena, including but not restricted to systematicity, large primitive vocabularies, large nonexclusive combinations of primitive vocabulary elements, and (even limited) recursion.

Chapter 2

CONNECTIONIST CONTENT

If the arguments of chapter 1 are correct, associationist connectionist models (such as ultralocal ones) yield the clearest alternatives to the LOT hypothesis. While it may be that such models cannot provide a general account of cognition, they may account for important aspects of cognition, such as low-level perception (e.g., with the interactive activation model of reading) or the mechanisms which distinguish experts from novices at a given skill (e.g., with dependency-network models). Since these models stand a fighting chance of being applicable to some aspects of cognition, it is important from a philosophical standpoint that we have appropriate tools for understanding such models. In particular, we want to have a theory of the *semantic content* of representations in certain connectionist models. In this chapter, I want to consider the prospects for applying a specific sort of "fine-grained" theory of content to such models.

According to the fine-grained theory I will consider, contents admit of degrees of *complexity*. Even the simplest

propositions (e.g., the proposition *that a is F*) are thought to be complexes of constituent concepts (e.g., the concepts *a* and *F*). What is required for a representation *r* to have a complex content, say, the proposition *that a is F*? On a fairly standard fine-grained conception, *r* must display "semantic dependence" on *other* representations which represent the concepts *a* and *F*, i.e., the constituent concepts of the proposition. What sort of relation between representations counts as semantic dependence? The most familiar examples are syntactic relations: the content of the English sentence "sugar is white" depends on the content of its syntactic parts "sugar" and "white". Another example of semantic dependence might loosely be termed "abbreviation": a syntactically simple symbol "p" in a logical formalism may have a complex, propositional content in virtue of being treated (by convention or otherwise) as "standing in place of" a sentence such as "sugar is white", and so depending semantically on the parts of that sentence.

Although virtually all nonconnectionist models in cognitive science, as well as many connectionist models, postulate relations such as syntactic complexity and abbreviation, many connectionist models appear *not* to support these relations. The best examples are the associationist (e.g., ultralocal) models. Nevertheless, for reasons I will describe there is at least an appearance that representations in these models

do have propositional contents. This generates a philosophical puzzle, at least for those sympathetic to the relevant fine-grained theories of content: how it is possible for a representation to have propositional content without displaying semantic dependence on other representations (e.g., without being either syntactically complex or an abbreviation)? My goal in this chapter is to explain how.

I will begin in section 2.1 by describing the intuitive idea of fine-grained theories, and the puzzle which certain connectionist models present for them. Then in section 2.2 I will attempt to show how the troublesome connectionist representations, even individual nodes, can have propositional content, and will discuss how this propositional content can even be *explicit* content. In the final section, I will briefly discuss some of the metaphysical commitments of the theory I propose.

2.1 Fine-Grained Content and Connectionism

Philosophers explore many distinct conceptions of content, and many distinct purposes for being concerned with content. One reason the contemporary literature is difficult is that it is unclear which conceptions of content are genuine competitors, and which conceptions are simply addressed to

different, but equally legitimate, purposes."[41] To help fix ideas, then, discussions of theories of content should begin with a specification of some guiding reasons for wanting a conception of content. Since I will be addressing a puzzle which arises out of certain connectionist models, I want to focus on a notion of content which is suitable for use in *cognitive science*, as it might (or might not) be opposed to common sense, for example. I will therefore be concerned with the content of *mental* representations of the sort which are postulated in cognitive-scientific models (e.g., physical tokens of functionally-specified propositional-attitude types--see section 0.1.1). I take it that, at a minimum, cognitive science appeals to content to specify these mental representations, and to express generalizations about their functional role--generalizations relating mental symbols not only to external conditions, via perception and action, but also to other mental representations, via inference.

In this section I want to describe two sorts of theories of content which seek to respect these constraints, one which treats contents as coarse-grained, and one which treats them

---

[41]Of course, the fact that contents serve different purposes (even apparently antagonistic ones) does not mean that a representation has different "kinds" of content, as many philosophers are quick to suppose. Often, the urge to postulate different kinds of content is merely a vestige of verificationism. We don't postulate that an object has different kinds of shapes simply because we are interested in shape for various reasons and have various methods (even apparently antagonistic ones) for amassing evidence about shape.

as fine-grained. After introducing the notion of *reference conditions* which is central to both theories (section 2.1.1), I will focus primarily on the notion of *semantic structure* which figures in the fine-grained theories (section 2.1.2). Then it will be possible to present the connectionist puzzle for these fine-grained theories (section 2.1.3).

2.1.1 Content and Reference

In discussing content it is common to begin with the relation between the content of a representation and its *referent*, or the existing entity, if any, that it is about. There is a close connection between content and reference, as is shown by the fact that no entity can have reference unless it has content, i.e., is a representation. I will assume that the referents of mental representations are existing objects, properties, and facts (which, I will also assume, are typically instantiations of existing properties by existing objects). An idea of Paris refers to, or is about, a certain existing city, an idea of prettiness refers to, or is about, a certain existing property, and a belief that Paris is pretty refers to, or is about, the existing fact that this city has this property.[50] I will limit my attention to representations

[50]See Barwise and Perry, 1981, for a defense of this conception of reference against the Fregean argument that sentences (and so, presumably, beliefs) refer not to facts but to (something like) truth values. What I call facts they call "situations" (which are not to be confused with their "abstract situations"). While the Fregean position is viable, the

which refer (or purport to refer) in these ways, and I will call these "referential" representations.

It might be tempting to begin to specify the content of a referential representation by specifying its referent. However, referents are not the whole of content, as is indicated by the familiar fact that not every referential representation even *has* a referent. An idea of my pet caterpillar does not refer, since there is no such thing; nor does a belief that Paris is in Germany refer, since there is no such fact.[91] Furthermore, there is sufficient reason to deny that referents are even *parts* of contents, at least in the case of many successfully referring representations. Suppose that tomorrow I will obtain a pet caterpillar, Crawlette. As a result, the idea of my pet caterpillar will undergo a change in reference, from having no referent today to having Crawlette as a referent tomorrow. Since the phrase

---

appeal to facts is more interesting for my present purposes. I will be considering (and rejecting) the position that contents are identical to referents, and this identification is not even initially tempting on the Fregean view. (In a moment, I will also consider the view that predicative ideas refer not to properties but to groups of objects.)

[91]What about predicative ideas, such as the idea of being pretty? Does this idea refer even if there are no pretty things--refer, say, to the property of being pretty? This raises controversial issues. Can properties exist without being instantiated? What about properties which are logically impossible to instantiate? Since my present point is independent of the proper answer to these questions, I will not to take a stand on them here. It would even be okay for my present purposes if these issues were sidestepped by taking the idea of being pretty to refer, not to a property, but to all and only the pretty things.

will not undergo a change in *content*, this shows that even for many successfully referring representations, content is not to be specified even in part by specifying referents themselves.[52]

Since objects, properties, and facts do not in general serve as the contents of (mental) representations, I want to adopt the standard philosophical terms "propositions" and "concepts" for whatever things *do* fit this bill.[53] One pressing question, of course, is: what *are* propositions and concepts, that they may serve as contents? While a full answer to this question requires consideration of a number of competing philosophical positions, I think we can place three weak, and so uncontroversial, constraints on them, given the discussion so far:

---

[52]This is compatible with there being *some* referential symbols for which content *is* partly or wholly constituted by referents themselves. It would take us too far afield to enter the debate over whether or not there are such cases, however. My opinion is that even what philosophers sometimes call "wide content" need not be specified by referents, but may be specified by reference *conditions*, of the sort to be described below.

[53]It is necessary to ward off certain associations from the psychological usage of the terms "concept" and "proposition". Often these terms are used to mean, not *contents*, but certain mental representations themselves. Although, as I will explain in section 2.3.3, there are ways to reconcile the psychological use of the terms with their semantic use, it should not be assumed from the start that the two uses coincide.

(1) Propositions and concepts should help to explain the fact that only entities with content have reference.
(2) Representations should be able to have propositions and concepts as contents even without having reference.
(3) Representations should be able to retain the same propositions and concepts as contents even through changes in reference.

Before searching for other constraints, a reasonable strategy is to try to find something which fits these three.

To satisfy these constraints, we should focus not on the referent of a representation but on its *reference condition*, or the way the world has to be in order for the representation to be about some existing entity (fact, object, or property). For instance, the reference condition of an idea of my pet caterpillar is that I have a pet caterpillar; if I have a pet caterpillar *x*, the idea refers to *x*, otherwise the idea fails to refer. Similarly, a belief that Paris is in Germany has the reference condition that Paris is in Germany.[54] Reference conditions meet constraints (1)-(3) on propositions and concepts: nothing can refer unless it has a reference condition; representations can have reference conditions

[54]Of course, in the case of beliefs and other symbols with truth value, reference conditions are commonly called "truth conditions". On the matter of terminology, it is perhaps useful to indicate that the distinction between referents and reference conditions is for all practical purposes the same as Carnap's distinction between "extensions" and "intensions" (Carnap, 1947). I shy away from this terminology, however, because I think "intension" is often also used for something more akin to Fregean senses (certain entities which are more finely grained than reference conditions).

without having referents; and representations can retain the same reference conditions even through changes in reference.

Because of these explanatory virtues of reference conditions, it is tempting to *identify* propositions and concepts with reference conditions.[55] On the most natural development of this idea, when we use a phrase or sentence *e* (in English, say) to identify the content of a representation *r*, we are saying that *r* has the same reference condition as *e*.[56] To say that *r* has as content the proposition *that checkerboards are squares* is to say that *r* has the same reference condition as the sentence "checkerboards are squares", while to say that *r* has as content the concept *being a checkerboard* is to say that *r* has the same reference condition as the phrase "being a checkerboard". It follows, on this theory, that if the two English sentences *e*1 and *e*2 have the same reference conditions, then the belief *that e*1 and the belief *that e*2 have the same content.

---

[55]See Stalnaker, 1984 for a defense of a conception of propositions as functions from "possible worlds", or possible ways for the world to be, to truth values. Concepts might be treated in a similar fashion, as functions from possible worlds to reference values (or perhaps to referents themselves).

[56]In everyday attribution, we are satisfied if the reference conditions are *nearly* the same. However, even in everyday attribution we are aware that attributions are more accurate the nearer the reference conditions are. When we mean to speak strictly and accurately, as we might for purposes of doing cognitive science, the ideal *is* sameness of reference condition.

2.1.2 Semantic Structure

Many philosophers resist the identification of contents with reference conditions on the grounds that propositions and concepts must be individuated more finely than reference conditions. The familiar argument for this is that sameness of reference conditions appears not to insure sameness of content. On widespread assumptions, for example, it is logically impossible for the following two mental representations to differ in reference:

Belief *B*1: *that checkerboards are squares.*
Belief *B*2: *that checkerboards are equilateral, right-angled polygons with a number of sides equal to the cube root of sixty-four.*

No matter what way the world is, either both are true or neither are, so they have the same reference conditions. Despite this, there is clear intuitive resistance to the idea that *B*1 and *B*2 are the same belief.

In ordinary cases, we are careful to *distinguish* *B*1 from *B*2. We suppose that people without mathematical training do not have belief *B*2, even if they do have belief *B*1. And we suppose that someone could have *B*2 even if they did not have *B*1, say, if they failed to realize that equilateral, right-angled, blahblahblahs *are* squares. Since reference conditions are not individuated finely enough to reflect these differences among mental states, some philosophers have sought

more finely individuated entities which do reflect these differences. The results are versions of what I will call "fine-grained" theories of content--theories according to which representations can differ in content even though they have the same reference condition, as this notion is normally construed (see, for example, Lewis, 1972; Evans, 1982; Cresswell, 1985). These theories are opposed, naturally, to "coarse-grained" theories according to which representations have the same content if they have the same reference condition (the theory that contents *are* reference conditions is, of course, one such theory).[57]

Intuitively, we have some idea of the respect in which the two beliefs differ: they are somehow "structured" out of different ideas.[58] There is some temptation to read this as

---

[57]I am not pretending that the reflections motivating fine-grained theories suffice to eliminate coarse-grained theories of content. To save a coarse-grained theory--e.g., to maintain the identification of content with reference conditions--one option is to insist that beliefs such as *B1* and *B2 do* have the same content, but differ with respect to some other dimension. I will have a little to say in section 2.2.1 about why the sorts of differences among mental states exemplified here may legitimately be reflected in content. For now, however, I am willing simply to assume and develop a fine-grained theory, for the sake of expressing the puzzle presented by connectionism. Of course, this dispute about what should and should not count as part of content may in the end be no more than a terminological issue, or else one to be finessed by adopting the verificationist's trick of treating the word "content" as referring ambiguously to various "kinds" of content.

[58]It is possible to adopt a fine-grained theory of content without appealing to a notion of semantic structure (Block, 1986). Such a view would not be faced with the connectionist puzzle to be described below. I will give fleeting attention to the view in section 2.3.2.

a claim about the *syntactic* structure of the physical representations which help to realize the attitudes and ideas--i.e., as the claim that these representations stand in literal whole/part relations. However, the claim is best taken as one about *semantic* structure, where this is to be explained in a nonsyntactic fashion. The reason is that it seems possible for there to be syntactically simple representations--ones without other representations as parts--which nonetheless are "structured" in the relevant sense. One way for this to happen is for a syntactically simple representation to be introduced as an "abbreviation" of a syntactically complex representation. It is best, then, to look for a nonsyntactic way of explaining semantic structure.

Philosophers sympathetic to fine-grained theories standardly express the difference between the two beliefs about checkerboards as follows: one can't have belief *B2* unless one has ideas of *cube roots*, *polygons*, etc., but one can have belief *B1* even without having these ideas. In other words, the two beliefs are in some sense dependent on the availability of different mental representations. Intuitively, the latter belief, unlike the former one, depends on the availability of a representation with the content *being a square root*, for instance. Also intuitively, it should be possible for a representation to depend on representations which are not its parts, perhaps as a building "depends" on

the ground below it. Therefore, dependence may be a first step toward understanding the notion of semantic structure without relying on claims of syntactic structure.

Of course, these intuitive appeals are no substitute for a philosophical account of what it is for one representation to be dependent on another, in the relevant sense. Although the connectionist puzzle for fine-grained theories is somewhat independent of the particular account adopted, it is necessary to have one on the table in order to describe the puzzle. Perhaps the foremost task of an account of dependence is that of avoiding a certain extremely "holistic" conclusion. Much of human inference appears to be holistic, in the sense that virtually any two premises can be combined in some rational inference or other. Given this sort of psychological holism, (virtually) every representation is such that its "functional role" depends on (virtually) every other representation (available to the thinker). The functional role of belief *B*1 *that checkerboards are squares* depends, in this way, on whether or not the thinker has representations for *Aristotle*, for *being polygonal*, etc. If this is the relevant kind of dependence, then fine-grained theories cannot distinguish representations (such as the two beliefs about checkerboards) in terms of their dependence relations. They would depend on the same set of representations, namely, (virtually) every one.

As a first step towards avoiding this result, I want to try to get clearer about *what it is* about a representation *r* which must be dependent on the availability of other representations, in order for *r* to be semantically structured. While psychological holism might show that *some* facts about a representation--such as its functional role--are dependent on (virtually) every other representation, it may be possible to identify some content-relevant facts which are not subject to extreme holism. A natural idea is that *r* is semantically structured iff *r*'s *content* (rather than its functional role) is dependent on that of representations with other contents. Since semantic structure is taken to be a determinant of content, this is indeed a claim which the fine-grained theorist wants to make. However, given that the fine-grained theorist is trying to *explain* content (at least partially) in terms of semantic structure, it is circular to explain semantic structure in terms of content. In specifying the relata of the relevant dependence relation, the fine-grained theorist needs to specify facts which, though relevant to content, are identifiable independently of content. Here the most natural strategy is to appeal to *reference conditions*.[59]

---

[59]Carnap's (1947) notion of intensional isomorphism also embodies dependence relations among reference conditions (i.e., his "intensions"). Similar positions are defended by Lewis (1972) and Cresswell (1985). In section 2.3 I will attempt to develop a metaphysical account of contents which is different from the account offered by these later authors.

On fine-grained theories, although reference conditions are clearly relevant to content, one can specify a representation's reference condition independently of its content. How might this suggestion be spelled out so as to avoid the holism problem?

There is one preliminary to spelling out this suggestion. While the fine-grained theories under consideration maintain that the reference condition of a semantically structured mental representation *r* depends on the subject's having *some or other* representation with a different reference condition *C*, they do not necessarily maintain (and probably should not maintain) that *r*'s reference condition depends on the availability of any *particular* token representation with reference condition *C*. It will help, then, to introduce a notation for grouping representations according to their reference condition. If a mental representation has reference condition *C* (whatever *C* may be), call it an *C*-referrer. The suggestion on behalf of fine-grained theories, then, is that a thinker's mental representation *r* is semantically structured out of *C*-referrers iff *r*'s reference condition is dependent on some *C*-referrer or other's being available to the thinker. In this case, I will say that the representation is "semantically dependent" on *C*-referrers. On the fine-grained theories I will consider, the content of a representation not only reflects its own reference condition, but also reflects

the reference conditions of other representations on which its reference condition depends.

As for holism, while the precise functional role of a representation may depend on that of (virtually) every other representation, its reference condition may depend on that of only a restricted set of other representations. We can display this possibility more clearly by explaining the relevant kind of dependence in terms of a certain kind of counterfactual claim. In speaking of a representation *r*'s reference condition as being dependent on *C*-referrers, I mean: were the *C*-referrers actually available to the thinker to have a different reference condition, then (as a result) *r* would have a different reference condition (holding fixed the actual functional relations between *r* and these *C*-referrers). Imagine someone with a mental representation *B*1, the realization of a belief that checkerboards are square. Suppose he also has available to him mental representations for checkerboards, for being square, for Aristotle, and for being polygonal. Intuitively, if the checkerboard-referrers or the squareness-referrers were to have different reference conditions--e.g., by being the effect of different perceptual states, of different phenomena in the world, etc.--then (holding fixed the functional relations between these representations and *B*1) *B*1 would also have a different reference condition. However, intuitively, even if the

Aristotle-referrers and the polygonality-referrers were to have different reference conditions, *B*1 might still have the same reference condition (holding fixed any functional relations between *B*1 and these representations).

Of course, these intuitive assessments of the counterfactuals might not be borne out by a proper philosophical theory of reference conditions: a theory which determines which particular representations have which particular reference conditions. Philosophers differ over what factors help to determine the reference conditions of a representation--its covariation with certain conditions, its functional role, its adaptational role, etc. This is a widely discussed and widely open question which would take considerable resources to address properly. I will not even begin to do so here, since the question does not appear relevant to the choice between fine-grained and coarse-grained theories of content. Defenders of coarse-grained theories have to address this question as well, since they, too, use the notion of reference conditions in understanding content. What I am concerned to show is that, contrary to a common opinion, holism does not raise more of a problem for fine-grained theories than it does for coarse-grained theories.

Suppose that (contrary to intuitive judgements) reference conditions turn out to be subject to an extreme holism, in that which reference condition a given representation has depends on which reference conditions (virtually) every other representation has. If so, then reference conditions would themselves be "fine-grained" in the sense which matters to defenders of fine-grained theories. For example, suppose that *B*1--a belief intuitively specified as the belief *that checkerboards are squares*--is held by someone without special mathematical training, and that *B*2--a belief intuitively specified as the belief *that checkerboards are equilateral, right-angled polygons with a number of sides equal to the cube root of sixty-four*--is held by someone with special mathematical training. If reference conditions are holistic, then (contrary to intuitive appearances) these beliefs have different reference conditions, as fine-grained theories would have it. On the other hand, if reference conditions are not subject to extreme holism, then we can make fine-grained distinctions among the two beliefs in terms of *which* representations their reference conditions depend on.[60]

---

[60]Although I have tried to sketch a particular strategy for avoiding the extremely holistic conclusion, it is enough for my present purposes simply to proceed on the assumption that there is *some* way of doing so. Neither the connectionist puzzle nor my solution to it depends on the particular strategy adopted.

To avoid extreme holism is not yet to claim that there can be *simple* representations, i.e., representations which are not semantically dependent on representations with other reference conditions. Even if a representation does not depend on *every* other representation, every representation might still depend semantically on *some* other representations. Every representation might be part of some (small or middle-sized) circle of mutually dependent representations, which would make a sort of "local holism" true. Perhaps the simplest representations are not, for this reason, absolutely simple. The connectionist puzzle arises independently of whether or not a fine-grained theory is committed to (absolutely) simple representations. It also arises independently of which sorts of representations are thought to be "simplest". Are some ideas in cognitively central systems (such as "checker" and "board") as simple as representations get, or are they in turn semantically structured out of representations in cognitively peripheral systems (e.g., the initial stages of vision)?[61]

---

[61]There may be some question as to how "central" and even partially "theoretical" ideas such as those referring to boards (or to caterpillars, or to quarks) could turn out to be simple, given their functional dependencies on peripheral and observational symbols. The reason they can (at least in principle) be simple is that their reference conditions might not depend on the reference conditions of *particular* types of peripheral symbols, where these types are individuated according to reference condition. Suppose a thinker's mental symbol "board" is triggered in part by certain symbols in his visual system which respond to rectangles presented under certain conditions. Even if these symbols were to have a different reference condition--e.g., even if they were to respond to circles instead of rectangles--it could be that "board" would have the same reference condition, if it has sufficiently strong causal connections to *other*, independently triggered, symbols (such as those responding to certain kinds of tactile pressure, or certain kinds of spoken

Since in presenting and addressing the puzzle I will mainly use examples of extremely "peripheral" connectionist representations, I will leave open this question about "central" representations.[62]

To frame the connectionist puzzle, the most important point about semantic structure (however it is ultimately to be construed) is its relevance to content ascriptions. In the previous section, I described a coarse-grained theory which claims that when we identify the content of a representation *r* by using an English phrase or sentence *e*, we commit ourselves to the claim that *r* has the same reference conditions as *e*. On the most natural development of the present fine-grained theory, we commit ourselves to *more* than sameness of reference conditions. We also commit ourselves to the claim that *r* and *e* have the same semantic structure--i.e., that *r*'s reference condition semantically depends on the availability of representations with the same reference conditions as those on which *e*'s reference condition depends.[63] For example, suppose "A B C" is a syntactically complex

---

communications from other people). (Again, this intuitive result might or might not be borne out by a philosophical theory of reference conditions.)

[62]Eventually, I will also have to say something about what sorts of entities propositions and concepts are, on the fine-grained theory (see section 2.3).

[63]At least, this is so when we mean to speak strictly. See footnote 56.

sentence which is semantically dependent (only) on the representations "A", "B", and "C".[64] Then for *r* to have the content *that A B C*, not only must *r* have the reference condition of "A B C", but this fact must depend (only) on representations with the same reference conditions as "A", "B", and "C". On this account, even if the sentences "A B C" and "D E F" have the same reference conditions, the belief *that A B C* and the belief *that D E F* might have different contents.

2.1.3 The Puzzle

Fortunately, the puzzle which connectionism presents for fine-grained theories of content can be described without focusing on intricate details of connectionist networks. We can display the problem by returning to the interactive activation model of reading (see section 0.2.1).[65] If nodes in connectionist networks are to be genuine representations, they must have semantic contents. For example, consider a particular node in the reading model, a feature node *f* the role of which is to become activated as a result of visual

---

[64] I am not using these letters as variables ranging over English formulae; rather, for purposes of illustration I am pretending that they *are* English formulae, arbitrarily chosen.

[65] I will use examples of local representations to drive the discussion, although I will indicate the respects in which the points generalize to distributed representations.

presentations, in word-initial position, of shapes like that of the top half of a circle, such as form the top of the letters "C", "G", "O", "Q", and "S". If *f* is a representation, it must have content, and a natural idea is that it has the content *that a top-half-of-a-circle-shaped figure is being visually presented in word-initial position.* What does it mean to say that *f* has this content? Well, on the coarse-grained theory of content presented in section 2.1.1, it means that *f* has (at least nearly) the same reference conditions as the English sentence "A top-half-of-a-circle-shaped figure is being visually presented in word-initial position". Call this sentence *e*. Then, on the coarse-grained theory, the claim is that sentence *e* and node *f* are true (or "veridical", or "faithful to the input", etc.) under thc same conditions. The situation is quite different for fine-grained theories, however.

On any fine-grained theory, specifying reference conditions is not sufficient for specifying content, since two representations may differ in content even if they share reference conditions. Furthermore, on any fine-grained theory, even if the English sentence *e* does adequately express the reference conditions of the feature node *f*, it is completely inappropriate to take *e* seriously as a specification of *f*'s content. On the fine-grained conception of content ascriptions I mentioned in the previous section,

to say that *f* has the content *that a top-half-of-a-circle-shaped figure is being visually presented in word-initial position* is (at least) to say that *f* is semantically dependent on representations for *halfness*, *circularity*, etc. However, *f* is simply triggered by an array of visual stimulations, independently of the influence of general knowledge about halfness, circularity, etc. In fact, the interactive-activation model has no knowledge of circles at all, and does not in any sense have a representation for *being a circle*. Therefore, on the fine-grained theory, *f* can't represent presented figures as being top-half-of-a-circle-shaped. Of course, it might simply be suggested that we have misidentified the content of *f*. To avoid wrongly attributing ideas of *being half*, *being a circle*, etc, we might even *draw* what we mean, saying that *f* represents visually presented figures in word-initial position as *being ∩-shaped*. But this still leaves the problem. The revised suggestion is that *f* has the content *that a ∩-shaped figure is being visually presented in word-initial position*. Even this content ascription requires the system to represent the (rather sophisticated) general concepts of *shape*, *vision*, etc. How can *f* have this content if the system lacks representations of these concepts? And if we can't take *such* ascriptions seriously, what ascriptions *can* we take seriously?

A defender of a fine-grained theory might simply respond that there *aren't* any ascriptions which we can take seriously. If verificationism is correct, it might follow from this response that such nodes have no content at all. Since it is established practice in cognitive science to treat such nodes as representations, however, this conclusion would weigh heavily against fine-grained theories of content. Even if verificationism is incorrect, and no such strong conclusion follows, the resulting notion of content would be unsuitable in connectionist theorizing. If there aren't any fine-grained ascriptions which we can take seriously, then there aren't any descriptions or generalizations involving fine-grained contents which we can take seriously, at least where certain individual nodes are concerned. This situation would be an embarrassment to any philosopher who has sympathy with the project of developing a notion of content which is suitable for use in cognitive science. To fulfill this project, it might seem that we need to abandon fine-grained theories in favor of coarse-grained theories, at least in the case of the troublesome connectionist representations. I hope to show how a fine-grained theorist can solve the connectionist puzzle without conceding ground in this fashion."

"Similar worries have appeared in other guises in philosophy, so the present puzzle may be better understood once these connections are drawn. I have in mind Donald Davidson's worries about content ascriptions to non-human animals, and Stephen Stich's worries about content ascriptions to small children and the mentally infirm (Davidson, 1984; Stich, 1984). While we have some inclination to treat animals of other species as having thoughts, we are easily persuaded that our content ascriptions are

It is possible that a defender of a fine-grained theory would show no surprise at the unavailability of serious content ascriptions for connectionist nodes such as feature nodes in the interactive-activation model. This might be explained by appeal to a familiar theoretical construct of fine-grained theories, namely, *simple concepts*. On many fine-grained theories of content, propositions are complexes of concepts, and these concepts are, in turn, either simple or else themselves complexes of simpler concepts. Opinions differ about which concepts are the simple ones. Some think many syntactically simple words in natural languages--words such as "caterpillar" and "board"--express simple concepts. Others think that simple concepts are expressed only by more purely observational representations, such as might be involved in early visual processing. We can abstract away

---

inaccurate. It is natural to suppose that a dog is capable of thinking that the person who normally feeds it is nearby. But at best, this specifies the reference condition of the thought, not its fine-grained content--is it so natural to suppose that the dog also has ideas of *personhood*, or of *normality*? Further attempts to hone in on the content by making substitutions for "person" and "normally" only make the problem worse, eventually appealing to general ideas of *perception*, or *time*, etc. As Stich emphasizes, the situation is the same for a child or a severely mentally handicapped person--consider their apparent thoughts about the people who normally feed *them*, and the bulk of their other thoughts. Both Davidson and Stich favor the conclusion that these apparent thoughts have *no* contents at all, except perhaps as a matter of our conventions. Even if we resist this extreme conclusion, however, the fine-grained theories under consideration are in the embarrassing situation of being unable to specify the content of the thoughts. My solution to the connectionist puzzle will also apply to these problems, although I won't trace out the connections.

from such disagreements of detail to isolate a relevant feature which is supposed to be shared by all representations which express simple concepts.

In ascribing content to such a representation, often the best we can do is to describe the referents of the representation. However, in doing so we generate the same kind of puzzle as that generated by the connectionist nodes: we normally employ concepts which need not be expressed by someone who grasps the target concept. Suppose, for illustration purposes only, that "caterpillar" in fact expresses a simple concept. If asked to say what "caterpillar" means, we might say that it picks out some furry worm-like animals which turn into butterflies. Given fine-grained strictures on content ascription, however, this would not be a serious specification of the *concept* expressed by "caterpillar", since (we are supposing) a representation could express that concept even in the absence of representations for *worm*, *animal*, *butterfly*, and so on. In such a case, the best we can do by way of a serious specification of fine-grained content is to repeat the representation: hence those inclined to think of "caterpillar" as expressing a simple concept often identify this concept as "the concept *caterpillar*". Given that at least some fine-grained theories are already prepared to countenance simple concepts, this device might be used to account for the

similar difficulties arising with respect to representations in early perceptual processing. (In fact, even someone who does not think that high-level representations such as "caterpillar" express simple concepts might be tempted to accept this idea for some nodes in connectionist models of low-level perceptual processing.)

The problem with this strategy is that some--in fact, most--contentful nodes in connectionist networks are interpreted *propositionally*. This precludes the general strategy of attributing to individual nodes the sorts of contents--namely, simple concepts--which some assume to be possessed by familiar syntactically simple words. If we are to find precedents for individual nodes with propositional content, we must look elsewhere. There *are* familiar syntactically simple representations which are propositional, as Robert Cummins notices:

> [I]t is perfectly obvious that a symbol can have a propositional content--can have a proposition as its proper interpretation--even though it has no syntax and is not part of a language-like system of symbols. Paul Revere's lanterns are a simple case in point. (Cummins, 1989, p. 18)

Examples of propositional but syntactically simple symbols which are parts of public language are the words "Mayday" and "Roger", familiar from telecommunications, or the symbol "p", which might in a particular logical notation be used to mean that Paris is pretty. Even without assuming that *all* public

symbols semantically depend on mental representations, however, it is at least intuitively plausible that all of *these* symbols are semantically dependent on other representations (whether mental or linguistic) in a way which reflects our complex specifications of their content. Intuitively, if Paul Revere's word "land" (or corresponding idea) had referred to helicopters, then the flashing of a single lantern would have been true if and only if the British soldiers were moving by helicopter, rather than by land (holding fixed the functional relations between his word and the lanterns, including in particular those which resulted from his conventions). As far as I know, *every* propositional representation in natural language--or in any form of public communication, for that matter--is semantically dependent on other representations in a similar fashion. This is the sense in which the commitment to connectionist nodes which are propositional without standing in semantic dependence relations is without obvious precedent. The philosophical challenge for fine-grained theories is to show how it is possible for syntactically simple representations without semantic dependencies to be propositional, or else to show why this is impossible.[67]

---

[67]The challenge is made more important by the fact that it is not restricted to individual nodes in connectionist networks. Many other connectionist models treat *groups* of nodes as symbols, but also do not admit of syntactic complexity or any similar relation of semantic dependence. Symbols in these models give rise to the same difficulty. Also, given that the puzzle arises from lack of semantic dependence, we can see why many features of connectionist networks are irrelevant. For

2.2 Simple Propositions

To meet this challenge, I want to show how a fine-grained theory of content can appeal to *simple propositions*. This can and should be done without prejudging the issue of which *particular* representations (if any) express simple propositions, just as we can abstract from disagreements over which representations (if any) express simple concepts (see section 2.1.2). If some propositions are simple, presumably this is because they share certain features with those *concepts* which are simple. A natural idea, given fine-grained theories' appeal to semantic dependence, is that simple concepts are those which may be represented by representations *without* semantic dependence on other representations. To return to an example from the previous section, it is sometimes held that "caterpillar" can mean *caterpillar* even if it were to become dissociated from other concepts such as *animal*, *butterfly*, and so on. Perhaps there are no other representations in particular on which the reference condition of "caterpillar" semantically depends. If so (and we are only

---

example, their being "hardwired" drops out. The problem would arise in the same way for models of cognition which, for example, postulated tiny people who read (only) syntactically simple symbols off of tiny monitors, compare them with similar patterns--rules?--in tiny library books, and write new ones onto the monitors according to what they find in the books. Most of the discussion, therefore, will be applicable not only to connectionist nodes but also to any symbols which appear to be propositional despite the lack of semantic-dependency relations.

imagining this for sake of illustration), on the present theory, this would mean that "caterpillar" has as content a simple concept. I suggest that we extend this general idea to the case of *propositions*, so that simple propositions are those which may be represented by representations which do not semantically depend on any representations with other contents. On this account, any connectionist node or other representation which satisfies this condition represents a simple proposition (again, the question of which ones do and which ones don't depends on the operative notion of semantic dependence).[66]

Although this idea is simple to state, it must be defended against several objections. Providing this defense will occupy me in this part of the chapter. My first aim is to show that simple propositions are consistent with the distinction between propositions and concepts (section 2.2.1). Then I want to distinguish the present proposal from one based on *de re* attributions of content (section 2.2.2). Finally, I will try to show how simple propositions can be represented explicitly by connectionist nodes (section 2.2.3).

---

[66]Toward the end of this section I will show how to extend a similar treatment to nodes which, though not genuinely simple (due to their involvement in local holisms), also give rise to the puzzle for fine-grained theories.

2.2.1 Propositions and Concepts

One potential worry about countenancing simple propositions is that doing so threatens the distinction between propositions and concepts. While not much seems to have been written about the distinction, it is at least initially plausible that propositions are to be distinguished from concepts by virtue of distinctive facts about their *structure*. If some propositions are taken to be simple, then, what would distinguish them from simple concepts? I will approach this question indirectly, by asking the question: which representations are supposed to have propositional content, and which representations are supposed to have merely conceptual content? It is easy to distinguish propositional representations from merely conceptual representations, *case by case*. The belief that Paris is pretty--or the wish that it were--has propositional content. By contrast, the idea of Paris, or of being pretty, has merely conceptual content. It is more difficult to formulate a semantically interesting *principle* by which the cases are to be distinguished.

A first stab at it is to say that propositional mental representations are those, like beliefs, which have a truth value (truth, falsity, and, if possible, undefined truth value). However, there are other sorts of propositional mental representations--desires, emotions, etc.--which have

no truth value. To cover these cases, one might try to use some more general notion instead of truth value, such as "success value": beliefs are successful if true, desires if satisfied, joy if appropriate, etc. It might then be said that propositional representations are those with success values. Without some principle of generalization from truth to success, however, this suggestion leaves unclear why the "reference values" of ideas aren't *also* success values: why isn't the idea of Paris successful since it refers to an existing object, and the idea of Atlantis unsuccessful since it fails to refer to one? A different suggestion might be that propositional representations are (or at least purport to be) about *facts*, as opposed to mere objects and properties. This also is inadequate. The idea of my favorite fact does purport to be about a fact; however, it is not a propositional representation but instead a conceptual one.

It is a striking fact that ideas are the only apparent examples of merely conceptual mental representations. However, it would not be correct simply to say that ideas have conceptual content while other mental representations--the propositional attitudes--have propositional content. As I argued in section 0.1.2, representationalism is committed to the existence of *propositional* ideas (and what I called propositional "symbols"--for present purposes these can be lumped with ideas). Nevertheless, by first characterizing the

difference between attitudes and ideas, as I will show, we can go on to distinguish propositional ideas from merely conceptual ideas. So the strategy I adopt will be undertaken in two steps.

As I suggested in section 0.1.2, propositional attitudes are those mental representations which, unlike ideas, standardly function as *units of reasoning*. Such representations have *rationality values*, i.e., degrees of rationality or irrationality, which can influence the rationality values of other representations or actions, or at least be influenced by other representations or perceptions. A belief that Paris is pretty--or a wish that it were--has a rationality value. By contrast, a mere idea of Paris--or of prettiness or the present time--is neither rational nor irrational. Nor is a *propositional* idea (e.g., *that Paris is pretty*) itself rational or irrational. It is hard to see how it could have a rationality value, since (by representationalism) it plays the same role in the belief *that Paris is pretty* that it does in the doubt *that Paris is pretty*, the same role in the hope *that Paris is pretty* that it does in the fear *that Paris is pretty*. Beyond drawing this connection between propositional attitudes and rationality values, I have very little to say about the proper conception of rationality values. I imagine that, at a minimum, having

rationality values is corequisite with having a role as a (potential) premise or conclusion of inference.[69,70]

---

[69]What is less clear is whether there is any way to distinguish inferential relations from non-inferential relations among symbols (e.g., association of ideas), short of appealing to the rationality values of the symbols. Nevertheless, and incidentally, we can use these ideas to defend the claim that some symbolic nodes in connectionist networks have propositional contents, rather than merely conceptual contents. This claim does not rest solely on the (legitimate enough) grounds that it is established scientific practice to interpret nodes propositionally. If a node is thought to be a symbol at all, it must have either propositional content or merely conceptual content. Now, suppose I am right that if a symbol has a rationality value, or figures as a premise or conclusion of inference, then it has propositional content. Therefore, if connectionist networks are to fulfill their appointed task of helping to explain certain rationally evaluable mental processes--e.g., the "tacit inference" involved in perception--they must contain some units of reasoning, i.e., some propositional symbols. Therefore, the only way for *all* symbolic nodes to be merely conceptual would be for them to *combine* to form propositional symbols. Not any methods of combination will do, however: just as not all collections of words are sentences, so not all collections of conceptual symbols are propositional symbols. A mechanism of syntactic combination, or something close, seems to be required. However, many connectionist models do not contain suitable mechanisms of combination. Given this, at least some of the nodes in these models must have propositional content if they are to be symbols at all.

[70]Incidentally, the present conception of the difference between propositions and concepts can be used by defenders of fine-grained theories to respond to a certain kind of objection on behalf of coarse-grained theories. A defender of a coarse-grained theory can reasonably ask why features of functional role--such as those which distinguish the two beliefs about checkerboards in section 2.1.2--should be reflected in content. After all, not *every* difference in inferential role counts as a difference in content. For example, the inferential role of a belief may change as it becomes *more strongly held*, but its content does not. Nevertheless, there are things which can be said in defense of considering functional role as being relevant to content. For one thing, some elements of functional role are *already* built into content, even on coarse-grained theories. Suppose my account of the difference between propositional symbols and merely conceptual symbols is even roughly right. If so, then the very difference between having propositional content and having merely conceptual content tracks a rather important difference in functional role: only symbols with propositional content have *rationality values* at all, so only they can figure as premises or conclusions of rational inference. Given this, there can be no *general* prohibition against building into content differences in functional role.

On, then, to the second step: distinguishing propositional ideas from merely conceptual ones. Here we can simply say that propositional ideas are those ideas which help to realize propositional attitudes without combining with other ideas. An idea *that Paris is pretty*, taken alone (i.e., without other ideas), can help to realize a propositional attitude (e.g., the belief *that Paris is pretty*). But an idea of Paris--or of prettiness, or my favorite fact--must cooperate with other ideas to realize a propositional attitude.

These results can easily be turned into an account of the difference between concepts and propositions which is fully compatible with the postulation of simple propositions. Propositions are those contents which are possessed either by representations (e.g., beliefs) with rationality values, or else by representations (i.e., propositional ideas or propositional symbols) which taken alone can help to realize them. Concepts are those contents which are possessed by representations (e.g., ideas) which must cooperate with other ideas to realize representations with rationality values. This is, I submit, all that is essential to a content's being a proposition rather than a concept, or a concept rather than a proposition. In particular, structure is inessential to the distinction. Since we can draw this distinction (at least in theory) even in the case of simple contents, we can postulate

simple propositions while maintaining a distinction between them and simple concepts.

2.2.2 Simple Propositions and *De Re* Attribution

Another potential worry about countenancing simple propositions is that doing so would introduce an element of coarse-grainedness into fine-grained theories. A natural question to ask is: if there are simple propositions, how are they to be expressed? Any *that*-clause we use will misleadingly commit us to postulating semantic dependencies where there are none. Recall the example of the interactive-activation model of reading, and in particular the feature node *f* (see section 2.1.3). One temptation was to say that *f* has as content the proposition *that a top-half-of-a-circle-shaped figure is being visually presented in word-initial position*. The trouble is that this erroneously attributes to the network representations which represent the concepts of *being a half*, of *vision*, etc. While it is possible to use such *that*-clauses along with disclaimers--e.g., "I hereby disavow the offensive commitments which normally attach to the claims I am making"--there is something vaguely dissatisfying about doing so.

One way to express this dissatisfaction is with the following question: what is the difference between using a

complex *that*-clause along with a disclaimer, and simply engaging in what philosophers call *de re* attribution? Even defenders of fine-grained theories admit that there are some cases of content attribution for which reference conditions are all that are relevant. Suppose, for example, that you and I are making dessert in my kitchen, and I think to tell you about my mother's belief that sugar is bad for the teeth. Pointing to the sugar jar, I might say "You, know, Mom thinks the stuff in this jar is bad for the teeth". Since I know that my Mom has never even *heard* of the jar I am pointing to, however, I only intend my content ascription as a *de re* specification of the *reference conditions* of her thought. Similarly, someone can use the apparatus of *de re* attribution to describe the feature node *f* as "meaning that a top-half-of-a-circle-shaped figure is being visually presented in word-initial position". Since defenders of fine-grained theories do not take *de re* attributions seriously as specifications of content, however, this would not be taken as a genuine answer to the question: what is *f*'s content? But by the same token, the postulation of simple propositions may seem to add nothing to the arsenal of fine-grained theories if we cannot express them without invoking *de re* disclaimers.

It may well be impossible for us to *express* simple propositions with ordinary complex that-clauses. But this would not prevent us from *specifying* or *describing* such a

proposition, simply as the (unique) simple proposition associated with such-and-such reference condition. This strategy is importantly different from employing *de re* attribution. It will help us to see the difference if we employ a notational device for shortening this sort of specification to take the form of a that-clause, while marking disclaimers of normal commitments to semantic dependence.

As I mentioned in section 2.1.2, saying that a representation *r* has the content *that A B C* normally carries the implication that *r* is semantically dependent on representations with contents *A*, *B*, and *C*.[71] To bracket this implication, we might bracket the representations used to specify the content. Thus, we might say that a representation *r* has as content the simple proposition *that [A B C]*. This would mean that *r* has the reference conditions of "A B C", without having the same *content*, i.e., without having the same set of semantic dependencies. To return to the feature-node example, we can express the content of *f* by saying that it means *that [a top-half-of-a-circle-shaped figure is being visually presented in word-initial position]*. This is merely a notational shorthand for saying that *f* has as content the proposition specified by (i) the reference condition of "a top-half-of-a-circle-shaped figure is being presented in

---

[71]As before, I am pretending that "A", "B", and "C" are English formulae, rather than treating them as variables ranging over English formulae.

word-initial position" and (ii) the null set of semantic dependencies.[72]

The brackets do signify a sort of attribution similar to *de re* attribution, in the sense that expressions with the same reference conditions may be substituted within brackets without a change in attributed content. However, such attributions differ from *de re* attributions in an important way: rather than signifying that *r* has *some content or other* associated with the reference conditions of "A B C", the brackets signify that *r* has a quite *specific* such content, namely, the unique simple one. Unlike ordinary *de re* attributions, therefore, we can take such attributions seriously, and literally, as precise specifications of fine-grained content. Moreover, the content so specified is genuinely fine-grained. It is true that any two representations with simple content have the *same* content if and only if they have the same reference condition. This is not a violation of fine-grainedness, however, since it does not mean that *every* pair of representations with the same reference condition have the same content. In particular, representations with simple contents do not have the same

[72]The proposed reference condition of *f* is intended simply as an illustration. The claim being illustrated is compatible with any of the main philosophical theories of reference conditions, which may assign different reference conditions to symbols such as *f*. The present claim may be put as follows: *whatever* the reference condition of *f*, the content of *f* is the simple proposition associated with that reference condition.

content as representations with complex contents. By contrast, countenancing simple propositions furthers the aims motivating fine-grained theories of content. Doing so *increases* fine-grainedness, by increasing the stock of contents which may be specified."

2.2.3 Explicit Content

There is another objection available to someone who wishes to deny the existence, or at least the utility, of simple propositions. It is reasonable to require that, to be a

---

"Given the bracketing convention, and the theoretical apparatus for which it is a shorthand, we are equipped to deal with symbols which exhibit the connectionist puzzle, but which are not absolutely simple (perhaps because they enter into a locally holistic circle of symbols). We can use the bracketing convention to describe the content of many such symbols, specifying propositions which are not absolutely simple, but which nevertheless are not expressed by an unbracketed *that*-clause. To do this, we might place brackets around only *part* of a content-specification, saying for instance that a symbol *s* has as content the proposition *that [A B] C*. Since "C" is unbracketed, this means that *s does* depend on a symbol with the same content as "C", and so is not absolutely simple. However, it means that *s* does not share the dependencies of "A B", namely, dependencies on symbols with the same content as "A" and "B". Instead, it would mean that *s* depends on a symbol with the same reference condition (but not the same content) as "A B". In the same vein, we might attribute the content *that [A B] [C]*, where the brackets around "C" signify that *s* depends on a symbol with the same reference condition as "C", but does not also depend on the symbols which "C" depends on. (In all these cases, as usual, it is being said that *s* has the reference condition of "A B C".) This strategy of using the bracketing convention may not always be applicable, since there may not always be a specification of a symbol's reference condition which admits of a pattern of bracketing reflecting the symbol's systematic dependencies. In these cases, however, we can always resort to a direct listing of these dependencies.

representation, an entity must have a content *explicitly*."[74] For a simple proposition to be the content of a connectionist node, then, sense must be made of the notion of representing such propositions explicitly. On some views, this is not possible. Dan Dennett, for example, has suggested that syntactic complexity is necessary for explicit representation:

> Let us say that information[75] is represented *explicitly* in a system . . . only if there actually exists in the functionally relevant place in the system a physically structured object, a *formula* or *string* or tokening of some members of a system (or "language") of elements for which there is a semantics or interpretation . . .. (Dennett, 1987, p. 216)

If Dennett is right, then there can be no connectionist nodes which represent propositions explicitly. But that would mean that fine-grained theories, even with simple propositions, fail to furnish a conception of content suitable for (many) cognitive scientific models.

Although I admit that Dennett's view is seductive, I think it is a mistake. The best way to see this is to focus on the role of the word "explicit". What does it serve to exclude? Since its clearest opposite is "implicit", we can begin by

---

[74]Sometimes, it is said that symbols themselves must be explicit. As natural as this sounds, however, it involves a category mistake, at best. "Explicit" is properly used to describe the relation between a symbol and its content. It makes sense to ask whether a symbol represents a proposition explicitly or only implicitly; but it makes no sense to say, strictly and literally, that a symbol is explicit, just as it makes no sense to say that a symbol itself is implicit.

[75]By "information", I suppose, Dennett means something similar to possible conditions (as well as, perhaps, impossible ones).

getting a grip on that notion. Dennett characterizes implicitness admirably in the next paragraph: "for information to be expressed *implicitly*, we shall mean that it is *implied* logically by something that is stored explicitly" (Dennett, 1987, p. 216).[76] This definition has more plausibility; unlike his definition of "explicit", it might have been lifted straight out of any dictionary."[77]

So where's the problem for Dennett's requirement of syntactic complexity? The problem is that there is no appropriate connection between syntactic complexity and the explicit/implicit distinction. There is an appearance of a connection, and this appearance probably explains the initial seductiveness of the requirement, but the appearance is illusory. Specifically, it might appear that syntactic complexity is necessary for there to be a distinction between

---

[76]What is it for some information--some potential condition--to be "implied logically" by other information--another potential condition? It is, I suppose, for it to be logically necessary for the first condition to hold given that the second one does.

[77]More generally, there are the various forms of what Cummins, 1989, calls "inexplicit" content. There are two varieties. First, there are conditions which are logically necessary given explicitly represented conditions *plus* some other conditions (e.g., facts about the domain). Second, there are conditions which are logically necessary given *only* conditions not represented explicitly (e.g., the state of control, the form of the representation, or the medium of representation), independently of what is explicitly represented. Since my subsequent remarks about implicit content will depend only on the notion of logical necessity common to all of these types of inexplicit content, my remarks can easily be expanded to take these into account.

explicit and implicit content. Consider the following train of thought:

> It is easy to see that the syntactically complex expression "Paris is pretty" explicitly represents the fact that Paris is pretty, since the parts of the expression match up with the objects and properties participating in the fact. It is easy to see that it only implicitly represents the fact that there is at least one pretty object, since the parts don't match up in the same way. The situation is quite different for syntactically simple representations. How could there be an explicit/implicit distinction for these?

The proper response is: easily. All that is needed for a representation to have some content explicitly and others implicitly is for it to have content *at all*.

Suppose that one (explicit or implicit) content of a connectionist node is *that [one is seeing shape S]*. Interestingly, no matter whether this content is explicitly or implicitly represented, it *follows* that there are other contents which are merely implicitly represented: *that [someone is seeing shape S]*, *that [one is seeing or smelling some shape]*, and so on. We know that these contents are merely implicitly represented, because they are *less specific* than another (implicit or explicit) content of the representation.[78] This suggests that an *explicit* content of a propositional representation is a content which is *maximally* specific (relative to its other contents). *This* is why the

---

[78]A content *C1* is less specific than another content *C2* iff the condition specified by *C1* is logically necessitated by the condition specified by *C2*, but not vice versa.

sentence "Paris is pretty" explicitly represents the fact that Paris is pretty but only implicitly represents the fact there is at least one pretty object. Although this account would undoubtedly have to be sharpened for technical reasons, I have said enough for us to see why it is preferable to Dennett's account in terms of syntactic complexity. The specificity account is potentially applicable in a uniform manner to *any* representation, including individual connectionist nodes, and makes manifest the relationship between explicitness and implication. The syntactic complexity requirement on explicit representation fails on both of these scores, and should not be enforced. There is no reason, then, why simple propositions cannot be represented explicitly.

2.2.4 Summary

Since simple propositions are unfamiliar, it is necessary to motivate belief in them. How should this be done? Consider an analogous case from arithmetic, namely, the "discovery" of the number zero, or of the negative numbers. Suppose we meet someone who is familiar only with the positive numbers, and we want to convince him that zero is a number. To motivate the belief that zero is in the running to be a genuine *number*, it would presumably be necessary to show that it would be "continuous" with the recognized numbers, in that it would share enough of their important features and

relations. To motivate the belief that zero exists, furthermore, it would presumably be necessary to show that this belief serves a useful arithmetical purpose. I suppose that belief in simple propositions can be motivated in an analogous way--by showing that they are continuous with familiar fine-grained propositions, and that they serve a useful semantic purpose.

I hope it is clear in what important respects simple propositions are continuous with other more familiar fine-grained contents. Like other contents, these propositions are individuated by associated reference conditions, by a set of associated commitments to semantic dependence relations (in this case, the null set), and by the presence of rationality values (to determine whether a content is a concept or a proposition). Like other contents, they may be explicitly represented. Furthermore, simple propositions may themselves be presupposed by other, more complex contents: for example, if one representation has the content *that [A B C]* and another representation has the content *that [D E F]*, a suitable syntactic combination of the two representations might have the content *that [A B C] and [D E F].*[79] Finally,

---

[79]This is an advantage of the present theory over any attempt to draw Gareth Evans' intuitive distinction between "conceptual content" and "nonconceptual content" (Evans, 1983). If it were true that connectionist nodes had contents of a "new", nonconceptual, kind, it would be mysterious how these contents could figure in contents of the "old", conceptually structured, kind.

postulating simple propositions furthers the aims motivating fine-grained theories of content.

While these are reasons for considering simple propositions, specified as I have specified them, to be in the running to be contents, more reason may be demanded for postulating them in the first place. It seems to me that we have as much reason to postulate simple propositions as we have to postulate more familiar fine-grained contents; the main reason is that these contents figure in cognitive-scientific explanations. I suppose that connectionist nodes without semantic dependencies--more generally, but more roughly, all syntactically simple representations which are not abbreviations--have simple contents. Since simple propositions may serve as contents of connectionist nodes, and since such representations seem to be required by research programs which show some promise of yielding true theories of at least some cognitive phenomena, good methodology dictates that we should believe in simple propositions, if we want a fine-grained theory of content at all.

2.3 Toward a Naturalized Fine-Grained Theory

In this final section of the chapter, I would like to address the question of what sorts of entities we are

committing ourselves to when we commit ourselves to fine-grained propositions and concepts which admit of degrees of complexity. First I give a rough description of what I take to be the standard conception of such contents, namely as mathematical trees of a certain sort (section 2.3.1). Then I attempt to express some rather elusive metaphysical worries about such a view (section 2.3.2). Finally, I present an alternative account according to which contents and propositions are certain sorts of properties of representations (section 2.3.3). I hasten to express my belief that the conclusions reached in sections 2.1 and 2.2 in no way depend upon the metaphysical conclusions in this section.

2.3.1 Contents as Set-Theoretic Objects

On fine-grained theories of content, as I have described them, propositions and concepts cannot be identified with reference conditions. Instead, they must be identified with entities which admit of degrees of complexity, to reflect the semantic-dependence relations among representations. What precisely might these entities be? The relation between complex contents and simpler contents is most naturally treated as an abstract correlative to the relation between

wholes and parts, perhaps the relation between (mathematical) trees and subtrees.[88]

Although there are different ways to pursue this idea, to a first approximation a content *C* might be taken to be a tree structure whose nodes are "filled" by reference conditions. The reference condition at the root node would be the reference condition of representations with content *C*. If *C* is a simple concept or proposition, then the tree consists only of the root node. If *C* is a complex concept or proposition, however, the root node has descendant nodes. In this case, the relation of ancestry in the tree corresponds to the relation of semantic dependence: every node in the tree is be filled with a reference condition such that representing *C* requires semantic dependence on representations with that reference condition. Since some dependencies are *mediated* by others (e.g., "checkerboards are square" depends on "boards" only through depending on "checkerboards"), some nodes would be mediate descendants of the root, others immediate descendants.

The entire structure would then be reduced, in the way of mathematical trees generally, to complex set-theoretic objects

---

[88]See Lewis (1972) and Cresswell (1985) for illustrations of the tree theory. These authors credit Carnap (1947) with inspiration for their views, although I am not sure that Carnap would have much sympathy with the search for objects to identify with contents.

whose members, ultimately, are the reference conditions which fill the nodes of the tree. Contents, on this view, simply are such set-theoretic trees. To have a handy name, then, I will call this version of the fine-grained theory the "tree theory" of content.

2.3.2 Some Metaphysical Worries

Metaphysical worries form one source of resistance to the tree theory, although these worries are very difficult to pin down. We can certainly imagine philosophers who would object to the tree theory on the grounds that it postulates abstract entities. However, such philosophers would object equally to a coarse-grained theory of content which postulates reference conditions. Like trees, reference conditions are supposed to be abstract objects. They are possible ways for the world to be; in other words, they are *properties* which may or may not be instantiated by the world. For the purposes of choosing between the tree theory and the coarse-grained theory, then, we can ignore this metaphysical worry.

A more subtle worry arises in its place, however. In section 2.1.1, I described my assumption that contents are used in cognitive science to specify mental representations and to express generalizations about the functional roles of these representations. Along with a smattering of

philosophers, I am inclined to go one step further, assuming that cognitive science appeals to the content of mental representations not only to *express* generalizations about functional role, but also to provide *causal explanations* of these generalizations.[61] Crudely put, then, I am interested in conceptions of content which at least leave open the possibility that content has causal powers. Given this, there is a tension between the tree theory and a certain general naturalistic conception of the sorts of things which have causal powers.

Although a fair amount of mystery surrounds the notion of causation itself, we can take as relatively unproblematic the notion of a *concrete* (i.e., "physical") object's having causal powers. Intuitively, a concrete object has causal powers just in case some fact about the object has causes or effects. Furthermore, there are developing theories which seek to explain how (first-order and higher-order) properties of (and relations among) concrete objects, despite being abstract, play a role in natural laws, and in natural-world cause-effect relations.[62] By comparison, we have no understanding of how

---

[61]Incidentally, this provides a source of resistance to fine-grained theories of content which simply *specify* contents in terms of functional role (see, for example, Block, 1986). By building a symbol's functional role into the specification of its content, such theories remove the possibility of causal laws which help *explain* functional role in terms of content.

[62]See Dretske (1977), Tooley (1989), and Armstrong (1983).

set-theoretic objects play any sort of causal role. Sets are not standardly thought to be concrete objects, and they are not standardly thought to be properties or relations (of whatever order) involving concrete objects. Set-theoretic descriptions can, of course, be used to *classify* causally active objects, properties, and facts, but this is not the same as the sets' *being* causally active. If contents are set-theoretic objects, then, it is unclear that there is any room for them to have genuine causal powers, and so unclear that they can underwrite causal explanations in virtue of content.[11]

A final metaphysical worry about the tree theory is that it is in a crucial respect *stipulative*. Given a fine-grained theory of content, there are many different families of mathematical tree structures which can equally effectively be identified with contents. Furthermore, there are many equally effective ways to reduce mathematical trees to sets. Given this, it is natural to wonder *which* of the many suitable set-theoretic objects *is* identical to a given content. While

---

[11]Although some properties have causal powers, the situation for the tree theory is not improved by focusing on a symbol's (alleged) *property* of having a particular tree as content. It is only (*n*th-order) properties and relations involving *physical* objects which lend themselves to our naturalistic understanding of causality. If there are such entities as sets and trees, presumably *their* properties do not have causal powers any more than they themselves do. In particular, relations between symbols and trees, such as the "expression" relation postulated by the tree theory, would be causally inert.

it is true that this question may be brushed aside with a stipulation, we should in a naturalistic spirit favor a theory which gives an objective answer to the question of what contents are."

2.3.3 Contents as Properties

I think that a more defensible fine-grained theory of content can be developed, one which treats propositions and concepts not as set-theoretic trees but as *properties*, properties which are (potentially, anyway) causally active. Which properties are supposed to be identified with propositions and concepts? I want to identify contents with certain properties of *representations*. Before discussing which properties of representations I mean, it is best to describe in general terms how this strategy is supposed to

---

"David Lewis raises this as an objection to his own tree theory:

> It may be disturbing that in our explication of meanings [as set-theoretic trees] we have made arbitrary choices . . .. Meanings are meanings--how can we *choose* to construct them in one way rather than another? The objection is a general objection to set-theoretic constructions, so I will not reply to it here. But if it troubles you, you may prefer to say that *real* meanings are *sui generis* entities and that the constructs I call 'meanings' do duty for real meanings because there is a natural one-to-one correspondence between them and the real meanings. (Lewis, 1972, p. 201)

The objection does bother me, although in the next section I try to do better than postulating *sui generis* contents.

Incidentally, the "general objection" to set-theoretic identifications which Lewis mentions was first cast by Paul Benacerraf (1965) against the theory that numbers are sets of a certain sort. I think that my treatment of contents has an analog for numbers, although I am not yet prepared to defend such a treatment.

work. On a view of content common to the tree theory and most coarse-grained theories, concepts and propositions are thought to be objects related to representations, typically by the *expression* relation. An alternative is that contents are *types* of representations--in particular, *semantic types*. Any token object belongs to many types: my desk is a token of the type of thing made of wood, the type of thing I own, the type of thing in Massachusetts, and so on. To say this is simply to say that any token object has many properties: my desk has the property of being made of wood, the property of being owned by me, the property of being in Massachusetts, and so on. Indeed, the natural view is that types *are* properties. On the view I propose, then, when we say that a representation has a certain content, we are saying that the representation belongs to a certain semantic type, or, what is the same thing, that the representation has a certain property. This is what enables the identification of contents with properties of representations. In turn, this will enable a fine-grained theory to avoid the metaphysical worries associated with set-theoretic objects, and to leave room for a genuine role for content in causal explanations.[88]

---

[88]By identifying contents with properties, or types, of symbols, we go some way toward adjudicating certain theoretical differences between philosophers and psychologists. Philosophers, by and large, think of concepts and propositions as abstract objects, while psychologists often think of concepts and propositions as concrete, mental representations. On the rare occasions when one side chooses to admit the sensibility of the other side's conception, it is normally supposed that the other side "means a different thing". However, we can unify both conceptions in a sensible way. Since concepts and propositions are types, they are,

If we identify the content of a representation with the property of having a certain reference condition, unfortunately, the result is a coarse-grained theory of content. Nevertheless, as I explained in section 2.1.1, reference conditions do fulfill *some* of the constraints on contents, and so it is appealing to identify contents with properties which are specified at least *partially* in terms of reference conditions. We can come near enough to a *complete* specification by exploiting the semantic dependence requirement, as I described it in section 2.1.2. This yields the following first approximation (which, coincidentally and thankfully, is also my final approximation):

> A referential representation's content is its property of
> (i) having a certain reference condition, and
> (ii) being such that were representations with certain other reference conditions not to be present, then it would have a different reference condition.

Ignoring subtleties,[11] a content is fully specified once the reference conditions in (i) and (ii) are specified. Simple propositions and concepts, naturally enough, are those properties specified by the null set of reference conditions in (ii).

---

strictly speaking, abstract. But since they are types of symbols, we can at least speak of mental representations as *instances* or *tokens* of concepts and propositions. Concept tokens and proposition tokens are psychological entities in people's heads in precisely the same (useful) sense that word tokens and sentence tokens are on pieces of paper, even though word, sentence, concept, and proposition types are abstract.

[11]For example, a third clause is needed to distinguish properties which are propositions from those which are concepts (see section 2.2.1).

Chapter 3

FRAMING THE FRAME PROBLEM

The upshot of the previous chapter is that it is possible for associationist connectionist networks to have propositional representations, and so, at least potentially, to implement propositional attitudes. This is a good thing, since I am interested in the fate of such networks as models of certain kinds of inferential processes: namely, the rapid, holistically sensitive inferential processes which seem characteristic of skillful activity (see sections 0.1.3 and 0.2.3). As I described in chapter 1, such associationist networks are subject to severe computational limitations, including the inability to support skill transfer through general rules which introduce and bind variables. These computational limitations do not rule associationist models out of court, however, for it may be that the inferential mechanisms which ultimately distinguish experts from novices have the same computational limitations. In effect, it may be that the generality provided by variable-binding is sacrificed for increased speed and specialization at well-practiced tasks (see Agre, 1989, for arguments to this effect).

On the other hand, perhaps the inferential mechanisms characteristic of expertise do require devices such as variable-binding (see Anderson, 1983; Singley and Anderson, 1989). If so, it appears, we would need to use classical models to implement rapid, holistically sensitive inference. As I mentioned in section 0.2.3, one worry for doing so is that it is difficult to see how processes which function in a classical fashion--e.g., by "matching" syntactically complex symbols, binding variables, and the like--can access large numbers of representations rapidly, given that these processes must be implemented with slow neurons. Unless something can be done to mitigate this problem, there is a danger that no classical models (and so, by process of elimination, no cognitive-scientific models) will be able to account for expertise--including the everyday skills which surely account for most of what we do (for arguments to this effect, see Dreyfus, 1979; Dreyfus and Dreyfus, 1986; Haugeland, 1982; Searle, 1986; Preston, 1988).

What can be done about this difficulty? For one thing, there is some room for hope that connectionist implementations of classical models will help to increase the number of representations which can be accessed in a short amount of time (e.g., through massive parallelism). In this chapter, I would like to pursue a complementary strategy, namely, that of searching for computational techniques which reduce the

number of representations which need to be accessed in implementing particular inferential processes. Call these "relieving" techniques. With luck, a combination of relieving techniques can *sufficiently* reduce the computational burden of humanly possible inferential tasks so that they can be implemented by classical models with (say) neural processors.[67]

To drive the search for relieving techniques, I want to address three specific problems for cognitive science which philosophers have posed under the name of the "frame problem". This term originated in artificial intelligence (AI) research (McCarthy and Hayes, 1969), and so philosophers interested in the frame problem have normally directed their attention to this discipline. I will follow this practice, although at a number of places I will indicate relevant implications for cognitive science generally. My first aim is to characterize the original frame problem, so that it may be compared to the philosopher's problems.

---

[67]We needn't find techniques which insure that a small enough burden is associated with *arbitrary* inferential tasks. Since the ultimate objective is to solve part of the mental-brain problem, we need only consider inferential tasks which humans can perform. In particular, for those inferential tasks which are specified in environmental terms (e.g., the task of inferring one's location with respect to objects in the environment, given one's previous location and one's movements), we need only consider environments in which humans can succeed (e.g., fairly stable environments in which the objects themselves do not move unpredictably).

3.1 Persistence and the Frame Problem

3.1.1 A Fable

Once upon a time there was a causation-computer, named C2 by its creators. Its only task was to read about simple events and to report their likely effects, in as much detail as it could. One day its designers arranged for it to learn that a bomb was in a room, resting on a wagon, and that the wagon was pulled through the doorway. C2 quickly reached the obvious conclusion that the bomb rode out of the room. "CONTRADICTION!" it printed, to the surprise of its teachers. "THE BOMB WAS BOTH IN AND OUT OF THE ROOM. CONTRADICTION! CONTRA"--they were forced to unplug it. Poor C2 could not understand that the time at which the bomb was out of the room was different from the time at which it was in the room.

Back to the drawing board. "The solution is obvious", said the designers. "Since states may change from one moment to the next, our next computer must represent the particular moments at which they obtain." They called their next model, the chronological-causation-computer, C3. C3 was told that the bomb was on the wagon at $t1$, and that the wagon was pulled a moment later, at $t2$. Then the programmers put it to the test:

"Tell us as much as you can about the effects at *t*3."
"THE WAGON WAS OUT OF THE ROOM AT *t*3."
"Anything else? Did anything happen to the bomb?"
"I DON'T KNOW. WHERE WAS IT WHEN THE WAGON WAS PULLED?"
"We just told you it was on the wagon, you tin ninny!"
"SURE, IT WAS THERE AT *t*1, BUT MAYBE THAT CHANGED BY *t*2."

Further questioning confirmed the worst: they had neglected to teach C3 how to tell which changeable facts persisted from one time to the next. "What color is the wagon?" "I DON'T KNOW--MAYBE IT CHANGED BECAUSE THE WAGON WAS PULLED." "What is your name?" "I DON'T KNOW--IT WAS 'C3' BEFORE YOU TOLD ME ABOUT THE ROOM." After a few more questions, mercifully, someone pulled the plug.

Back to the drawing board. "We might try giving it 'frame axioms'", said the designers, "which put a border around the effects of an event." They soon realized that this was hopeless, however, since the number of frame axioms would mushroom. They would have to teach their next model that reading about a wagon does not change its color, that pulling a wagon does not change one's name or change the number of pink elephants in the world, and so on. This presented the "frame problem": how to design a system which could, unlike C3, infer the persistence of nonchanges, but which could do so *automatically*--that is, without explicitly storing or accessing frame axioms for them.

Before long, the programmers discovered various ways for a system to infer automatically the persistence of nonchanges. Their favorite was the suggestion that representations of facts should refer not to particular *moments* but to time *intervals*. Thus was born a chronological-causation-computer-for-persistence, named C3P. C3P was given the same problem that had stumped C3. When C3P learned that the bomb was on the wagon at *t*1, it generated this internal representation:

R: THE BOMB IS ON THE WAGON FROM *t*1 ONWARD.

R did not need to be updated with each passing moment to handle persistence, since *R* itself meant that the bomb was on the wagon at *t*2, *t*3, and so on. This allowed C3P, unlike C3, to infer the bomb's motion, when it was told that the wagon was pulled at *t*2. The programmers also gave C3P the ability to "snip" representations such as *R*, by representing finite intervals. For example, when C3P learned that the bomb was taken off the wagon at *t*100, it substituted "TO *t*99" for "ONWARD" in *R*. As a result of all of this, C3P was able genuinely to ignore facts that it took to be unchanged by a given event, focusing only on purported changes. Since there was no longer a need for storing and accessing frame axioms, the designers of C3P were satisfied that they had solved the frame problem.

All was calm, all was bright, until one night three wise men arrived from the East. C3P received no homage from them,

however--much less any expensive gifts. The first wise man deemed the frame problem "a new, deep epistemological problem" which "whatever it is, is certainly not solved yet". The second wise man intensified the point, suggesting that the frame problem is "foisted on unsuspecting epistemology by misguided presumptions underlying AI as a discipline." Needless to say, the programmers found this completely mystifying. "You may suppose that you have solved the frame problem," explained the third wise man, "but in fact you are begging it. How could the depth, beauty, and urgency of the frame problem have been so widely misperceived?" In answer to his own question, he pronounced, "It's like the ancient doctrine of the music of the spheres. If you can't hear it, that's because it's *everywhere.*" Satisfied that their hosts were completely at a loss for words, the wise men bid them farewell. As they left, the first wise man turned and issued the ominous warning, "If there is ever to be a robot with the fabled perspicacity and real-time adroitness of C3P*O*, robot-designers must solve the frame problem."[**]

[**]I have transcribed the words of the three wise men from the reports of Daniel Dennett (whose original "R2D2" fable is the model for mine), John Haugeland, and Jerry Fodor, respectively (Dennett 1987, pp. 42-43; Haugeland 1987, p. 93; Fodor 1987, p. 142).

3.1.2 Persistence and Sleeping Dogs

The frame problem as it is most commonly construed in AI was first described and named by John McCarthy and Pat Hayes (1969). These authors were interested in exploring a certain formalism for reasoning about change, called the 'situation calculus'. In the situation calculus, changeable facts are represented as being relative to particular moments of time, as in the chronological-causation-computer, C3. The facts which are represented as obtaining at a given moment are said to constitute a 'situation'. Given that an event *E* occurs in a situation *S*, and that certain surrounding conditions hold in *S*, the system's task is to calculate what is true in the next situation, *S*+1. It does so by applying conditional rule-symbols which describe purported effects of *E*-like events, given that certain facts obtain in *S*. McCarthy and Hayes called these rule-symbols "axioms", although these symbols needn't be unsupported or irrefutable.

McCarthy and Hayes discovered, though presumably not the hard way, that the situation calculus deals with nonchanges very inefficiently. Such a system makes no inferences about what is true in situations unless these inferences are sanctioned by axioms. Consequently, a system needs axioms relating each event of which it has an idea (e.g., *E*) to each

changeable fact of which it has an idea. This is true even of facts which are not (purported to be) changed by a given event. These "frame axioms" have (to a near enough approximation) the form: "if *E* occurs in *S* and *F* is true in *S*, then *F* is true in *S*+1". Without such an axiom, the system would not infer that *F* persists, as illustrated by C3 in the fable. It is difficult to see how inferential processes which access huge numbers of frame axioms could be implemented (in brains or in known computers) without serious degradations in speed. For this reason, McCarthy and Hayes posed what they called the "frame problem": how can the persistence of nonchanges be inferred *without* accessing frame axioms? In order to minimize confusion with other problems (to be discussed) which have come to be called the "frame problem", I will adopt the more descriptive term "persistence problem" for this original frame problem (see Shoham 1988).

The standard strategy for solving the persistence problem has usefully been labeled the "sleeping-dog strategy".[11] According to the sleeping-dog strategy, instead of using frame axioms a system should assume *by default* that a fact persists, unless there is an axiom specifying that it is changed by an occurring event (given existing conditions). In this way,

[11]The strategy originated in Fikes and Nilsson, 1971, and has since appeared in a number of more sophisticated guises, e.g., Shoham, 1988. The term is due to Haugeland, 1987.

given that an event E occurs in situation S, the system can use axioms to infer new facts existing in $S+1$, and then simply "copy" the remainder of its beliefs about S over to $S+1$. In turn, the copying process can be avoided by abandoning the framework of momentary situations, in favor of that of extended time intervals, as illustrated in C3P. If a representation specifies a fact as holding over an interval of time, then the representation need not be accessed at all unless an axiom of change becomes applicable to it (Shoham, 1988). By modifying both the situations and the calculations of the situation calculus, the sleeping-dog strategy allows the persistence of facts to be inferred "automatically", that is, without accessing representations of the facts. A system can let sleeping representations lie, unless there is a positive reason to wake them.

The sleeping-dog strategy is a relieving technique, in the sense defined at the beginning of the chapter, and so it appears valuable for purposes of developing classical models which can implement rapid, holistically sensitive inference.[90]

---

[90]Jim Higginbotham has suggested to me that the sleeping-dog strategy is a special case of a more general strategy which is applicable, for example, in the following sort of case. Imagine an agent who is placed on a disk which rotates at a constant speed, and who (for whatever reason) has to keep track of certain of his momentary, egocentrically-specified relations to objects which are off the disk and out of sight. Since these relations change continuously, the agent has continuous "positive reason" to access and update his corresponding representations. But since these relations change predictably, the system can employ a regular procedure for these updates. In this case, the system can adopt a "sleeping-procedure strategy": don't consider changing the procedure

Although the sleeping-dog strategy seems to reduce the computational burden of virtually any task in reasoning about change, it does not by itself serve to "relieve" the suspicion that the burden associated with humanly feasible inferential tasks is too large to be implemented in brains by classical models. Short of providing a complete model of human cognition, there is not much which can be done to relieve this suspicion. What we can do is examine the specific reasons

---

unless there is positive reason to do so. The sleeping-dog and sleeping-procedure strategies might then be seen as special cases of the nearly trivial "sleeping-thing strategy": don't consider changing anything (e.g., one's wallpaper) unless there is a positive reason to do so.

While I think this is right, there is good reason to give special attention to the sleeping-dog strategy (formulated as a principle specifically governing the direct alteration of beliefs about changeable facts). If we focus on the general sleeping-thing strategy, I think, we remove the possibility of placing substantive constraints on what it is to be a "positive" reason. In the general case, maybe a positive reason to consider changing something is simply that one is in the mood to do so, or, even more weakly, simply that one finds oneself about to do so. With respect to the sleeping-dog strategy conceived in the standard way, however, we can motivate at least a skeletal account of what it is for a given computational system to have a positive reason for considering the revision of a belief *that p*: perhaps the sort of system under consideration must have an "axiom" whose antecedent is believed to be satisfied, and whose consequent is *that not-p*. Furthermore, we can add that the axiom should not be a frame axiom--that is, its antecedent should not presuppose *that not-p*. This account can be elaborated by requiring the "axioms" to have theoretical or observational support, to admit of certain degrees of confidence, etc.

Furthermore, the sleeping-dog strategy has special interest as a relieving technique, at least by comparison with the sleeping-procedure strategy. Generally speaking, there are more representations (of changeable facts) than there are procedures, so the sleeping-dog strategy has greater potential for reducing computational load. Indeed, the sleeping-dog strategy has merit even in Higginbotham's spinning-disk example. Although in this case it is possible that the agent must continually access and update a huge number of representations, without the sleeping-dog strategy the agent would have to access and update an even *huger* number of representations, e.g., about the unchanging but changeable aspects of his situation.

which have been given for this sort of suspicion, and attempt to respond to them individually. Accordingly, the rest of this chapter is devoted to an examination of three attempts to display the frame problem as a deep, difficult problem: Dennett's problem of ignoring obviously irrelevant knowledge, Haugeland's problem of efficiently keeping track of salient side effects of occurrences, and Fodor's problem of avoiding the use of certain "kooky" concepts. Each problem consists of considerations which seem to threaten the project of reducing the number of representations which inferential processes need to access.

In a negative vein, I will argue that their problems bear nothing but a superficial similarity to the original frame problem of AI (see McCarthy and Hayes 1969). Of course, it must be conceded that the terminological issue is unimportant. (This is one reason for using the term "persistence problem" for the original frame problem.) But the point is more than terminological, for it weighs against the philosophers' use of their frame problems to disparage the sleeping-dog strategy. The primary negative claim of this chapter, then, is that the sleeping-dog strategy is not susceptible to criticism based on their new problems.

More positively, I will argue that their problems are easily solved by slight variations on familiar AI themes.

These solutions will take the form of various relieving strategies, as desired. Finally, along the way I will devote some discussion to more difficult problems confronting the classical framework in AI and cognitive science. Although these unsolved problems will, I think, require enormous effort to overcome, understood properly they are not "principled" problems in a certain relevant sense. Although they may (or may not) lead one to suspect that cognitive scientists will never *actually* provide a complete account of human cognition, they do not lead one to doubt the partial accounts which may (or may not) be provided. They are difficult problems for cognitive science as a human activity, but not difficult problems for cognitive science as a type of theory.

3.2 Relevance and the Frame Problem

3.2.1 The Relevance Problem

As I mentioned, my introductory fable is a twist on a fable with which the first wise man, Daniel Dennett, introduces the frame problem of AI (Dennett 1987, pp. 41-42). I will first retell his tale, and then explain how it is misleading in this role. The robots in Dennett's fable are charged with the task of mentally testing a plan, given a goal to be reached and some idea of the initial conditions under which the plan is

to be executed. Each of them comes complete with these three states:

G: the goal of saving its spare battery from a live bomb.
I: knowledge of the initial conditions that the battery and the bomb are on a wagon in a room.
P: the plan of pulling the wagon out of the room (to remove the battery).

Plan testing also requires a fourth element, a set *R* of "inference rules". To test a plan, one tries to find a sequence of rules in *R* which allows the goal to be inferred from the plan and the initial conditions. In other words, one searches for an "inferential path" from the plan and the initial conditions to the goal, one for which each step along the way is sanctioned by an inference rule.[91] Very roughly, if such a path exists, the plan passes the test.

Dennett begins with a simple robot, R1, which can recognize "the intended implications of its acts", but not "the implications about their side effects". In other words, in testing a plan, R1 uses only inference rules which correspond to intended effects of the plan. Since *G* is an intended effect of *P*, of course, *P* passes R1's test. So R1 proceeds to pull the wagon out of the room, without recognizing the tragic side effect due to the fact that the bomb is also on the wagon. Back to the drawing board go the designers; out

[91]For a good introduction to the AI literature on "search", see chapter II of Barr and Feigenbaum (eds.) 1981. The term "operators" is standardly used for inference rules as well as other goal-reaching devices which do not concern me here.

pops the robot-deducer, R1D1, which can test its plans for side effects. It does so by removing all restrictions on which inference rules and initial conditions it can consider in testing a plan. As a result, in searching for an inferential path from *P* to *G* it "deduces" everything it can: that *P* "[does] not change the color of the room's walls", that *P* "cause[s] [the wagon's] wheels to turn more revolutions than there [are] wheels on the wagon", and so on. Boom! Therefore, the designers install in their next robot a method for tagging implications as relevant or irrelevant to its goals. They call the new model R2D1, the robot-relevant-deducer. The relevance tags don't help, however, since not only does R2D1 waste time inferring all the same irrelevant implications, but it also generates more inferences to the effect that they are irrelevant. "All these robots suffer from the frame problem," Dennett concludes. "If there is ever to be a robot with the fabled perspicacity and real-time adroitness of R2D2, robot-designers must solve the frame problem".

R1D1 and R2D1 do seem to illustrate the original frame problem--the persistence problem--since they engage in explicit inferences about nonchanges such as the color of the walls. The persistence problem requires one not to use frame axioms to infer the persistence of nonchanges. My claim is that a good dose of the sleeping-dog strategy would cure this

ill, and I will argue for this claim throughout the course of this chapter. However, these robots suffer from a further problem which is not even addressed by the sleeping-dog strategy: not only do they bother with the noneffects of their plans, but they also bother with many genuine effects which are obviously irrelevant to their goals, such as the number of revolutions of the wagon's wheels. The extra problem facing their programmers, then, is how to design systems which test plans without bothering with obviously irrelevant inferences.

This problem may be generalized in a straightforward way, since there are other kinds of goal-oriented searches besides plan testing. In order to generate a plan, for example, one may search for an inferential path from the initial conditions to the goal which requires performing some actions. In order to generate subgoals for a current goal, one may search for an inferential path to the goal which requires that certain subgoals be reached. From this general perspective, Dennett's problem becomes that of designing a system which finds inferential paths between initial conditions and goals without considering inferences which "obviously" do not point in the right direction. I will call this the "relevance problem".

3.2.2 Relations to the Frame Problem of AI

Despite the similarities between the persistence and relevance problems, it is something of a mystery why, in Dennett's hands, the shift takes place. He seems to feel that the original frame problem is merely an instance of the more general relevance problem. Thus, he calls the relevance problem a "broader" problem than the "narrowly conceived" original frame problem (Dennett 1987, p. 43). Although this may have some initial appeal, I think it should be resisted.

First, consider what Dennett can say in defense of the claim that the persistence problem is an instance of the relevance problem. A first attempt might be to argue that the desirability of ignoring noneffects of an event follows from the desirability of ignoring all irrelevant knowledge. The situation is not so simple, however. Often, noneffects *are* highly relevant to one's goals: in Dennett's fable, for example, pulling the wagon does not change the fact that the battery is on the wagon, and this is directly relevant to the robot's goal. Therefore, the robot might need to access the knowledge that the battery will stay on the wagon.

Nevertheless, it is possible for Dennett to reply that, even if noneffects are often relevant to a system's goals, processing them with explicit frame axioms is irrelevant. However, this substitution of "irrelevant processing" for "irrelevant knowledge" forces an unwelcome shift in the construal of the relevance problem. What is "irrelevant processing" supposed to mean? Useless processing? But if a robot needs to know about a certain (relevant) noneffect, a corresponding frame axiom might be very useful in supplying this knowledge. Of course, given that systems can use the sleeping-dog strategy instead of frame axioms, the latter are too costly. But being too costly is not the same as being irrelevant. If it were, *any* problem about processing costs would be a problem about irrelevant processing. On this view, for example, electrical engineers debating the relative processing virtues of various home computers would be discussing an "instance" of the relevance problem! But then the relevance problem would no longer be *Dennett's* problem of accessing useful knowledge at the right time. Therefore, appealing to the irrelevance of processing noneffects fails to show that the persistence problem is an instance of Dennett's relevance problem.

There is a more direct reason not to assimilate the persistence problem to the relevance problem. The persistence problem arises completely independently of goals, planning,

action, or problem-solving. It deals purely with causal reasoning--keeping track of change. In my fable, C3 and friends are "pure predictors"; the only "goal" they ever have is to report as much as they can about the effects of an event. As a result, *every* effect is "relevant" to them, and no effect is irrelevant. Therefore, no instance of the relevance problem can arise for pure predictors like C3; there are no irrelevant effects to ignore. Since the persistence problem is present in its full force for C3, it cannot be an instance of the relevance problem. Nevertheless, the point remains that if there are ever to be smart robots such as R2D2 and C3P0, the relevance problem must be solved.

3.2.3 The Role of Bidirectional Search

The task facing the plan-tester is, as I have described it, that of searching for an inferential path from a plan and some initial conditions to a goal. In this respect it is rather like walking through a labyrinth, searching for an unobstructed path from the entrance to the exit. Now, compare three strategies for negotiating a labyrinth. First, there is "forward search": starting at the entrance and walking around (marking one's path, of course) until one happens upon the exit. Second, there is "backward search": starting at the exit, and trying to make one's way to the entrance. Third, there is "bidirectional search": searching forward while a

partner searches backward, until one finds a path marked by the other. Bidirectional search is clearly the more efficient strategy, in general (see Barr and Feigenbaum (eds.) 1981, pp. 46-53).

From this perspective, it appears that a major defect of Dennett's robots is that they engage only in forward search. His robots start with their plan *P* and initial conditions *I* and keep making inferences from these (and from their consequences, and so on) until they happen upon their goal *G* (or its negation). As a result, they infer consequences more or less at random, with respect to the goal, and so suffer from the relevance problem. We can account for one aspect of R2D2's fabled perspicacity and real-time adroitness if we suppose that it uses bidirectional search instead. Supposing this, how would R2D2 test *P*?

We can imagine R2D2 first searching backward from *G*. The procedure is to look at some inference rules of the form "IF <condition>, THEN *G*", and to mark these conditions as plausible parts of paths from *P* to *G*. (Recall that *G* is the goal of saving the battery from the bomb.) This set of inference rules is likely to refer to the condition that the battery and the bomb are far apart, but is unlikely to refer to conditions regarding the number of revolutions of a wagon's

wheel, or the color of the walls."[11] So the locations of the battery and bomb would be marked as "relevant" to *G*.

At this point, R2D2 can ask itself the question: what happens *to the whereabouts of the battery and bomb* if I roll the wagon out of the room? More precisely, R2D2 can let the details of this question *guide* its forward search from this plan. That is, instead of looking at all the rules of the form "IF . . . A WAGON ROLLS . . ., THEN <consequence>", it can look only at those with potential consequences for the positions of batteries and bombs. Presumably, it finds inference rules such as these:[12]

IF A WAGON ROLLS, AND *x* IS IN THE WAY,
THEN *x* IS PUSHED ALONG.
IF A WAGON ROLLS, AND *x* IS ON THE WAGON,
THEN *x* RIDES ALONG.

R2D2 therefore checks whether it believes that the battery and bomb satisfy *x* in the antecedents of these rules. It finds that, in fact, it does believe that the two are on the wagon, so it infers that the two will ride along, and will not be far apart. Finally, it infers that the battery will not be saved, and can try to find a better plan based on what went wrong with this one.

---

[11]As I explain in section 3.2.1, "relevance holism" creates a difficulty here, but one which can be solved.

[12]I omit nuances such as the temporal factors mentioned in the introduction and the exceptions discussed in section 3.2.1.

As I mentioned above, the relevance problem arises for tasks other than plan testing, such as subgoal generation and plan generation. Given that R2D2 can paint the wagon, draw the drapes, or pace up and down the room, what keeps it from considering these options in generating a plan to rescue its battery? Bidirectional search does. R2D2 can search backward from the goal, to find subgoals and actions most likely in general to lead to the goal. It can then direct its forward search from the initial conditions, to determine which of these subgoals and actions are most likely to be suitable under these conditions. Other subgoals and actions should be considered only if none of these are suitable or if subsequent plan testing rules them out.

Although bidirectional search is a relieving technique in that it greatly reduces the number of representations which need to be accessed in problem solving, it does not itself bring these computational costs to a minimum. In my illustration, I vaguely described R2D2 as looking at "some" inference rules of the form "IF <condition>, THEN *G*". But which? If it looks at them all, it is likely to bother with many irrelevancies. I discuss this problem in the next section, in connection with "relevance holism". First, however, I want to discuss briefly another problem related to the relevance problem.

Although Dennett casts the relevance problem as a problem about finding knowledge relevant to one's current goals, it may be suspected that there is a deeper problem about how to make the right goals current at the right times. However, I am not aware of any attempts to show what the "problem" is. We might suppose that some goals are always current in R2D2, e.g., the goal of staying out of danger, and that some goals are triggered by certain conditions, e.g., given that there is a potential danger, R2D2 can generate the goal of finding out if any valuables are in danger and removing them from the danger. Once R2D2 learns that there is a live bomb in the room (i.e., a potential danger), but that there is some time to work with (so R2D2 itself is not yet in danger), R2D2 can search for valuables near the bomb (i.e., in danger). We can imagine that it can discover that the battery is near the bomb either by quickly looking around the room, or else by being told this, as in Dennett's fable. Consequently, it can generate the goal of removing the danger and, as I have described, it can generate and test plans to meet this goal.

3.3 Holism and the Frame Problem

3.3.1 The Holism Problem

The second wise man, John Haugeland, construes the frame problem as arising from the fact that inferential relations

in the real world are holistic: what it is reasonable to infer from a given condition may depend on many other "surrounding conditions". First, virtually any inference can be warranted by virtually any condition, if the right surrounding conditions hold. From the premise that a wagon is pulled, for example, one may infer that a bomb moves (if there is one on the wagon), that one pulls a muscle (if the load is heavy), that the wheels will squeak (if they aren't oiled), that one will please a coworker (if he has asked for the slab on the wagon), and so on. Second, virtually any inference can fail to be warranted by virtually any condition if the wrong surrounding conditions hold. As Haugeland points out, there are many possible situations in which pulling a wagon might fail to make a bomb ride along, even though the bomb is on the wagon:

> But what if [the bomb] is also tied to the doorknob with a string? Or what if, instead of [rolling], [the wagon] tips over? (Haugeland 1987, p. 85)

This holism leads to Haugeland's problem:

> The so called frame problem is how to 'notice' salient [inferences] without having to eliminate all of the other possibilities that might conceivably have occurred had the facts somehow been different. (Haugeland 1985, p. 204)

In other words, the problem is to come up with an efficient algorithm for respecting the fact that what may be inferred from a given condition may depend on virtually any surrounding condition. Such an algorithm would have to make tractable the number of surrounding conditions a system must check, without

blinding it to the "salient" ones. In order to distinguish this problem from others that have gone by the name "frame problem", I will refer to it as the "holism problem".[94]

The holism problem intensifies the relevance problem. My illustration of bidirectional search in section 3.2.3 proceeds under the assumption that the inference rules associated with R2D2's goal of saving the battery from the bomb do not refer to the precise number of revolutions of a wagon's wheel, or the color of the walls, or any other "obviously" irrelevant conditions. However, if the bomb is activated by the squeaking of the wagon's wheels, the precise number of revolutions of the wheels may be of crucial relevance. Even the color of the walls may be relevant, if the room is painted in such a way as to camouflage the door. As a result of this holism, to deal with the real world R2D2 is likely to need inference rules to handle these possibilities, raising the combined "relevance-holism" problem: how can a system know which knowledge is relevant to a goal in its particular situation, without having to think about a vast number of possibilities?

---

[94]In The Modularity of Mind, Fodor anticipates Haugeland's treatment of the frame problem as a problem about holism. He writes that one of the things that "makes [the frame] problem so hard" is that "which beliefs are up for grabs depends intimately upon which actions are performed and upon the context of the performances" (Fodor 1983, p. 114).

3.3.2 Relations to the Frame Problem of AI

As Haugeland points out, the sleeping-dog strategy does not provide a solution to the holism problem. But more than this is needed to show that something is wrong with the sleeping-dog strategy, of course. (After all, the sleeping-dog strategy also "fails" to solve the problem of world hunger.) Haugeland therefore makes a stronger claim, to the effect that the sleeping-dog strategy *raises* the holism problem. The sleeping-dog strategy requires there to be "positive indications" to the effect that certain facts are changed by an event, so that the system can focus only on these facts. These positive indications are provided by inference rules (e.g., the non-frame axioms of the situation calculus). Therefore, he concludes, it is the sleeping-dog strategy which "raises formidable design questions about how to get the needed positive indications for all the important [inferences]", i.e., the holism problem (Haugeland 1985, p. 206). On closer inspection, however, it is easy to see that it's *not* the sleeping-dog strategy which raises the holism problem; the problem arises for *any* system which has to make inferences about the real world, whether or not it uses the sleeping-dog strategy. For example, in my introductory fable the computer C3 does not use the sleeping-dog strategy. Nevertheless, of course, it must make inferences, and these

inferences must be sensitive to salient surrounding conditions, so it must face problems about inferential holism.

As a consequence, something more is needed to show that the sleeping-dog strategy is inadequate for the problem it's intended to solve, namely, the persistence problem. Perhaps Haugeland's idea is that the persistence problem cannot be solved without simultaneously solving the holism problem. Since he does not even attempt to provide reasons for bringing inferential holism into discussions of the frame problem, there is room for speculation about why he is tempted to do so. Perhaps the reasoning goes like this: to be a solution to the persistence problem, a system must ignore the facts which are not changed (by an event), so it must be able to tell which facts are changed, so it must respect the holism of change, and, more generally, the holism of inference. The problem with this argument is fairly subtle; to display it I must invoke a distinction between domain-general "process-and-form" problems and domain-specific "content" problems. I will devote more attention to this distinction than may at first appear necessary, because it will prove to be crucial later in this section, in my defense of a solution to the holism problem.

Much research in AI proceeds on the assumption that there is a difference between being well-informed and being smart.

Being well-informed has to do, roughly, with the content of one's representations--about their truth and the range of subjects they cover. Being smart, on the other hand, has to do with one's ability to process these representations and with packaging them in a form that allows them to be processed efficiently. The main theoretical concern of artificial intelligence research is to solve "process-and-form" problems: problems with finding processes and representational formats which can qualify a computer as being smart.

Of course, in order to build computers which can deal with the real world, we must also solve "content" problems involving figuring out which particular representations computers should have, so that the computers qualify as being well-informed about a variety of domains. It is neither surprising nor worrisome that AI has not solved all these content problems, for they are not, in the first instance, AI's problems. One can make headway into process-and-form problems in the AI laboratory, but to make headway into content problems, one must incorporate empirical investigations in particular domains ranging from medical diagnosis to the mechanics of middle-sized objects to sociology to chess to laundromats to train stations. These investigations appear to demand enormous resources; if we decide not to allocate these resources, it may even be impossible for AI to succeed at constructing intelligent and

well-informed systems. But this is a practical rather than principled problem, in the specific sense that it does not indicate that the classical AI framework is incorrect. It seems a reasonable division of labor, then, for AI to pass domain-specific bucks to domain-specific sciences."[88]

Accordingly, the persistence problem is posed as a domain-general process-and-form problem. In other words, it is not about which particular facts a system should take to be unchanged by which events. Consider again the frame axiom proposal (see section 3.1). Frame axioms turned out to be a bad idea, not because they didn't capture reliable information about nonchanges (we may suppose that they did), but because there were too many of them. The persistence problem therefore arises regardless of how reliable or unreliable a system is about which facts are unchanged. As a result, to solve it all we need to do is to design a system which has the

---

[88]Couldn't we avoid having to gather all of this information for the computers, by designing them to investigate the world for themselves as children do? No, for two broad reasons. First, setting computers loose in the world involves implanting them in robots; but we don't yet know how to build robots which can see, feel, hear, hop, skip, and jump well enough to cross a city street safely. Second, there is the "blank slate" problem. It appears impossible to learn efficiently about a domain unless one already has some reliable information about what sorts of data to concentrate on, what sorts of hypotheses to try out, etc. Thus, building robot learners requires endowing them with considerable domain-specific innate knowledge, which requires us to engage in domain-specific investigations after all. Add to this the explicit instruction (e.g., "book learning") which must be gathered and presented to children, and it appears that a robot's need to be spoonfed extensive amounts of domain-specific knowledge is unsurprising and rather humanlike.

capacity to ignore the facts which are not changed, if it knows which facts really are unchanged.

It is this fact which shows that the holism problem does not lurk behind the persistence problem. To be a solution to the persistence problem, a system only needs to ignore the facts it *thinks* are not changed by an event. But to do that, the system needn't be able to tell which facts really are changed. Since a solution to the persistence problem needn't insure that systems are *right* about which facts are changed, it needn't insure that systems have the capacity to keep track of the holism of change. So the sleeping-dog strategy can solve the persistence problem without solving the holism problem. Of course, I am not denying that we need to solve the holism problem in order to get intelligent machines that can deal reliably with the real world. In the rest of this section I focus on attempts in AI to solve this very problem. The point here is merely that the fate of this problem is irrelevant to the fate of the sleeping-dog strategy.

3.3.3 The Role of Heuristic Search

At root, the holism problem is this: for any set of conditions one wishes to make inferences from, there are always too many potentially applicable inference rules to consider, rules which may require one to check virtually any

surrounding conditions. Returning to the labyrinth analogy, the problem is that, from any fork, there are so many paths that one can't follow them all. If one knows nothing about the particular labyrinth one is in, one must select a path more or less at random. This is called "blind search" in AI. However, in some cases one can use specific information about the labyrinth to help one select the paths which are likely to be the best to follow. This is called "heuristic search". For example, one might know that the better paths in a certain garden tend to be wider, while those in another tend to be better lit. Such heuristics can help one to achieve better results than blind search (see Barr and Feigenbaum (eds.) 1981, pp. 58-63).

Now, when a computer is searching for inferential paths, it can use similar heuristics to avoid blindly checking every inference rule. For example, associated with each inference rule might be some measure of its general reliability. The inference rule "IF A WAGON IS PULLED, IT ROLLS" might, for instance, be deemed more reliable than "IF A WAGON IS PULLED, THE COLOR OF THE WALLS CHANGES". In addition, or instead, each inference rule might make reference to the antecedent probability that it will "apply", that is, to the antecedent probability of the surrounding conditions it presupposes. Take the rule "IF A WAGON ROLLS, AND *x* IS ON THE WAGON, THEN *x* RIDES ALONG". As Haugeland says, this rule can fail if *x*

is tied to the doorknob, but then the antecedent probability of such failure might be deemed to be very low.

Given some such metric, a computer can constrain searches by looking initially only at the set of rules with the best marks (the size of the set depends on how many rules can be processed at the same time). It can thereby focus on the rolling of the wagon rather than the potential change of color of the walls, and it can assume "by default" that *x* is not tied to the doorknob."[6,7] If this set doesn't get it where it wants to go, it can try the next best set, and so on down the "search hierarchy".

If one's special concern is relevance holism, one might prefer (instead, or in addition) to use heuristics regarding

---

[6]"Occasionally, when the stakes are high, it may be advantageous for a system to go into a more careful mode in which it avoids making some default assumptions, explicitly checking the surrounding conditions instead. I ignore this nicety here, since to the degree that a system needs to be careful, the holism problem is made less important: if the stakes are so high that a system needs explicitly to check surrounding conditions, it can hardly be faulted for doing so.

[7]"AI researchers have had mixed success in trying to develop a "nonmonotonic logic" for reasoning with default assumptions (for a review of this literature, see Shoham 1988). From the perspective adopted here, however, default (or nonmonotonic) reasoning is an ordinary example of heuristic search, which is generally thought not to require the development of a corresponding "logic". This is one way of seeing that we may not need nonmonotonic logic (as opposed to nonmonotonic reasoning), so that the shortcomings of nonmonotonic logics may not be important. If some in AI have not appreciated this point, it is perhaps due to placing too much emphasis on the distinction between heuristics and "epistemology" (i.e., inference) offered in McCarthy and Hayes 1969 (for an example of this, see Janlert 1987, pp. 2-3).

the general *usefulness* of inference rules. For instance, the rule "IF A WAGON ROLLS, AND *x* IS ON THE WAGON, THEN *x* RIDES ALONG" may be deemed to be generally more useful than the rule "IF A WAGON ROLLS, THEN THE NUMBER OF REVOLUTIONS OF ITS WHEELS IS PROPORTIONAL TO THE DISTANCE". This may be so even though the former is less reliable (since *x* might be tied to the doorknob) and less likely to be applicable (since the wagon might be empty)."

Although Haugeland doesn't discuss heuristics as an approach to the holism problem, Jerry Fodor, the third wise man, registers this complaint:

> So far as I can tell, the usual assumption about the frame problem in AI is that it is somehow to be solved 'heuristically'. . . . Perhaps a bundle of such heuristics, properly coordinated and rapidly deployed, would suffice to make the central processes of a robot as [holistic] as yours, or mine, or the practicing scientist's ever actually succeed in being. Since there are, at present, no serious proposals about what heuristics might belong to such a bundle, it seems hardly worth arguing the point. (Fodor 1983, pp. 115-116)

Fodor appears to be insisting that the trouble with the idea of heuristic search is that it raises the hard question: *which* heuristics should be used to establish search hierarchies of inference rules?

"A good illustration of this method is in Holland, et al. 1986. Their "strength" parameters reflect the past usefulness of rules, and are used to constrain search.

It is unclear whether Fodor construes this as a domain-general process-and-form problem or as a domain-specific content problem. He seems to be asking for a domain-general answer when he calls for a "principled solution to the frame problem" (Fodor 1983, p. 116), although he doesn't attempt to explain the difference between principled and unprincipled solutions. Looked at this way, however, "serious proposals" about heuristics are a dime a dozen. I've just seriously proposed three principled heuristics, regarding the general reliability of an inference rule, its antecedent probability of applying, and its general usefulness. Of course, these principles leave open the various domain-specific problems about which inference rules are generally more reliable for dealing with the real world than which others, about which conditions in the real world are antecedently more likely to hold than which others, and about which inference rules are more likely to be useful than which others. Perhaps, then, Fodor is referring to the difficulty of these domain-specific "hierarchy problems".

How is a computer to establish the search hierarchies of inference rules necessary to solve hierarchy problems? Well, if we could set robots loose to gather data for themselves, they could rely on their own past experience, experience of which conditions have in fact obtained most often, or of which inference rules have in fact been most reliable and useful.

But, as I mentioned above, we are not currently able to do this. Typically, then, a system must rely on the hierarchies we program into it. Can Fodor argue that the solution to the "frame problem" escapes our grasp by swinging away on this loose end? After all, how do we know which hierarchies to program into a reasoning system? Alas, for many domains, we don't! Hierarchy problems are domain-specific content problems; to solve them, we have to do a lot of science. In this respect, hierarchy problems are no deeper than any other content problems, say, the "price problem": how are computers to know the prices of objects in the real world? Well, we've got to tell them, since we can't very well turn them loose to find out for themselves. And for us to know, we've got to split up and do a lot of domain-specific investigations: you've got to find out about the prices of wagons, I've got to find out about the prices of bombs, etc. Similarly with hierarchy problems: you've got to find out how often wagons malfunction, I've got to find out how often bombs are put on wagons, etc. If AI is ever to build a well-informed computer, it must incorporate the findings of experts in wildly diverse domains. The important point is that AI's "problems" of ranking conditions according to their relative probabilities, and of ranking rules according to their relative reliability and usefulness, are no more surprising or principled than its "problem" with specifying the prices of objects.

3.3.4 Summary

Before moving on, it may be helpful to summarize the main conclusions thus far. First, the relevance problem and the holism problem have nothing important to do with the frame problem as it is understood in AI, namely, the persistence problem. As a result, it is improper to use them in arguments against the sleeping-dog strategy. Second, the two problems, construed as domain-general problems, are easily solved by appeal to two familiar AI relieving techniques: bidirectional and heuristic search. Finally, although AI does not have a complete solution to certain domain-specific problems, the musings of the three wise men have not shown this to be a deep, epistemological problem: AI can simply incorporate the results of the domain-specific sciences.

3.4 Kookiness and the Frame Problem

3.4.1 The "Fridgeon" Problem

The third wise man, Jerry Fodor, raises a novel and interesting objection to the sleeping-dog strategy based on the kooky predicate "fridgeon", defined as follows: *x* is a fridgeon at *t* iff *x* is a physical particle at *t* and Fodor's fridge is on at *t*. Fodor points out that when he turns his

fridge on, he makes billions of changes--namely, he turns each particle in the universe into a fridgeon. Therefore, he argues:

> If I let the facts about fridgeons into my database . . ., *pursuing the sleeping dog strategy will no longer solve the frame problem.* . . . [A] strategy which says 'look just at the facts which change' will buy you nothing; it will commit you to looking at indefinitely many facts. (Fodor 1987, pp. 144-145; emphasis Fodor's)

The point is quite general. As Fodor explains, "there are arbitrarily many kooky concepts which can be defined with the same apparatus that you use to define perfectly kosher concepts," namely, the apparatus of "basic concepts" and "logical syntax" (Fodor 1987, pp. 145-146). "So," he continues, "the problem--viz. the FRAME problem--is to find a RULE that will keep the kooky concepts out while letting the non-kooky concepts in" (Fodor 1987, p. 146; emphasis Fodor's). But this would be tantamount to "a rigorous account of our commonsense estimate of the world's taxonomic structure," which would require "formalizing our intuitions about inductive relevance" (Fodor 1987, pp. 147-148). It's no wonder, then, that Fodor claims the frame problem is "too important to leave to the hackers" (Fodor 1987, p. 148)!"

"Some of Fodor's readers have been struck by similarities between his "fridgeon" problem and Nelson Goodman's infamous "grue" problem (Goodman, 1965). I will discuss the relevance of the "grue" problem in section 3.3.4.

3.4.2 Three Kinds of Memory

Before turning directly to Fodor's problem of formalizing inductive kookiness, it will help to get clearer about what a system should do in the face of kookiness. What I will argue is that a system should represent kooky facts implicitly in its representations of nonkooky facts. The basic idea can be explained by reference to the way people (like yourself) deal with the mental predicate "FRIDGEON". If Fodor is right, then you must keep representations of fridgeon facts out of your "database". But this doesn't mean you must keep the *definition* of "FRIDGEON" out of your memory; if you did, you wouldn't even be able to understand Fodor's argument! On a natural view, then, you must have something like a mental dictionary, in which you can store the definition of "FRIDGEON". (For simplicity, we can suppose that this dictionary is wholly separate from the database of "facts", although it is not necessary for my purposes.) If (for some odd reason) you want to check whether Nancy-the-Neutron is a fridgeon, you must first find "FRIDGEON" in your mental dictionary, and then check your database to determine whether Nancy satisfies the definition--that is, whether Nancy is a particle and whether Fodor's fridge is on. Given that "FRIDGEON" appears in your mental dictionary, then, representations of fridgeon facts needn't appear in your

database. So you don't need to update them when you discover that Fodor has turned his fridge on. The same is true for an AI system with both a dictionary and a database. When Fodor turns his fridge on, the system only needs to change *one* representation, namely, its representation of the state of Fodor's fridge.

The most obvious objection to this strategy is that even representations of fridgeon facts must sometimes be explicit. Otherwise, one could never use the predicate "FRIDGEON", as you are in thinking about Fodor's argument. In the example, once you find "FRIDGEON" in your dictionary, and check whether Nancy satisfies the definition, you still must infer explicitly that Nancy is a fridgeon. In other words, apparently, you must put the representation "NANCY IS A FRIDGEON" in your database. Since this representation is explicit, however, it needs to be updated explicitly, when Fodor turns his fridge on. It might seem, then, that the distinction between the dictionary and the database cuts no ice. The proper response to this objection is to appeal to a third kind of memory, which cognitive scientists call "working memory". Working memory is a temporary storage space, for representations which are being used at a given time. The important thing about working memory for present purposes is that once representations in working memory are used, they can be erased. Now, while it is true that fridgeon

facts sometimes need to be represented explicitly, they need only be explicit in working memory, not in the long-term database. Therefore, after generating and using the explicit representation "NANCY IS A FRIDGEON", you can simply erase it, without worrying about updating it. The same is true for an AI system with a working memory.

But Fodor can also object to this. The situation is different when a system is *told* that Nancy is a fridgeon--that is, when this is new information. If the system simply erases this representation from working memory, it will lose the information about Nancy. So, apparently, it must first *copy* the representation into the database, in which case it needs to worry about updating the copy. The response to this objection is simple. If the system is to keep fridgeon facts out of the database, it must translate representations of them into nonkooky representations (using the dictionary), and copy these nonkooky representations into the database. So, when a system is told that Nancy is a fridgeon, it should put the representations "NANCY IS A PARTICLE" and "FODOR'S FRIDGE IS ON" into the database.

3.4.3 How to Rule out Kooky Predicates

Even given the viability of keeping kooky mental predicates in the dictionary and in working memory, the "fridgeon"

problem has not been addressed. For how does a system know which predicates to keep there, and which to allow into the database? Mustn't it follow a rule which, as Fodor claims, codifies "our intuitions about inductive relevance"? Not obviously. I agree with Fodor that no one knows how to formalize inductive kookiness, but I disagree with his claim that we need to do this in order to save the sleeping-dog strategy. As Fodor himself insists, kooky predicates are defined in terms of basic predicates, so representations involving kooky predicates can always be left implicit in representations involving only basic predicates. Suppose, then, that a system follows this rule: allow only *basic* predicates into the database, and keep all *defined* predicates in the dictionary and in working memory.[100] Even though this

---

[100]Strictly speaking, this rule may have to be modified in order to be implemented in familiar models. The distinction between basic and derived predicates is a *semantic* distinction of considerable intricacy (see section 2.1.2). Given this, it is difficult to see how a computational system could classify its predicates literally as basic or as derived. However, a number of computational tests can combine to generate a useful classification which is near enough to the basic/derived distinction: call it the distinction between "quasi-basic" and "quasi-derived" predicates.

The first step is to classify predicates as either syntactically complex or syntactically simple. Since this is a formal distinction, nearly enough, I presume it is not difficult to understand how a system could draw it, nearly enough. Syntactically complex predicates may be classified as quasi-derived.

The harder task is to draw a distinction which can do duty for the distinction between basic and derived syntactically simple predicates. It may appear that we need a test for when a syntactically simple predicate (e.g., "fridgeon") is literally *defined* by a syntactically complex predicate. However, all that is actually required is that there be a test for syntactically simple predicates for which the system is *disposed to substitute* a particular syntactically complex predicate, in any situation the system judges to be both likely and important. (This seems to be the situation we find ourselves in when Fodor tells us that

rule does not formalize kookiness, it is generally applicable to any kooky predicate Fodor chooses to define.

Call a system using this rule a "basic system", since all of its inferential processes are carried out over representations involving only basic predicates. Although a basic system does not need to appeal to inductive relevance in order to exclude kooky predicates, if it is to count as well-informed about the real world then it needs to know which particular "basic representations" to infer from which particular others. Call this the "basic learning problem". It may appear that my appeal to basic systems simply begs Fodor's questions, since the basic learning problem is similar to Fodor's problem of formalizing inductive relevance.[101] If this *is* Fodor's question, however, it deserves begging, for it is deprived of any interest. Given the possibility of basic systems, Fodor cannot support his (interesting) claim that the sleeping-dog strategy raises special problems about kooky predicates. All he can claim, then, is that the sleeping-dog strategy must work hand-in-hand with a solution

---

"*x* is a fridgeon" *means* "*x* is a particle and Fodor's fridge is on".) In many cognitive models such dispositions are realized in such a way (e.g., as "strengths" of productions in production systems or as "weights" of connections in connectionist networks) that processes may access and modify them. I presume, then, that it is possible for computational processes to access these dispositions in distinguishing quasi-basic from quasi-derived predicates.

[101] I thank Joelle Proust for pressing this point.

to the basic learning problem. But this is no surprise. The basic learning problem arises for *any* system which has to make inferences about the real world, whether or not it uses the sleeping-dog strategy (compare the discussion of C3 in section 3.3.2). Therefore, the sleeping-dog strategy does not raise or intensify the problem. More importantly, Fodor has not shown any principled difficulties with solving the problem. If we want well-informed robots, then we must do two things: we must engage in lots of domain-specific scientific investigations about what may be inferred from what, and we must occupy ourselves with issues surrounding how machines can learn as children do. The basic learning problem is a familiar example of an unprincipled, domain-specific content problem (see section 3.3.2).

Another objection is that the rule which defines basic systems is a bit too strong. It not only keeps kooky predicates out of the database, but also excludes nonkooky defined predicates, like "MY BULGARIAN GRANDMOTHER" and "VEGETABLE CRISPER". The problem is that if one often uses these predicates, one might need to have representations involving them at one's mental fingertips--that is, one might need to have them explicit in the database. In other words, it might take too much time and energy to deal with all the basic predicates each time one needs to use one of these complex predicates. Fair enough. The rule needs to be

weakened in the following way: allow only representations involving basic predicates into the database, except for representations (involving defined predicates) that are so useful that you need to have them at your fingertips. In other words, when a particular combination of basic predicates recurs very frequently in the course of problem solving, the system may introduce into the database an abbreviation for it (i.e., a derived predicate). As amended, however, the rule needn't mention anything about "our commonsense estimate of the world's taxonomic structure."

One last argument on behalf of Fodor: he can object that weakening the rule may allow fridgeon facts back into the database after all. If individual fridgeon facts were (somehow) vitally important to a system, it might indeed need to have fridgeon information at its fingertips. But then it would be forced to update many representations when Fodor turns on his fridge. This is true. For such a system, however, "FRIDGEON" would not be a kooky predicate at all--at least, it would not be something the system *should* want to rule out of the database! A system with kooky enough needs *would* have to update indefinitely many beliefs; that's just tough kookies.[102] The sleeping-dog strategy is not supposed

[102]Even if a system had the kooky need to allow fridgeon facts into the database, it would not necessarily run into computational problems. Once the system has determined that Fodor's fridge is on, it can form the *one* explicit representation "*ALL* PARTICLES ARE FRIDGEONS". When it changes its representation of Fodor's fridge, it can change this *one*

magically to eliminate this possibility, but only to help minimize the number of updates, given a fixed set of needs. I conclude, then, that Fodor has not shown that the sleeping-dog strategy faces a problem about formalizing our intuitions about inductive kookiness.

3.4.4 The New Riddle of Induction

Although Fodor doesn't explicitly mention Nelson Goodman's "new riddle of induction" (the "grue" problem), he certainly nudges the reader toward the conclusion that AI must solve it to solve the frame problem. As I have mentioned, he (incorrectly) touts his "fridgeon" problem as a problem about "inductive relevance". This can make it seem similar to the problem of showing why "grue" facts are irrelevant to inductive inference. Elsewhere, Fodor writes, "the frame problem is not distinguishable from the problem of nondemonstrative confirmation" (Fodor, 1983, p. 138). To philosophers, at least, the "grue" problem is the paradigm

---

representation. Surely this is computationally feasible! But Fodor can object that even this move leaves implicit all the billions of *particular*-fridgeon facts, say, the fact that Nancy-the-Neutron is a fridgeon, and so on. In order to establish that Nancy is a fridgeon, the system has to infer this from the general fact about all particles. Perhaps each one of the billions of fridgeon facts might be *so* vital to a system's microsecond-to-microsecond welfare that, in the heat of the moment, it cannot afford to waste time on this inference. Such a system would be very likely to fail. But if *this* is Fodor's frame problem, it should simply be ignored. Any human would fail, too, in the same situation. So AI could design a machine as smart as any human, without running up against this problem.

paradox of nondemonstrative confirmation. Whatever his intentions on the matter, then, it is common for Fodor's readers to come away with the distinct impression that he has shown the frame problem to be or to include the "grue" problem.

For all its distinctness, this impression is wrong. Although Fodor's problem and Goodman's problem are both problems about kooky predicates, they are only superficially similar. "Grue" is, like "fridgeon", a kooky predicate, which may be defined as follows: *x* is grue if and only if *x* is a green thing that has been examined or *x* is a blue thing that has not been examined. Suppose that every emerald ever examined is green. According to the definition, then, every emerald ever examined is grue. The trouble with "grue" is that, unlike the nonkooky predicate "green", it is not "projectible" to unexamined emeralds. "Green" is projectible, since the fact that all examined emeralds are green helps confirm the hypothesis that all unexamined emeralds are green. But the fact that all examined emeralds are grue does not help confirm the hypothesis that all unexamined emeralds are grue; by the definition, this would amount to the hypothesis that all unexamined emeralds are *blue*. The problem Goodman poses, then, is to show why "grue" and its like are not projectible from examined instances to unexamined instances, while "green" and its like are projectible.

By contrast, Fodor's problem simply has nothing to do with projectibility, since "fridgeon", unlike "grue", is perfectly projectible. For if at least one particle is a fridgeon, it follows that Fodor's fridge is on, so it follows that all particles are fridgeons. Therefore, even if Fodor is right that the sleeping-dog strategy converts the frame problem into a serious problem about inductive relevance, it would not follow that the frame problem would include the problem of avoiding the projection of "GRUE"-like mental predicates.[103]

Whether or not Goodman's problem has anything to do with the frame problem, AI must solve something like it in order to design good reasoning systems. Cast in Fodor's computational terms, the problem is to find a "rule" that will keep kooky predicates (like "grue") from being projected while letting nonkooky predicates (like "green") be projected.

---

[103]Conversely, the mental predicate "GRUE" does not create the (alleged) problems "FRIDGEON" does. Using "FRIDGEON" rather than nonkooky predicates leads to Fodor's problem, because doing so (apparently) increases the number of representations one must update. Using "GRUE" rather than "GREEN", on the other hand, does not (even apparently) increase the number of updates. When one observes an object *x*, one can classify it as "GREEN" by storing the single representation "*x* IS GREEN". Even if one instead classifies it as "GRUE", however, one need only store the single representation "*x* IS GRUE". (One wouldn't *also* need to store "*x* IS GREEN", since this may be left implicit in "*x* IS GRUE" and "*x* IS OBSERVED".) Likewise, one can project "GREEN" to an unobserved object *y* by adding the single representation "*y* IS GREEN". Even if one instead projects "GRUE", however, one need only add the single representation "*y* IS GRUE". (One wouldn't also need to add "*y* is BLUE", since this may be left implicit in "*y* IS GRUE" and "*y* IS UNOBSERVED".

Fortunately, AI can follow Goodman's own lead in finding such a rule (Goodman, 1965, chapter 4). Goodman's proposal centers around the "entrenchment" of a predicate. Roughly, a predicate is entrenched in proportion to the number of times it and its superordinates (e.g., "colored" for "green") have actually been projected in the past. Again roughly, Goodman suggests that a predicate (like "green") is projectible if it is much more entrenched than any "competing" predicate (like "grue"). Cast in terms of a "rule", what Goodman is saying is, roughly: "Project the predicates you've projected often in the past, and don't project ones which lead to conflicting projections". It is hard to believe that AI would run into much difficulty with implementing this sort of rule! To exclude "grue", for instance, all that needs to be done is to endow a system with an "innate" disposition to project "green" (an endowment presumably shared by humans), and a disposition to favor entrenched predicates over conflicting, nonentrenched predicates.

If Goodman's "grue" problem sends shivers down philosophical spines, this is probably due to the worry that Goodman's appeal to entrenchment merely *describes* our inductive practices, without *justifying* them. Why *should* we project "green" rather than "grue"? Because, he says, this coheres better with our established practices. Surely this is reasonable; it follows from a more general principle of

conservation of existing theories and practices until something goes positively wrong with them. But the worry is that even if we *had* normally projected "grue" much more often than "green", this practice would still have been less optimal than our existing practice. Fortunately for AI, however, even if Goodman *does* need to say more to dispel this worry, AI doesn't. AI might succeed in building a system with our inductive practices, without even beginning to justify these practices. As Pat Hayes writes in response to Fodor, "We [in AI] wouldn't need to solve the philosophical problem of other minds in order to build a conversational program, or the problem of the *ding an sich* in order to build a vision system, either" (Hayes[1987], p. 134). The musings of the third wise man, far from leading to a new, deep problem for AI, leads to an old, deep problem for philosophy.

3.4.5 Summary

In two respects, the "fridgeon" problem shares the fate of the relevance problem and the holism problem. First, none of them are properly identified with the frame problem as it is understood in AI (i.e., with the persistence problem), and none of them weigh against the sleeping-dog strategy. Second, they are all easy to solve. Therefore, neither Dennett nor Haugeland nor Fodor succeeds in demonstrating a deep, difficult problem for AI. However, they are left with a deep,

difficult problem of their *own*, namely, the problem of framing the frame problem: why should one suppose that what they are talking about is the *frame* problem, and why should one suppose that it's a *problem*?

In the absence of a complete classical model of cognition, it does not appear possible to eliminate the suspicion that such models will prove incapable of implementing human inference in all of its complexity. Indeed, this suspicion is healthy, when (in the hands of talented, sensitive, wise men and women) it gives rise to attempts to demonstrate principled problems for cognitive science. Such problems have an important role in the development of a science--they help to generate solutions, and provide for an understanding of the explanatory role of these solutions. I have tried to isolate a number of complexity-relieving techniques which solve the frame problems in their various guises. It is an empirical, and entirely open, question whether these techniques suffice to reduce complexity to such a degree that classical models can be implemented in brains. The ultimate fate of functionalism and token physicalism, as an approach to the mind/body problem, hangs in the balance.

REFERENCES

Agre, P., 1989: "The Dynamic Structure of Everyday Life", Ph.D. thesis, MIT A.I. Lab.

Anderson, J., 1983: The Architecture of Cognition, Harvard University Press, Cambridge.

Armstrong, D., 1983: What Is a Law of Nature?, Cambridge University Press, Cambridge.

Barr, A., and E. Feigenbaum, ed., 1981: The Handbook of Artificial Intelligence, vol. 1, William Kaufmann, Inc., Los Altos, CA.

Barwise J., and J. Perry, 1981: Situations and Attitudes, MIT Press, Cambridge.

Benacerraf, P., 1965: "What Numbers Could Not Be", Philosophical Review.

Block, N., ed., 1980: Readings in Philosophy of Psychology, Harvard University Press, Cambridge.

Block, N., 1986: "Advertisement for a Semantics for Psychology", in Midwest Studies in Philosophy, vol. 9.

Carnap, R., 1947: Meaning and Necessity, University of Chicago Press, Chicago.

Chalmers, D., 1990: "Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation", manuscript.

Clark, A., 1989: Microcognition, MIT Press, Cambridge.

Cresswell, M., 1985: Structured Meanings, MIT Press, Cambridge.

Cummins, R., 1983: The Nature of Psychological Explanation, MIT Press, Cambridge.

Cummins, R., and G. Schwarz, 1987: "Radical Connectionism", Southern Journal of Philosophy (supplement).

Cummins, R., 1989: Meaning and Mental Representation, MIT Press, Cambridge.

Davidson, D., 1969: "The Individuation of Events", in Essays on Actions and Events, Oxford University Press, Oxford, 1980.

Davidson, D., 1975: "Thought and Talk", in Inquiries into Truth and Interpretation, Oxford University Press, Oxford, 1984.

Davies, M., 1989: "Concepts, Connectionism, and the Language of Thought", Proceedings of the International Colloquium on Cognitive Science.

Dennett, D., 1984: "Cognitive Wheels: the Frame Problem of AI", in Pylyshyn, ed., 1987.

Dennett, D., 1983: "Styles of Mental Representation", in The Intentional Stance, MIT Press, Cambridge, 1987.

Dretske, F., 1977: "The Laws of Nature", in Philosophy of Science.
Dreyfus, H., 1979: What Computers Can't Do, Harper Books, New York.
Dreyfus, H. and S. Dreyfus, 1986: Mind Over Machine, Free Press, New York.
Evans, G., 1983: The Varieties of Reference, Oxford University Press, Oxford.
Fikes, R., and N. Nilsson, 1971: "STRIPS: A New Approach to the Application of Theorem proving to Problem Solving", Artificial Intelligence.
Fodor, J., 1968: "The Appeal to Tacit Knowledge in Psychological Explanations", in Fodor, 1981.
Fodor, J., 1975: The Language of Thought, Harvard University Press, Cambridge.
Fodor, J., 1981: Representations, MIT Press, Cambridge.
Fodor, J., 1983: The Modularity of Mind, MIT Press, Cambridge.
Fodor, J., 1987a: "Why There Still Has to be a Language of Thought", in Psychosemantics, MIT Press, Cambridge.
Fodor, J., 1987b: "Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres", in Pylyshyn, ed., 1987.
Fodor, J. and Z. Pylyshyn, 1988: "Connectionism and Cognitive Architecture", in Cognition.
Goodman, N., 1965: Fact, Fiction, and Forecast, Bobbs-Merrill, Indianapolis.
Haugeland, J., 1985: Artificial Intelligence: the Very Idea, MIT Press, Cambridge.
Haugeland, J., 1982: "The Nature and Plausibility of Cognitivism", in The Behavioral and Brain Sciences.
Haugeland, J., 1987: "An Overview of the Frame Problem", in Pylyshyn, ed., 1987.
Hayes, P., 1987: "What the Frame Problem Is and Isn't", in Pylyshyn, ed., 1987.
Holland, J., K. Holyoak, R. Nisbett, and P. Thagard, 1986: Induction, MIT Press, Cambridge.
Janlert, L., 1987: "Modeling Change--The Frame Problem", in Pylyshyn, ed., 1987.
Johnson-Laird, P., D. Herrmann, and R. Chaffin, 1984: "Only connections: A critique of semantic networks", Psychological Bulletin.
Lewis, D., 1972: "General Semantics", in Philosophical Papers, v. 1, Oxford University Press, Oxford.
McCarthy, J., and P. Hayes, 1969: "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in B. Meltzer and D. Michie, eds., Machine Intelligence 4, Edinburgh University Press, Edinburgh.
Preston, E., 1988: "Representational and Non-Representational Intentionality", Ph.D. thesis, B.U. Philosophy Dept.
Pylyshyn, Z. (ed), 1987: The Robot's Dilemma: the Frame Problem in Artificial Intelligence, Ablex, Norwood, NJ.

Rumelhart D., and J. McClelland, 1982: "An interactive activation model of context effects in letter perception", in Psychology Review.
Rumelhart D., G. Hinton, and J. McClelland, 1986: "A General Framework for Parallel Distributed Processing", in J. McClelland and D. Rumelhart, ed., Parallel Distributed Processing, v. 1, MIT Press, Cambridge.
Searle, J., 1983: Intentionality, Cambridge University Press, Cambridge.
Schwarz, G., 1987: "Explaining Cognition as Computation", M.A. Thesis, University of Colorado Philosophy Dept.
Shoham, Y., 1988: Reasoning about Change, MIT Press, Cambridge.
Singley, M., and J. Anderson, 1989: The Transfer of Cognitive Skill, Harvard University Press, Cambridge.
Smolensky, P. 1988: "On the Proper Treatment of Connectionism", in The Behavioral and Brain Sciences.
Smolensky, P. 1989: "Connectionism, Constituency, and the Language of Thought", manuscript.
Stabler, E. 1983: "How Are Grammars Represented?", in The Behavioral and Brain Sciences.
Stalnaker, R. 1984: Inquiry, MIT Press, Cambridge.
Stich, S. 1984: From Folk Psychology to Cognitive Science, MIT Press, Cambridge.
Tooley, M. 1989: Causation, Cambridge University Press, Cambridge.
van Gelder, T. 1989: "Compositionality: Variations on a Classical Theme", manuscript.