Information, Computation, and the Nature of Cognition: A Critique of Computational Approaches to Understanding and Creating Minds

by

Michael Karl Wilhelm Happold
December 11, 2000

A dissertation submitted to the Faculty of the Graduate School of State University of New York at Buffalo in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
Department of Philosophy

Major Professor: Barry Smith, Ph.D.

Table of Contents

List of Figures and Equations

Abstract

Information, Computation, and the Nature of Cognition: A Critique of Computational Approaches to Understanding and Creating Minds

Cognitive scientists generally subscribe to an information-processing model of mind and implement this model through computational methods. Information processing is understood to be generation and composition of informational primitives into more complex pieces of information in response to signals extracted from the environment. Computationalism offers a powerful methodology for carrying out information processing, with abstract tokens standing in for pieces of information, and new information structures being created through application of rules. This purely syntactic model of cognition is unable, however, to explain the nature of semantic information. Modifications of Shannon information theory and applications of principles of natural selection fail to provide a non-syntactic account of the nature and origin of semantic information. Purely syntactic explanations of semantic information fail to capture the representational capabilities of true cognitive agents. Relying on computational explanations of cognitive behavior leads not only to an explanatory gap, but also to practical failings. These failings constitute the specific and general frame problems. To avoid these problems, a model of cognition that focuses on the adaptive capabilities of its realizing hardware must be adopted. Evolutionary and ecological models only provide pieces of a final theory. A high-level model for cognition can be found in the field of complex and self-organizing systems, a branch of

dynamic systems theory. One possible characterization of the lower-level mechanisms responsible for the high-level behavior is given by Edelman's Theory of Neuronal Group Selection. What emerges is a theory with strong similarities to behaviorism and teleological functionalism, although without the deficiencies of either theory. This theory offers the foundation for a new understanding of the nature of representation and information.

Introduction

If there is any view that unites Dynamic Systems Behaviorism the disparate fields gathered under the banner of cognitive science, it is that the mind/brain is an information processing machine. Defenders of folk psychology insist that the mind must be understood in terms of propositional attitudes and their content. Functionalists contend that the mind consists of a series of algorithms, gathering information and running computations on the assembled data. Connectionists disagree with functionalists about how this information is stored and processed, but not about whether information processing is what is going on when we think. Direct realists go so far as to claim that the world is pre-labeled, and that the mind is just an information gatherer in its relation to the world. Even eliminative materialists, who are willing to discard the trappings of folk psychology and functionalism, believe that the mind/brain is just a highly sophisticated information processor. Painting a discipline with such a broad brush is likely to cover over nuances of specific theories, but the underlying unity of cognitive science is only to be found in the overarching theme of information processing.

This theme has been popularly embodied in the computer metaphor of mind. While this metaphor is not embraced by all cognitive scientists, it is emblematic of cognitive science's devotion to the notion of the mind/brain as information processor. In keeping with this metaphor, cognitive science has followed a research strategy that focuses primarily on the formal properties of mental states, ignoring to a large degree their environmental context

and evolutionary considerations. Many cognitive scientists accept methodological solipsism as their research methodology, and those who do not do so explicitly often implicitly adhere to its dictates. They are encouraged in their efforts to study in isolation specific aspects of the mind/brain by the many successes that this divide-and-conquer strategy has yielded in the past few decades. Why tamper with success?

This dissertation takes direct aim at the notion that the mind/brain ought to be studied as if it were a computer. The mind/brain is not a computer because it is not an information processor in the sense proposed by cognitive scientists. Proving this is currently an impossible task, for it would require a fully developed theory of the mind/brain. Nonetheless, the information processing theory of mind harbors inherent flaws that prevent it from achieving its own comprehensive explanation of consciousness. The disciplines of cognitive psychology, linguistics, and neurobiology have all made great strides while embracing information processing, and while avoiding the presentation of a theory of consciousness. Although this avoidance has been due in part to reasonable scientific caution, I will argue that a more important reason is that the special characteristics of the mind/brain distinguishing it from computers stand in the way of the progress of current information processing theories.

The information processing approach is implemented by cognitive scientists primarily along computationalist lines. Computationalism is the view that cognition is computation, or at least realizable as computation even if not actually implemented as such in living organisms. Computation is the application of formal rules over abstract tokens that act as symbols, and computationalism is the natural expression of the information processing approach because it allows for a direct mapping between chunks of information and these

tokens. The informational structure of thought is therefore decomposable into its primitive elements, and what role a piece of information plays in cognition is determined by the rules the token it maps to takes part in.

Even if computationalism enables researchers to analyze how information is expressed and used in cognitive systems, it nonetheless brings with it severe limitations. Computational systems are plagued by a host of problems. These include lack of adaptability to their environment and inability to handle anything other than toy or severely constrained worlds. These practical problems are indicative of theoretical problems, including the specific and general frame problems and the symbol-grounding problem.

To take computationalism's place, I will sketch an evolutionary-ecological theory of the mind/brain, drawn in part from Gerald Edelman's (1987, 1989) Theory of Neuronal Group Selection (TNGS) and from recent developments in dynamic systems theory. While many of Edelman's arguments against information processing models are flawed, he has developed a deeply suggestive theory of the brain's development and functioning that calls into question the validity of the computer model of mind.

Whereas Edelman's TNGS offers an account of the neural principles that drive the organization of the brain, dynamic systems theory provides an explanation of the relation between the local neural interactions and global, psychological states. These two theories are combined to provide an account of mind that eschews the information processing model of mind embodied by computationalism, replacing it with a model based on the special behavior characteristic of neural systems.

Among philosophers who have taken issue with the computational theory of mind and the information-processing model underlying it, Searle stands out as one of the few who

has connected the flaws of these two theories. His Chinese Room argument provides the underlying intuition of this dissertation that there is something missing from the attempt to understand cognition in terms of syntactic information processing, and this missing element is a fully developed theory of semantic information. Searle's own positive theory, however, does little to unravel the mystery of cognition and consciousness, merely declaring these to be biological phenomena. Searle is right to turn to biological phenomena to understand cognition, but wrong to imagine that it is something special about biological matter rather than mechanisms that gives brains their unique qualities.

Timothy van Gelder's exposition of a dynamic systems alternative to computationalism is both the inspiration and departure point for the dynamic turn advocated in this dissertation. Van Gelder has drawn up the preliminary sketch that others have tried to fill in—among them, Horgan and Tienson. Cognitive systems are but a subset of dynamic systems, and van Gelder, in illustrating the differences between computational and dynamic systems, has glossed over important distinctions within the set of dynamic systems that make cognition possible. The question to be answered is what is it about biological dynamic systems that gives rise to cognition. This dissertation attempts to begin answering this question.

Chapter 1 examines the information processing model and various attempts to develop a theory of semantic information. Among these are Dretske's (1981) theory of natural information, the molecular Darwinist (Eigen 1992; Küppers 1990) account of biological information, and Sayre's (1986) application of the concept of mutual information from Shannon information theory. Each is found wanting, exposing the explanatory gap in cognitive science between syntactic and semantic information.

Chapter 2 analyzes computationalism and its numerous incarnations to determine whether they can explain the nature of cognition. Key to this analysis is Searle's (1980) Chinese Room argument and the efforts of computationalists to counter it. The dynamic systems critique of computationalism is introduced here as well.

Chapter 3 focuses on connectionist claims to have overcome the deficiencies in traditional computationalist approaches to modeling cognition by using important principles of neural architecture and behavior. The claim to exceptionalism is examined in the light of connectionist successes, and is found to be exaggerated. One principle of connectionism, the superposition of representations, is found to be an important exception to traditional computationalist methods.

Chapter 4 presents the specific and general frame problems, as well as an argument for how they arise from computationalist methods. Recent efforts at solving the specific frame problem are considered, and although they present concrete progress, they do not provide keys toward solving, and in fact exacerbate, the general frame problem. The principle of the superposition of information is identified as a possible way out of the frame problem.

Chapter 5 examines externalist alternatives to the computationalist understanding of information: Millikan's version of teleological functionalism, Dretske's Representationalism, and Gibson's ecological approach. The former two ground their externalism in the evolutionary history of organisms, whereas the latter locates information external to the observer in the optic array and its relation to the environment. Evolutionary externalism bears severe defects that make it unworkable as a theory of semanticity. On the other hand, ecological externalism, in it more moderate form, offers promising ideas as to how organ-

isms perceive and cognize the structure of their environment. A brief look is taken at Tye's Representationalism, a more comprehensive view being given in the final chapter.

Chapter 6 reintroduces the dynamic systems critique of computationalism and explains how it offers an alternative for understanding the nature of mind. Edelman's (1987) Theory of Neuronal Group Selection is presented to fill out the details left blank by dynamic systems theory, and a similar move by Thelen and Smith (1994) is examined. Finally, an alternative understanding of the relevance of dynamic systems theory to explaining the nature of mind offered by Horgan and Tienson (1996) is critiqued.

Chapter 7 provides a sketch of a theory of mind, Dynamic Systems Behaviorism, and demonstrates how it avoids various objections leveled against behaviorism and functionalism. Finally, the nature of representation and information is reexamined in the light of this theory, with a comparison of Dynamic Systems Behaviorism to Michael Tye's version of Representationalism.

Chapter 1 Cognition as Information Processing

What do DNA, the spiking of neurons, newspaper reports, and computer databases all have in common? Even if one has but dabbled in the disciplines that are thought to comprise cognitive science, the answer is obvious: They are all forms of information. While cognitive science's overarching theme is, as the name denotes, the study of cognition, it is already assumed that cognition is a species of information processing. What is so thrilling about the information processing model of mind is the connection it makes to two very different fields: computer science and evolutionary biology. The former holds out the promise that mind can be made silicon, the latter is an account of how the mind might have been made flesh. If the mind is an information processor, then computers not only can help us model its behavior but also allow us to create artificial minds. And if DNA and thoughts are both units of information, then perhaps the same mechanisms of selection that apply to the former also apply to the latter. Both approaches demystify the mind by removing its almost magical uniqueness and explaining its functioning in mundane terms. Information processing in the form of DNA to RNA transcription existed long before minds. We should not be too surprised that Nature found a way to do real-time transfer and processing of information. Nor should we be too surprised to find that we can reconstruct Nature's achievement in the form of silicon, copper, and plastic.

More than any other general perspective on cognition, the idea that the mind is an information processor unites the disputing camps of cognitive science. While proponents of picture-theories of representation eschew computational explanations (Kosslyn 1983), and

symbolic AI is under assault by connectionists, all agree that the mind is processing information. So if it is a scandal, as Andy Clark concedes (1997), that cognitive scientists have yet to agree on what computation is, then it is an absolute outrage that there is no commonly agreed upon formulation of what information is. In fact, it is a rarity to see even an informal definition attempted in works explicitly presenting information-processing models of mind. So what is it that these researchers are agreeing upon?

There is a commonsense understanding of information that we all share. Information is a fact or set of facts about the world. Newspapers provide us with information insofar as they tell us what is going on in the world. But this commonsense understanding is not without ambiguity, an ambiguity too severe for it to be the basis for a research program. People often speak of being provided 'false information', which translates into 'false facts' if we are to abide by the above definition. In everyday speech, 'information' stands in at times for 'facts', and at other times for 'representations' or 'communications'. That its meaning is ambiguous is of little consequence to the purposes of everyday communication.

At the opposite extreme is Claude Shannon's (1949) definition of information, a mathematical formulation that expresses informational content in terms of the syntactic novelty of a signal or communication. Shannon Information Theory focuses exclusively on the syntactic aspects of a signal, and in so doing makes little contact with the sense of the term 'information' used by cognitive scientists. For a signal to have informational content according to Shannon Information Theory, it need not be 'about' anything. Similarly, Algorithmic Information Theory confines itself to considerations of encoding possibilities for signals and not what the signals mean.

Neither of these meanings of information expresses precisely what the cognitive scientist intends when he speaks of information-processing mechanisms in the mind, although each captures certain aspects. Information, as the commonsense notion would have it, does relate in some way to the production of representations. Yet information processing, as a scientifically studied phenomenon, should be explicable in mechanistic, quantifiable terms. The cognitive scientist looks to explain the origin and transfer of information, so he can't rely on the ambiguities of folk psychology. But this information is inherently 'about' the world; it is what an organism possesses when we say it has knowledge. Mathematical theories of communication do little in the way of explaining the origins of knowledge.

In this chapter, I will show how the few efforts at explaining the nature of information and information processing made by cognitive scientists have failed to address the most serious issues. This failure has been caused, and concealed, by a systematic conflation of meanings of the term 'information'. The various meanings of the term 'information' correspond to various levels of analysis of the mind/brain's functioning, as well as to connections with fields such as molecular biology. In their efforts both to explain the nature of DNA and neural functions to a lay public and to unify the sciences that investigate the mind/brain, researchers have ignored significant distinctions between levels of information processing. It is this conflation that has led to the problem of the homunculus in explanations of the mind/brain, as well as the misguided efforts to free cognitive science of this problem by casting the mind/brain as an evolved computer. So long as a folk psychological notion of information processing is imputed to neurophysiological structures, the homunculus problem will remain. This applies to computationalist approaches as well, though this is a topic for a later chapter. For now, I confine myself to showing that the notion of information that

cognitive scientists assume is at best too ambiguous to serve their purposes, and at worst self-referential. The latter point I take to be the crux of John Searle's criticisms of imputing intentionality to computers. On the former point, I can but express astonishment that so few have asked the question: what does it mean for a mind to be an information processor? The answers that have been given yield a notion of information that leaves indeterminate the difference between information-processing and non-information-processing systems.

1.1 What Is Information?

The literature of cognitive science provides a number of possible definitions of the term 'information':

1. Information is a measure of the novelty of a message; this is found more in accounts in Information Theory, though it is often used as a foundation for semantic notions of information, such as that found in (Dretske 1981).
2. Information is a causal/structural relationship between micro- and macrostates. Eigen and his colleagues (Eigen 1992; Küppers 1990) have presented this definition to describe the DNA/Cell-structure relationship, although Dennett (Dennett 1995) conflates it with his use of the term.
3. Information is the content of representation (Sterelny 1990).
4. Information is a fact or set of facts about the world. This corresponds roughly to the common-sense notion of information. It is not the same as a true representation, because the information lies in the object or event appearing to a sentient being, not

in its mind. Thus, a newspaper carries information regardless if anyone reads it (see Dennett 1998 for an analogy between information and ore, needing to be extracted and refined).

5. Information is a *normal* causal relationship between the way the world is, and what goes on in a sentient being's mind (Dretske 1981).

6. Information is a veridicality relationship between object and representation, with veridicality defined in terms of Shannon's notion of high mutual information (Sayre 1986).

Perhaps the most unenlightening definition of information is that of Cummins, who, in defining an information-processing system states:

> an *information processor* is simply a symbol manipulator. *Symbols* are distinguished from other things by the possibility or actuality of systematic semantic interpretation. To alter symbols is therefore to move from one meaning to another, and that is why symbol manipulation is information processing (Cummins 1983, 34).

Either Cummins has merely stated the truism that information processing is the movement from one meaning to another (since symbols are meaningful by definition), or he is attempting to define semanticity in terms of syntax. What remains to be answered is how symbols acquire semantic interpretation *within* the system bearing them. A programmer can give a systematic semantic interpretation of the widget detector that he programmed, but does it follow that the widget detector's states have semantic content? If so, does a flow diagram of the widget detector, which also can be given a semantic interpretation, have semantic content?

The common, implicit definition of information seems to be as stated in 4. Pieces of information are facts about the world. Sometimes these facts emanate as waves, such as

sound waves. The waves themselves are not information. They are just physical phenomena. But they *carry* information. Information is that which *could* be deciphered from the wave about the world in which it originates. There need not be any beings capable of deciphering and representing the information, but it is carried nonetheless. To rephrase a philosophical cliché, if a tree falls in the forest and there is no one around to hear it, the sound waves it causes do carry information.

This informal definition of information gives rise to significant problems. It leaves the question 'What is information?' unanswered; is information a physical property of a process or entity, and if so, what is the nature of that property? Do all physical processes/entities carry information, and how much? I will examine one attempt to answer this latter question, Shannon Information Theory, and demonstrate that it does not serve cognitive science's purposes. Shannon Information Theory enables us to quantify the informational content of a signal in terms of the signal's syntax. It does not enable us to quantify the semantic content of a signal. Shannon information theory tells us about the probability of a set of symbols occurring as a message. It does not tell us the range of what these symbols are about. As such, Shannon Information Theory itself assumes the encodability of physical signals, and measures this encodability. To give an example, suppose we ask how much information the rings of trees carry. Quantifying this according to Shannon Information Theory requires a scheme for encoding the rings as if they constituted a message. Supposing we produce an appropriate coding, we still might ask: what is this message about, what does it tell us? Are the rings of a tree information about the age of the tree? Are they information about the climate in which the tree grew?

In short, we are asking whether the quantity of the semantic content of a particular message is bounded in any way, such as by, say, the possible representations sentient beings can draw from it. We can surely agree that the rings of a tree do not carry information about the cereal that Al Gore ate this morning, so there are undoubtedly limitations to the quantity of semantic content a message carries. But we must be careful not to arbitrarily draw these boundaries based on what seems absurd. One of the favorite examples of popularizers of chaos theory is the butterfly effect concerning weather systems. Because weather systems are likely to be chaotic dynamical systems, small perturbations in the systems can cause them to go from stable to unstable states—or as expressed in Lorenz's famous paper (1993), the flap of a butterfly's wing in Brazil might stir up a tornado in Texas. Such a major event might be reflected in the rings of Texan trees, or at least contribute to a climate that is reflected in them. Do then the rings of a tree carry information about every minor variable that contributed to their specific formation?

Even if there are limitations that can be specified in a principled manner, if we identify 'information about' with a physical property, we are conceptually multiplying the physical properties of an object without there being any physical differences corresponding to these different properties. It would seem that the property of carrying X amount of semantic information might not differ physically from the property of carrying Y amount of semantic information. If two tree rings have the same pattern, yet have a different number of causal sources, then the quantity of information carried differs without a physical difference between the two rings. Similarly, there may be no physical difference between the two rings, yet each has an entirely different causal history. Such a case is improbable, but not impossible. Thus, proponents of information-processing models are faced with the sticky

question of why we ought to impute a causal difference between a signal that carries X amount of information about W and a signal that carries Y amount about Z when there is no physical difference between the two. What is more important, there seems to be no reason to impute a causal difference between a signal carrying information and one that does not. If this last point is correct, information-processing models are unnecessary; they can be superseded by direct explanations in terms of physical structure.

In his discussion of the computational theory of mind, Kim Sterelny (1990) adopted a three-level stratification of explanations of the mind/brain, drawn from David Marr's (1982) similar differentiation of the levels of information processing devices into the computational, the algorithmic, and the hardware levels. Sterelny holds that information is a phenomenon found only at that highest or ecological level. This is the level of intentional systems, or persons. It is the level that folk psychology describes. I will temporarily adopt the stance that what folk psychology describes of persons is roughly accurate. The question is then how to explain the emergence of such behavior, and whether information processing models apply to levels below that of the person, say to neurons or neuronal maps.

Attacking information-processing models of mind is hardly a new endeavor. The philosopher John Searle (1992) has made it the recent focus of his criticisms of cognitive science and artificial intelligence. In answering the questions whether the brain is a digital computer and whether the mind is a computer program, Searle has argued that there simply isn't a sense in which the brain processes information and therefore these two questions border on incoherence. Crucial to being an information processor, in Searle's estimation, is having syntactic operations, and the brain does not have these because

> syntax is essentially an observer-relative notion. The multiple realizability of computationally equivalent processes in different physical media is not just a sign that the processes are abstract, but that they are not intrinsic to the system at all. They depend on an interpretation from outside. (Searle 1992, 209)

Similarly,

> “functional organization” and “information”... have no causal explanatory power. To the extent that you make the function and information specific, they exist only relative to observers and interpreters. (Searle 1997, 176)

Information and syntax are observer relative properties, and thus not inherent in the physics of the brain. They are found in computers because computer designers put them there. The conclusion we are to draw is that what isn't in the brain's physics isn't in the brain. According to Searle, the notion that syntax in the brain is the basis for consciousness is nonsense, because syntax does not have any causal powers. Nonetheless, the brain is an intentional system, and this inherent intentionality is just a brute fact of brain biology. Notice that this argument subtly changes the domain in question from physics to biology. Intentionality is presumably not inherent in the brain's physics, intentionality not normally being considered a proper subject of study for physics (even though physicists are concerned with the effects and nature of observation). It is unclear why this does not disqualify the description of the mind/brain as an intentional system in the same way that it disqualifies the information-processing account. Before diving fully into Searle's critique, however, I ought to first define what is meant by information.

1.2 Shannon Information Theory

Just what constitutes information is the source of a great amount of confusion, and not merely among philosophers. For example, Edelman is notorious for attacking the idea that the brain processes information or that it invokes any algorithms, although it will be shown

that this critique of Edelman rests on a misunderstanding of what he and his colleagues mean by the terms ‘information’ and ‘algorithm’. Most often, authors merely assume that there exists a shared and well-defined notion of information such that any effort on their part to explain what they mean by information is superfluous. It is my contention that this assumption is false, that cognitive science lacks a clear idea of what information is and how it arises. This does not mean there are no theories of information. I will endeavor to show that these theories do not provide cognitive science with a foundation for its information-processing models.

Claude Shannon inaugurated what is now referred to as Shannon information theory (or just information theory) in his work “The Mathematical Theory of Communication” (1949). Shannon’s mathematical theory concerns only the syntactic aspect of information, the measure of which is termed entropy. It is a concept similar to that found in thermodynamics, but not to be strictly identified with it. Entropy in communication theory is sometimes defined as a lack of knowledge. Thus, if a message contains a great deal of novelty, i.e., what the message carries is not already known, then it has high entropy. This means that the quantity of information is observer-relative in the sense that it depends on what the recipient of a message already ‘knows’ or expects in the message. There is no absolute measure of Shannon information. Entropy in Shannon Information Theory is given by the function

Equation 1-1. Entropy

$$H = -K\sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant representing the unit of measure, p_i is the probability of event i, and n is the number of possible events (Shannon 1949, 50). This function is at its maxi-

mum when the values for p_i are equal, and zero when any p_i is 1. The more unlikely the message (the lower its probability), the greater the informational content is. Just what constitutes a 'message' is not strictly defined, and thus the notion of information is still a bit vague. Note also, this definition makes no reference to the meaning of the message. In fact, meaningless messages can have high informational content, or as Warren Weaver has expressed it: "two messages, one heavily loaded with meaning and the other pure nonsense, can be equivalent as regards information" (Weaver 1949, 12). Thus, Shannon information theory is concerned only with the quantity of information in a message as determined by its probability, not with the everyday sense of information as data about the world, and so the term 'information content' should not be confused with the semantic content of a word or sentence.

As the molecular Darwinist Bernd-Olaf Küppers (1990) has pointed out, the Shannon definition of information seems paradoxical, since information is generally regarded as knowledge while information theory defines it as entropy or lack of knowledge. The paradox dissolves when you view Shannon information as potential information. What it measures is the amount that could be learned from a message, without making any judgment about the meaning of the message. This seems to suggest that Shannon information is information only in virtue of its knowability by an observer. While the formal definition of information given above does not include the recipients of a message in the equation, the notion of 'message' is heavily laden with reference to knowing subjects. Shannon information at least requires the possibility of potential knowers, or, more accurately, potential decoders, since the message need not carry any semantic information in order to be known.

Shannon Information Theory offers the cognitive scientist little in the way of explaining the origin of information as cognitive scientists understand it, namely in the commonsense way of representations of the world. While this is often acknowledged, it is nonetheless thought to provide a rough measure of the possible ways to cognize the environment from which the information stems (Goonatilake 1991, 12; Holzmüller 1984). Nor has this realization deterred the likes of Fred Dretske or Kenneth Sayre from trying to derive a more suitable definition of information from Shannon Information Theory, efforts that will be examined later. First, however, I turn to the work of the molecular Darwinists, Manfred Eigen and his colleagues, to determine whether their expositions of the semantic and pragmatic levels of information offer an adequate account of its origin.

1.3 Biological Information

It has become standard parlance to speak of DNA and genes as if they were books or blueprints containing information about how to construct a phenotype. These books and blueprints also have readers, the RNA strands that translate the code into proteins, which in turn form the organism, enable it to move, and catalyze biochemical reactions among other functions (Brown 1989). According to this understanding of genes, information about the organism is stored in the genes's amino acids. So how does this information arise, and in what sense is it semantic information? Molecular Darwinists (Küppers 1990; Eigen 1992) have sought to answer these questions by applying the principle of natural selection to the molecular foundations of genetics.

Molecular Darwinism is the theory that “biological information has arisen by the selective self-organization and evolution of biological macromolecules” (Küppers 1990, 81).

Selection determines the arrangements of the 'symbols' in the genetic code, and, therefore, the information that is stored in the genes (Eigen 1992, 13). But what makes these nucleotide symbols forms of semantic information?

Following Weizsäcker (1971), Küppers holds that semantic information is information that is 'understood'. The relation of 'understanding' is a relation between microstates and macrostates. To explain this relation, he uses the unfortunate analogy of syntactic analyses of sentence structures. The microstates of a word are the letters, the macrostate is the word itself. Similarly for a sentence, the microstates are the words and the macrostate is the sentence itself. Although these relations are syntactic, what makes them semantic, to Küppers mind, is that they have some "mutual understanding" (Küppers 1990, 49). The mutual understanding is that the structural information of genetic microstates, nucleotides, is "contained completely in the structural information belonging to the concept 'nucleotide chain'" (Küppers 1990, 45). But structural information is merely a Shannon information-theoretic measure of the possible variability in a structure. The molecular Darwinist account of semantic information fails to adequately distinguish between syntactic and semantic information. There is, however, an additional 'pragmatic' aspect of information, which the molecular Darwinists consider to be necessary for semantic information to arise.

Molecular Darwinists understand the pragmatic aspect of genetic information to be as follows:

> The pragmatic aspect of information reveals itself wherever a message or an event, in the widest sense, alters the recipient. By "alters" we mean here both any structural change in the recipient and any willingness induced in the recipient to carry out a goal-directed action. (Küppers 1990, 88)

The pragmatic aspect of neuronal information could be its capacity to alter the structure of the population or map of neurons in which it participates. Küppers contends that the pragmatic aspect of information makes the semantic aspect "effective" (1990, 48). What emerges, vaguely, is a causal account of semantic information, where the causal relations between microstates and macrostates must be selected for.

There are a number of departures from the common sense notion of information in the molecular Darwinist approach. At the semantic level, biological structure is information only by virtue of its being normally expressed—the semantics of DNA is causal connection between its structure and the structure of the phenotype. Both the statistical sense of normal and Millikan's sense of 'having been selected for this particular function' are applicable here. DNA that is not normally correlated with or does not have the normal function of producing a phenotypic structure, such as junk DNA, does not have semantic content, even if there is a possible abstract mapping between it and a possible phenotypic structure. Contrast this with the intentional-systems-level/folk-psychological understanding of information. Information has an abstract rather than a necessary physico/causal relation to the structure it is 'about'. From the molecular Darwinist perspective, God could have seeded messages about how to construct alternative phenotypes into junk DNA, yet it would not have semantic content. Yet these messages would have semantic content from a folk psychological perspective.

Folk psychology also admits of information that has no effect on its recipient. Redundant information is one case of this. You can tell a person how to drive to Arizona when he already knows how and this is still passing information in the folk psychological sense, although it is not novel or useful to the recipient. However, from the molecular Darwinist

perspective, a biological message or event is information only if it normally causes a change in the structure of the recipient. It might be objected that the directions to Arizona normally cause a change in the recipient's informational structure, if not his brain processes, or normally enable him to find his way to Arizona. But here 'normally' must be understood in its teleonomic—the function of directions to Arizona is to enable people to reach Arizona—and not its statistical sense, as the latter would imply that only novel information is truly information. But the former sense would imply that only useful information is information. That would rule out information such as trivial facts (assuming we do not take the vacuous position that the function of such facts is to inform). In each case, the folk psychological notion of information as facts about the world must be radically altered for it to match the molecular Darwinist understanding.

So which understanding of information should prevail? I will argue that neither position explains the origin and nature of meaning and intentionality in the mind/brain. The causal-evolutionary account of the molecular Darwinists merely describes the interpretations we make of genes and other biological structures, interpretations based on causal connections that we discover. A nucleotide or chain of nucleotides is not a representation of a phenotype, but rather a mechanism for producing it. The question remains whether the firing of a neuron is a representation. In contrast to the molecular Darwinist theory of semantic information, which identifies semantics at the level of molecules and genes, Fred Dretske has sought to develop a causal-evolutionary theory of semantic information at the level of neural structures (Dretske 1981).

1.4 Why Dretske's Theory of Information Fails to Provide a Foundation for Cognitive Science

Although lacking a formal definition of information, and, hence, a definition of what it means to be an information processor, computational cognitive scientists nonetheless attribute several properties to information. Information must be in some sense extractable from the environment (or internal states). It must provide a reasonably reliable representation of the environment (or internal states). And it must be storable and manipulable in a manner that ensures that it is both retrievable and non-degenerate in the sense that noise introduced into the information is not so severe that the mapping is lost or is unusable. Yet, these properties are seemingly inconsistent. The first suggests that information is the input to a cognitive agent. The second implies that information is the mapping of input to the external world. The third hints that information consists of the physical symbols that are stored, retrieved, and manipulated.

In *Knowledge and the Flow of Information* (1981), Fred Dretske attempted to develop a theory of information derived from Shannon's that is both amenable to cognitive science's needs and resolves the apparent paradoxes in these needs. Dretske's theory distinguishes information proper from both Shannon information and meaning:

> signals may have a meaning but they carry information. What information a signal carries is what it is capable of "telling" us, telling us truly, about another state of affairs. Roughly speaking, information is that commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it. (Dretske 1981, 44)

He nonetheless formalizes his notion of information in terms of the Shannon functions for novelty of a signal and the average entropy, but only as a comparative function to compute whether one message contains more or less information than another. Dretske-information

is not measured in terms of the probability of a message, but in terms of its lack-of-knowledge reduction or possibility-elimination. For example, the more detailed the directions are that a gas station attendant gives you about where in the city of Chicago you can find a hotel, the more possible places for its location his instructions have reduced from your original knowledge that there is a hotel in Chicago, and thus the more information he has provided.

Dretske stipulates that the recipient of the message need not know how many possibilities were reduced, that the information content does not depend on this knowledge. Nonetheless, the number of possibilities reduced does depend on your prior knowledge about where a hotel might be found. You need not know that there are, say, 1 million possible locations for a hotel in Chicago for the reduction of possibilities to be 999,999 locations. This is a significantly smaller reduction than if you had only known that there is a hotel in Illinois. Dretske's example of comparing information content in messages does not address this aspect of the subject-relativity of information. Dretske compares two messages, one telling you that Denny lives in Madison, WI, and another telling you that he lives on Adams Street in that city. Since the number of places where Denny could live in Madison is less in the latter message than the former, the equivocation of the message is less and it carries more information. Dretske appropriates the term equivocation from Shannon information theory, using it to mean, in this instance, how many places could have been meant given that a particular place was mentioned. Saying only that Denny lives in Madison leaves open the possibility that the communicator meant Denny lives on a street other than Adams. In Shannon information theory, equivocation is "the uncertainty as to what symbols where transmitted when the received symbols are known" (Pierce 1980).

This sets the stage for Dretske's exposition of the semantic level of information. Dretske places three conditions on a signal for it to have the informational content in the semantic sense that s is F:

> (A) The signal carries as much information about s as would be generated by s's being F.
> (B) s is F.
> (C) The quantity of information the signal carries about s is (or includes) that quantity generated by s's being F (and not, say, by s's being G). (Dretske 1981, 63-64)

Condition (A) sets a lower bound on the amount of information a signal can carry if it is to carry the information that s is F. Thus, if s's being F 'generates' N bits of information, then a signal must have at least N bits of information to carry the information that s is F. Despite its similar formulation, condition (C) is meant to designate an important difference from condition (A). Condition (A) stipulates a lower bound on the informational quantity of a signal. Condition (C) requires not only that the right amount of information be present, but also that the right information is carried. If s's being G generates the same amount of information as s's being F, then a signal carrying the information that s is G satisfies condition (A) for s's being F as well as s's being G. It only satisfies condition (C) for s's being G, because only that information is being carried, and not the information that s is F.

Drestke offers the following definition of informational content to meet all three conditions:

> *Informational content*: A signal r carries the information that s is F = The conditional probability of s's being F, given r (and k), is 1 (but, given k alone, less than one). (Dretske 1981, 65)

The variable k quantifies what the recipient of a message already knows. In other words, the recipient might have knowledge that narrows the possibilities of what a message might indicate. If your partner in a card game tells you that his card is either the King of Hearts

or the Ace of Clubs, and you hold the Ace of Clubs in your hand, your prior knowledge makes the probability that he holds the King of Hearts to be 1.

What Dretske is trying to achieve with these definitions is to establish a relationship between information, truth, and causality such that a message *carries* information *about s* being F just in that case where it is true that s is F and that information is about s by virtue of its being caused by *s*. Dretske's paradigmatic example is the question whether the neurons in a frog's brain that fire when a bug is in view are indeed bug-detectors:

> The fact that a small moving bug on a light background causes a certain set of neurons to fire in the frog's brain and this, in turn, triggers a response on the part of the frog ("zapping" the bug with its tongue) does not mean that the neurons, or the frog, are receiving information to the effect that there is a bug nearby . . . It seems clear that certain neurons are labeled "bug detectors," not simply because a moving bug *causes* them to fire, but because, in the frog's natural habitat, *only* a moving bug (or relevantly equivalent stimulus) causes them to fire. (Dretske 1981, 35)

Dretske insists that the conditional probability of an event given the message must be 1 if the message is to bear information, because otherwise it would violate his Xerox principle, namely if A carries the information that B, and B carries the information that C, then A carries the information that C. If we were to allow lower probabilities, then a chain of messages would result in the final message not bearing information (falling below the accepted conditional probability, because these lower probabilities are all multiplied), something Dretske considers absurd.

So what is the upshot of Dretske's view of information? It is an account of the difference that information makes in cognition. According to Dretske, for information to be of any interest in studying the mind, it must have causal efficacy. A signal bearing information processed by the brain must have a different effect than a signal that bears no such information. It is important to remember that Dretske differentiates information-bearing signals

from non-information-bearing signals according to whether the condition they are about actually obtains in the world. Thus two signals could have the exact same physical properties, yet one bear information and the other not. So how could the fact that one bears information make a causal difference? Dretske offers discrimination learning in rats as one example of information making a difference. A rat hears a tone and is either rewarded or punished according to whether it, say, rings a bell. It originally has some internal structure S that is occasioned by the tone, and this structure takes on functions through the course of training that it originally did not have. These functions cause the rat to ring the bell when, and only when, the tone sounds. Dretske argues that the difference between S prior to training and after training is that S has become an information-bearing structure. To think otherwise, he contends, is to ascribe magical properties to the structure S. The change in the rat's behavior cannot be ascribed to changes in neural structure, rather, "[w]hat explains the perceptual state's newfound causal power is, in other words, its semantic, informational or intentional properties; not what it *is*, but what it is about" (Dretske 1990, 123).

Here Dretske is simply wrong in his contention that neurophysiological data cannot explain the change in the rat's behavior. The structure S has different causal powers because it has changed during learning. Long-term potentiation is a likely candidate for a neurophysiological explanation of the change. Dretske's error rests largely on an assumption of modularity about the neural processes involved. Presumably, if there were a modular perceptual state that always occurred when the tone was heard by the rat, and it was this perceptual state alone that resulted in the rat's changed behavior, then Dretske's explanation would be preferred to no explanation. But we are not stuck with an *a priori* assumption of modularity in explaining the rat's behavior from a neural perspective. There is a large body

of neurophysiological research documenting changes in neural structure correlated with learning. It is, in fact, Dretske's explanation that invokes magic when he contends that the semantic properties of a neural structure have causal powers. And as Brian Cantwell Smith has pointed out, a satisfactory account of learning in the rat can be given simply by reference to the causal relationships between S, the tone, and the prescribed behavior (Smith 1990).

A further problem with Dretske's explanation of the role of information in cognition is his insistence that a message only bears informational content if the condition it is about obtains with probability 1 whenever the message is sent. Given that Dretske has rejected the Shannon information theory formula of information as averaged over all possible messages, to say that a condition obtains with probability 1 does not mean an average over all possible instances, but simply that whenever P occurs, it only has informational content if the condition it is about obtains. Dretske is strictly identifying informational content with truth conditions. But suppose we are training a rat, and we only establish an association between message and conditions of 0.95. If the rat still learns the association even when a false message is presented, then Dretske's theory is empirically false.

Dretske's account of the nature of information fails to establish a link between the signaling found in dumb, low-level neural processes and the everyday sense of information as what is found in newspapers, scientific articles, and recipe books. Nonetheless, Dretske's work has highlighted distinctions between knowledge, belief, and what information might be, distinctions that are often papered over in artificial intelligence research. If a computer has proposition X in its database, the temptation is to say the computer believes proposition X and it knows proposition X, treating the two mental states as equivalent. Clearly, cogni-

tive science must address the important questions of what is information, how does it arise, and what is its relation to knowledge and belief if it is to explain human cognition. So what alternatives to Dretske's scheme are available?

1.5 A Misapplied Concept: Sayre's Use of Mutual Information to Explain Semantic Information

Like Dretske, Sayre attempts to provide an account of semantic information by extending concepts from Shannon information theory. In place of Dretske's causal explanation of semanticity, Sayre proposes that the reliability of a signal, understood as the mutual information between source and terminus of the signal, establishes its semantic content. Because reliable information channels provide a selective advantage, nature selects, according to Sayre, for greater reliability; this establishes for Sayre an evolutionary grounding for the appearance of semanticity.

The mutual information measure of an information channel is defined as the difference between the a priori entropy of a signal and the signal's equivocation. The a priori entropy of a signal is the measure of entropy at the source without knowledge of the signal arriving at the terminus. A signal's equivocation is a measure of the uncertainty remaining about what symbols where transmitted from the source given the output signal at the terminus. If the terminus is a perfect indicator of the source, then the equivocation is 0 and the mutual information is equal to the a priori entropy. Noise in the information channel is a possible cause for increased equivocation.

Sayre uses the human visual system to elucidate how mutual information underlies semanticity and intentionality:

> In visual perception . . . the relationship of intentionality is quite literally a relationship of high mutual information between a set of objective circumstances and a representation in the cortex of the perceiving organism. The representation picks out that particular set of circumstances, by virtue of its being the only object in the perceptual scene with which the representation shares that relationship. Through such a representation the organism's perception is directed upon a specific object, which is thereby the object to which the representation refers. By sharing identically in its particular structure, the representation is true of the corresponding object.
>
> But these are precisely the characteristics—intentionality, reference, truth, direction upon an object—that serve as paradigms of semantic features in CS [Cognitive Science] literature. (Sayre 1986, 136)

Increasing the reliability of a channel results in increasing the degree of identity between representation and object. Thus, Sayre subscribes to a form of direct realism, a point he is eager to make.

Sayre's account is vulnerable for this very direct realism. The visual system, by selecting aspects of the visual signal, reduces the amount of information available at the terminus. The most obvious of these reductions is from the 3-D world of the object to the 2-D world of the representation (Daugman 1986). As Daugman points out, one cannot truly compute a measure of fidelity when the dimensionality between what is being compared changes. Furthermore, Sayre's measure of reliability is unidirectional: There is a distinct source and distinct terminus which cannot be reversed. But information theory holds that the relation of mutual information is simply between two probability distributions and what is designated as 'input' and 'output' is irrelevant (Daugman 1986, 140).

Finally, even Sayre's reliance on evolutionary theory to somehow legitimize his thesis is misplaced, because the accuracy of a representational structure is only one factor in determining its selective value. Other factors include how fast a representation can be developed and how expensive (in terms of what sorts of physical structures are required) it is to develop it. While these issues will be addressed at length in a later chapter, it is clear

that mutual information does not provide an explanation for the origin or nature of semanticity. So the question remains, given the rather obvious and devastating flaws in Sayre's theory, whether any account of semantic information processing is available to the cognitive scientist.

1.6 Confusing Symbols, Representations, and Information

Efforts at explaining intelligence by computationalists have focused largely on the nature of symbols and how they are grounded, despite the common references to intelligent systems as *information-processing* systems. This is due to an identification of symbols embedded in cognitive systems with information. To process information is to manipulate grounded symbols, as Cummins (1983) has suggested. The problem of the nature of information then reduces to the problem of how symbols can be grounded, which is still a nontrivial problem. What attempts at solving the symbol-grounding problem most clearly demonstrate, however, is not how the mind/brain links its representations to the world, but that information cannot be identified with grounded symbols.

'The symbol-grounding problem' is Stevan Harnad's term for the collection of dilemmas that are implied by Searle's Chinese Room (1980) example and the Frame Problem (Harnad 1990). The primary difficulty is how meaningless symbols can come to have meaning, or how syntactic systems give rise to semantics. As Harnad (1990), Searle (1980, 1997) and others have pointed out, the interpretability of a system as having semantics does not imply that those semantics are intrinsic to the system. More recently, Searle has argued that the interpretability of a system as having syntax does not imply that the syntax is intrin-

sic to the system. In fact, no syntax is ever intrinsic to a system, since it is merely a categorization of the behavior of a system by an outside observer. Even if we are to assume that Searle is wrong about the possibility of syntax being intrinsic to a system, we are still faced with the overwhelming problem of how syntactic structures acquire meaning. Theories such as the Language of Thought that have previously dominated cognitive science assume, we shall see later, the existence of a built-in interpretation, sometimes likened to a compiler for generating machine code from high-level languages used in computers (Fodor 1975). But merely presupposing a built-in interpreter does little to explain how the interpretation of symbols is accomplished, as is now widely acknowledged.

Harnad's solution is to link symbols to the world via two forms of what Harnad considers to be non-symbolic representation: iconic and categorical representations. The former are "analogs of the proximal sensory projections of distal objects and events" (Harnad 1990, 335). While Harnad does not offer much in the way of neural correlates of either form of representation, a straightforward example of iconic representations might be the firing of the simple cells in the visual cortex in response to the presence of edges in the visual field. Harnad's example stimulus, however, is far more complex. He defends a prominent theory in cognitive psychology, namely that iconic representations of such complex objects as horses enable us to carry out discrimination between objects in the world without necessarily being able to identify their category. Categorical representations arise from reducing the properties of iconic representations to the invariant features, e.g., to those properties that single out horses as horses, allowing for identification of objects rather than mere discrimination. Harnad further contends that he and his colleagues ought not to be required to provide neural correlates just yet, since this would stifle cognitive theories by forcing them

to rely on quite preliminary research in the neurosciences. This will become an issue later when I examine whether cognitive scientists such as Harnad are proposing impossibility engines, after Christopher Cherniak's phrase (1991) for explanations of neural function that defy neuroanatomical constraints.

Simply naming categorical representations is not sufficient to elevate them to symbols, because symbols are considered to have systematicity and compositionality: They are assigned meaning only as part of a system and are capable of being composed into more complex symbolic structures, respectively. Harnad argues that compositionality is achieved merely by stringing categories together into "propositions about further category membership relations," and that this is sufficient to form systems of meaning. Therefore, the symbol-grounding problem dissipates through a bottom-up process of gradual abstraction.

Harnad acknowledges one major difficulty with his account, namely that the mechanisms for producing categorical representations are largely unknown. A far greater problem lurks—that the structures for storing and manipulating symbols in the brain are completely unknown, that no symbol has yet been identified in any biological neural network—but Harnad requests that this be overlooked for now. In addition to these two obvious obstacles, Harnad's account has further confused the relation between representation, symbol, and information. Information makes its appearance in a cognitive system not as symbol, but as iconic representation. The bottom-up process of symbol-grounding does not produce information; rather, it strips it away. Thus categorical representations are developed by ignoring all information except that which pertains to invariant features. Moving to symbols compounds this process of sensory deprivation, for symbols need make reference only to the few invariant features relevant to the proposition in which they appear. Understanding the

proposition ‘The horse is lame’ does not require the full informational content that identifying an animal as a horse does, since in the former case we may ignore the features that distinguish a horse from a zebra. We are merely given that whatever we are talking about is indeed a horse. The invariant features that distinguish horses from other animals might be thought to be implicit in the proposition, but then what is explicit in iconic representations is implicit in categorical ones as well. But even this is unnecessary. One could simply know that a horse is a quadrupedal animal, and thus know what it means for a horse to be lame, without being able to identify a horse. Symbols are thus two steps removed from the information the senses provide. That might be seen as a recasting of information, or a restricting of information, or information about information. We are left to guess.

What is missing from Harnad’s and the symbolic account is not so much how categorical representations are formed, but what is the relation between information in the physical signal and the iconic representation which is an ‘analog’ of it. Harnad contends that his treatment of iconic representations is free of homunculi, but that leaves us guessing why he calls them representations. If they are simply the transduction of one signal into another, then every such event in nature is a representation. The energy transferred to a rock as heat when a sledge hammer strikes it might then be thought of as a representation. Harnad’s examples suggest that he means some sort of mental imagery. An icon is not merely the transduction of an external signal by the senses—it is an ‘image’ derived from this transduction. But images have ‘aboutness,’ they are ‘of’ objects in the world, and are not merely caused by objects in the world.

Cognitive scientists typically view the ‘aboutness’ of representations as a function of their mapping to features in the world. Hence, Harnad uses the term ‘analog’ to describe

icons. Each icon has its own mapping, which is independent of the mapping of other icons; if it is not, then Harnad has not established the base level he wishes for his bottom-up approach. Remember, systematicity is to arise out of compositionality, and the latter is made possible by the simple nature of icons. What is lacking is an account of how this mapping inheres in the representation. To solve this problem, the notion of information has been used as a magic quantity, inhering in both external signals and representations alike. The passing of information from the external signal to the representation establishes the mapping. But cognitive science has not given us an account of what this magic property is, nor how it is passed. Assuming the mapping to be inherent in the representation is to drive the homunculus back into the machine. I will argue in later chapters that the only way to drive it out is to abandon the notion of representation as abstract mappings and instead accept systematicity at even the lowest levels. The challenge then is to develop an account of how these low level 'representations' arise without recourse to compositionality. But first I turn to the difficulties facing the spartan account of information that cognitive science has relied upon.

1.7 Problems with the Information-Processing Model of Mind

Regardless of the problems with cognitive science's methodology, denying that the mind/brain is an information processor will undoubtedly strike the reader as preposterous. Who would deny that humans gather, communicate, and discover information about their world and themselves? We create external representations of the world, such as drawings and maps, so surely we have internal representations that we are merely committing to paper. To reject this is to endorse behaviorism, an almost universally undesirable alternative.

John Searle has argued that there is no information inherent in the physics of the brain (Searle 1992). The level of phenomena investigated by physics is but one level that we could look at for information processing. Searle does not deny that humans do indeed process information, but how this would come about is left as a mystery, since Searle contends it cannot be found in the brute facts of the brain's biology. We are left with a series of choices as to where information might arise:

1. The sub-neuronal level: Might information arise at the level of mitochondria or other constituents of the neuron?
2. The neuronal level: Is the action potential of a neuron a transformation of information? Is it computation?
3. The neural population/map level: Are neural maps representations of the external world?
4. The gross structure level: Does representation arise within, for example, the hippocampus as a whole?
5. The whole brain level: Is the activity of a portion of the brain a representation only within the context of the whole brain?
6. The embodied brain level: Must the context be extended to the brain+body for there to be representation?
7. The environmental level: Must the embodied brain be considered in its environmental context for there to be representation?

Alternative characterizations of the possible levels of the mind/brain have, of course, been proposed. One popular characterization is that proposed by Simon and Newell (New-

ell 1982) in which they distinguish between the physical architecture level, the program level, and the knowledge level. Simon and Newell's hierarchy, however, is inadequate for asking the question of what level information arises at, because it presupposes which level it is.

The level of physical architecture is simply the structured guts of the machine, whether this is the connectivity of the brain's neural nets or the computer's internal configuration. Information does not appear at this level, because there is nothing more to it than how the physical parts fit together. What the physical architecture does is run a program, and it is at the program level that information and computation occur. Though the physical parts are carrying out this computation, the symbols that correspond to information are properly at the level of the program. The knowledge level is that which is generally described by folk psychology-full-blown intentional systems reasoning about their environment. So Simon and Newell place the origin of information at the level of the program, though it occurs also at the knowledge level. For this to be true about humans, one has to assume that there is a corresponding program level in the human mind/brain. It means that the mind/brain is a computer, following computational algorithms. I will soon investigate reasons to believe that the mind is not software on the brain. Because the Simon and Newell hierarchy does not provide a non-question-begging framework to ask at what level information arises, and the hierarchy I have proposed does not make similar assumptions about the existence of mind/brain programs, this seven-level stratification will serve in the meanwhile for investigating the origin of information.

Roger Penrose (1989) has made a case for assigning the origins of consciousness to the subneuronal level, arguing that microtubules, in displaying quantum-level effects, could

hold the key to the mystery of consciousness. The response to one who would put the origin of information at the subneuronal level is the same as that to Penrose. Microtubules, and similar constituents of cells, are common to cells throughout the body. Why then don't these cells and cell structures (like the foot) demonstrate consciousness? Similarly, why wouldn't they be considered information-processing systems?

The neuronal level is a more likely candidate for the origin of information, though this possibility has fallen more and more out of favor (Churchland and Sejnowski 1992). While connectionist models view the neuron as a computational element, taking inputs and transforming them according to (in most cases) a nonlinear squashing function into output, information as representation appears only in distributed form. A representation is a combination of activity of neurons across the net, not the input-output mapping of a single neuron. Still, research into the visual cortex has revealed single cells, known as simple cells, that respond to the presentation of an edge of a certain angle at a certain position (e.g., the center of the simple cells' receptive field), as well as complex cells which respond to an edge of a particular angle regardless of its position (Carlson 1994, 158). Also, many complex cells increase their rate of firing when the edge moves perpendicular to its angle of orientation, and so are thought to act as motion detectors (Ibid). It would seem plausible to hold that simple and complex cells represent edges of specific orientations. Proponents of the view that representations are compositional, that is, complex representations are composed out of simpler representations, certainly find this research in line with their predictions. Given connectionism's well-known neural implausibility, research of this type ought to be heeded more readily than the connectionist's contentions about the distributed nature of representation.

Yet, connectionism's insistence on a holistic approach to neural systems, rather than the traditional engineering approach of decomposing systems to understand constituent components (in this case, representations), sidesteps the initial objection that can be raised to neuronal level accounts of representation. If the neuron's activity is in fact a case of representation, who or what is 'viewing' the representation? The neuron itself is an unlikely homunculus, too dumb a component to understand what it is representing. Representations at the level of the neuron can only be representations in virtue of being part of a system that interprets them, that understands the relation between the spiking patterns and the world. For connectionist systems, representations emerge at the level of the network. It is not a question of what interprets the activity of a neuron, because the meaning of the neuron's activity is provided by the system's meaningful behavior, assuming it has such behavior.

Unfortunately, connectionism's solution to the neuronal-level homunculus problem is only a temporary dodge. That connectionist systems require a human interpreter will be demonstrated in the following sections, so I will not dwell immediately on the details of why connectionism fails. For the present purpose of estimating the level at which representations appear, the ability to collapse artificial neural networks into the function that is being approximated shows that they differ little from explanations of single cell behavior. Whether a researcher interprets an artificial neural network as parsing English or recognizing written words does not change its fundamental nature. Artificial neural networks are needed not so much because they simulate how the brain works-they don't-but because researchers don't know exactly how the brain does what it does, and a universal function approximator allows simulation without explicit knowledge of what the mechanism at hand is. If the exact form of the function being computed by a cell is known-assuming that is

what a cell is doing-the artificial neural network is simply not needed. Artificial neural networks simulate only the apparent computational work of a real neural system. They approximate its behavior-and nothing more. That they need an interpreter is a direct implication of this fact, as they are not embedded in real neural systems. An artificial neural network that approximates a Gabor transform or Difference of Gaussians is only simulating a cell in the visual system in the mind of the researcher carrying out the simulation.

Connectionist systems are level-4 systems, although they are not truly neural maps or populations. Perhaps true neural populations, or maybe level-5 systems (gross neural structures), can do the work of the homunculus. This would seem to be the implication of Daniel Dennett's Multiple Drafts theory (1991). Dennett argues that we should replace the well-worn image of the mind/brain as a Cartesian Theater-a place where representations are put on display for some supposed viewer-with the idea of the mind as either a collection of competing and cooperating demons or a series of multiple drafts of representations. (The correlation between this theory and Edelman's is superficially striking. More will be said on this later.) Dennett's theory resides at the level of cognitive psychology, with little being said about how the brain might realize such a schema. He has strayed very little from his original argument about how to get rid of the homunculus problem, which is that as one analyzes the mind/brain at finer and finer levels of granularity, the systems become dumber and dumber, until finally the homunculus is gone. But the problem of the nature of information does not disappear by pushing it farther down. In fact, it becomes more and more insoluble.

It is tempting to view the release of neurotransmitters as a form of information transfer, and to consider this the most basic level at which information is passed. But what is it about

neurotransmitters that makes them vehicles for information? Is it that they are released in response to environmental stimuli? Not all neurotransmitter release occurs because of environmental stimuli. Perhaps neurotransmitters are bearers of information because they are part of a chain of information transfer that begins with either external or internal stimuli. Thus, either photons striking the retina or hormones indicating states of the body could be the initial source of information, passed on by neurons signaling to one another via neurotransmitter release. So the initial information comes from a non-cognitive source.

At this point, the cognitive science de facto position that information is akin to representation, or serves as the content of representation, or is grounded-symbols, utterly founders. Semantic information of this sort is simply not carried by physical signals such as photon emission. Otherwise we would have to concede that rocks not only carry semantic information, they are information processors because there are transitions between their information-bearing states (in other words, any change in a rock's current state would be information-processing). Unfortunately, some cognitive scientists (McCarthy 1979) see nothing wrong with saying that objects such as thermostats have representations.

Even adopting such an absurd position would not save cognitive science's account of information. The kind and amount of semantic information being carried by a signal would be indeterminate, and so could not serve an account of the role information plays in cognition. The rings of trees presumably contain information about the causal forces that resulted in their individual patterns, but, as I have pointed out, there are two obstacles to this view. The amount of information that would have to be considered present is nearly infinite. Yet the vast majority of this information would not, and could not, have any causal efficacy. Information can therefore be present in a signal yet have no effect (and this undermines

entirely Dretske's position). So in what sense is it there? Second, it will often be in principle impossible to determine the type of information present. This is due to what might be called semantic equivocation. This happens when two or more causal sources produce indistinguishable effects. One answer to this is to consider tree rings as carrying information about all their possible causes. But this is to break the connection between information and the way the world actually is, as well as to multiply the inefficacious information carried by a signal.

If there ever was a ghost-in-the-machine explanation of cognition, it is the information-processing paradigm. Harnad is correct to point out that cognitivism has allowed mentalism in the backdoor, but he is wrong to think that he can get rid of it and still be a cognitivist. Granted that the notion of iconic representation plays a useful role in cognitive psychology, yet it is a bit like doing population genetics without molecular genetics to back it up: explaining a surface phenomenon by reference to entities of which we have no idea if they are real.

The question is not how the homunculus disappears, as it is generally agreed that molecules are not intelligent. The question is how it appears. Dennett has replaced the homunculus known as the unity of consciousness with slightly lower-level homunculi that are not much dumber. In a very important sense, Dennett has not even gotten rid of the Central Meaner, the homunculus that does all of the understanding for the neural system and thus provides the appearance of a unity of consciousness. He has merely replaced it with a Central Arbitrator. Because output devices on humans (i.e., mouths, hands, etc.) are in limited supply, not all of the demons can access them at once. Just as with a computer, a system of priority and arbitration must be enforced. Humans impose their priorities on the computer

as its priority system. What in the human mind/brain carries out this function for the multiple competing drafts? It is not a random system, as our trains of thought do have coherence. The Central Arbitrator might be considered dumber than the Central Meaner by virtue of imputing the task of understanding to the individual demons. It might simply be a selection system that allows the 'strongest' demon to express itself. But this means that all Dennett has succeeded in doing is multiplying the Central Meaner. His effort is like that of the Sorcerer's Apprentice, furiously whacking at the unyielding broom, only to find that the shards have taken on a life of their own and behave in the same manner as the original.

1.8 Theories of Syntactic Information Do Not Provide a Basis for a Theory of Semantic Information

Cognitive scientists regularly speak of sentences, utterances, and signals 'carrying information,' assuming not only that the notion of 'information' is well-defined, but also that the idea of something 'carrying' information is well-understood. Loosely, what is meant by 'information' is some set of *facts* about the world (or at a more basic level, some set of *indicators* of the state of the world). Thus, what is meant by information in this context is any signal that bears content, in contrast to Shannon's notion of information as a measure of the novelty of signals regardless of their content. When pressed for a theory of semantic information, however, few cognitive scientists undertake to sketch even preliminary details, and those who do have generally offered some permutation of Shannon's theory of syntactic information. The efforts to recast Shannon's original theory into an account of the nature of semantic information generally identify the semanticity of a signal with one of three aspects: its causal connections, its reliability, or the structural relations between sender and receiver of the signal. In this chapter, I have examined one representative of each of these

approaches: Dretske, Sayre, and the molecular Darwinists respectively. Each representative fails to explain the origin and nature of semantic information, though each for a different reason.

Dretske's causal account places too stringent of a restriction on what signals qualify as carrying information about the world, requiring a perfect correlation between signal and what the signal is about. Even with this restriction, the carrying capacity of any given signal is still theoretically infinite, carrying information about everything in its causal chain. What makes the signal about a particular event and not about all of the events normally causally connected with it is not explained. At best, Dretske explains by virtue of what a signal 'carries' information, but not what that information is.

Sayre's use of the concept of Shannon mutual information as a measure of reliability to explain the semanticity of signals falters on the very notion of the reliability of a signal. One might define the reliability of a signal as its faithfulness in capturing relations between objects and events in the world, but then one would be begging the question in defining semanticity in terms of it. Reliability is also a notoriously difficult notion to tease out when applied to sensory organs that map from a 3D world to 1- or 2D signals.

The molecular Darwinist approach is simply to cast the syntactic aspects of signals into biological terms as relations between biological structures. What makes these relations semantic rather than syntactic is unclear. What is also unclear is why semantic information is restricted to biological structures.

The failure to provide a theory of semantic information is of no little consequence for cognitive science: it entails the inability to develop a theory of cognition. Dennett has argued that intelligence arises through the conjunction of dumb building blocks into

smarter and smarter components. Yet, if one cannot explain how the dumb building blocks can represent *anything*, then conjoining them will do little toward explaining more sophisticated representations. Without a theory semantic information, one cannot explain the activity of dumb neurons in terms of their transferring information-about-the-world between one another.

In the next chapter, I turn to how information processing plays a role in the development of computational theories of mind. What will be argued is that computationalism inherits the indeterminacy of the information-processing perspective, making it impossible to tell when a system is actually doing computation. I will also investigate dynamical systems challenges to computationalism, and show that the crux of the disagreement is whether a system is doing information processing at levels lower than that of the global explanation of a systems function, which corresponds in cognitive systems to folk psychological explanations. Computationalism will be shown to be at best a method of describing simple systems, implying nothing about the nature of the system.

Chapter 2 The Failure of Computationalism and the Computer Metaphor to Explain the Nature of Mind

2.1 Information, Representation, and the Computer Metaphor

If one were to summarize the disparate efforts of cognitive scientists to understand cognition and perception, it would be in the form of the question: How does the mind/brain represent the world? And even if the notion of information remains ambiguous, we cannot level a similar charge that cognitive scientists have not attempted to clarify what representations are. Cognitive science has inherited more than two thousand years of philosophical thought upon the subject of representation, and in its brief history has surpassed the prior millennia in offering detailed models of representational systems. Cognitive scientists even have an account of how transitions between representations occur, computationalism, and a unifying metaphor for the mind, that of the computer (with all the usual caveats that there are dissenters to this view). Yet there remains a gaping hole in the computational/representational perspective. This gap is the role of information in representations and, by extension, computations. It is only by virtue of carrying information that mental events are representations, and only by virtue of processing information that computations are plausible models of mental processes. Models in cognitive science generally require us to assume that the mental structures they posit are indeed representations. Such an assumption merely

covers up what is truly at issue, namely how information turns brain events to representations and computations.

2.1.1 The Nature of Representations

So what is meant when a cognitive scientist says X is a representation? A representation is generally regarded as an encoding of a domain such that there exists a correspondence between the code and the subject of coding that can be deciphered in a nonarbitrary manner. One straightforward way of developing a representation is to draw a picture of an object, keeping all relevant aspects of the picture in the same relation as the object's properties. What is relevant is determined by the goal of the representer, since all representation is a reduction in dimensionality of the feature space of an object. For example, drawings and photographs reduce 3-D objects to 2-D renderings.

Neuroscientists often point to the motor and sensory homunculi as examples of relatively straightforward representations—the activity patterns of the primary motor and sensory cortices in relation to what part of the body is stimulated appear (to the observer) to form body maps or homunculi. The motor and sensory homunculi illustrate not only the dimensionality reduction inherent in representations—the number of neurons in the motor and sensory cortices is far less than the number of objects they represent—but also the feature distortion that is common to representations. Although hands and lips constitute a smaller fraction of the human body than do legs and arms, the homunculi have more neurons associated with the former two than with the latter two. Representations, by having a focus on a reduced feature set, distort some of the relations between features of their

objects. In the case of the motor and sensory homunculi, the relation of surface area and volume between lips and legs is distorted.

Cognitive psychology is essentially the study of representations as defined above. The debate between proponents of prototype and exemplar theories of memory is a case in point, as what is at stake is not whether the mind constructs representations, but whether categories are representations as statistical means, as advocated in prototype theory (Rosch 1977), or collections of representations of token members of categories, as understood in exemplar theory (Nosofsky 1986).

This is not to say that some cognitive scientists would not object to this definition of representation, but that is because they take a narrower view of representation for the sake of their particular research within cognitive science. Thus, Fodor contends:

> The full-blown Representational Theory of Mind (hereinafter RTM...) purports to explain how there could be states that have the semantical and causal properties that propositional attitudes are commonsensically supposed to have. In effect, RTM proposes an account of what the propositional attitudes are. So, the further you are from Realism about propositional attitudes, the dimmer the view of RTM that you are likely to take. (Fodor 1993, 273)

Fodor identifies representations with propositional attitudes because he is a Realist about the latter. The role that propositional attitudes play, however, is subsumed under the definition of representation given above. In rejecting the Representational Theory of Mind, authors such as Stich are not rejecting the notion that the mind is a representation machine. Rather, they reject the notion that the representations the mind uses are in fact what folk psychology purports, namely propositional attitudes. Stich's Syntactic Theory of Mind (1983) contends that the mind manipulates symbols, which themselves represent items in the world, but these symbols do not have the semantic and causal properties of proposi-

tional attitudes. Allen Newell has proposed a general law to capture the commonalities between kinds of representations:

> This is the essence of representation-to be able to go from something to something else by a different path when the originals are not available. We can cast this as a general law:
> The representation law:
> decode[encode(T)(encode(X))] = T(X)
> where X is the original external situation and T is the external transformation. (Newell 1990, 59)

2.1.2 How Information Serves as the Notion of Representation for the Computer Metaphor

Regardless of the specific form of the representational theory, cognitive scientists have implicitly defined *representation-of-the-world* to be synonymous with *information-about-the-world*. What is essential to being a representation is not its syntactic form, but its semantic relation to the represented domain. The representation need not be simple as in the cited examples. It can involve complicated mathematical transformations of the represented domain, so long as the brain is capable of carrying out the necessary computations. But a series of constraints on the reliability, storability, and accessibility of a representation leads to more specific notions of what the mind/brain is. In particular, they lead to a computer model of the mind. No matter how sophisticated a calculator the brain might be, if information degrades beyond a certain point, correspondences to the initial domain will be lost. Since humans have memories reliable enough to get them through daily life, as well as having assisted them through their evolutionary past, memory must be stored, computationalists reason, in a manner similar to that of the extremely reliable computer, give or take a little noise. Specific memories have specific locations (or, if you are a Connectionist, distributed locations, but a collective location nonetheless). Memory recall must also be sim-

ilar to that of a computer, namely accessing the appropriate location or address. Cognitive scientists look upon recall as retrieval. Putting information in its appropriate place is also likened to a computer's input mechanisms and data pathways.

Even the evolution of the mind/brain is compared to the design and construction of a computer. Explaining the posited innate structures is done in terms of a computer's internal code and compilers: just as the computer scientist creates the correspondence between machine code and primitive actions in the computer, which then becomes part of the computer's inherent behavior, so too evolution builds in certain codes that run the behaviors of living organisms. Jerry Fodor (1975) has used this analogy to deflect criticism that his Language of Thought thesis implies an infinite regress of mental languages. Just as the computer does not suffer such a regress, because the programmer establishes the underlying correspondence between the inherent code and the behavior of the machine, so evolution performs that service in the brain. Evolution is a *deus ex machina* in the role of programmer of the machine.

The influence of the information-processing formula is greater than merely suggesting metaphors. If the mind operates on information gathered about the world, it cannot assume that that information is always correct. It must have some way of testing that information, of reversing the transformation from stimulus to stored data, thus transforming that data to probing behavior. These transformations can only be explained in terms of computations on data, for while folk psychology may be able to tell us why someone did something, it cannot tell us how that process occurred in the mind/brain. That someone has a belief tells little about how it influences the person's behavior. Computationalism, on the other hand, does provide the explanation of how states in the mind transition to overt behaviors, or

other mental states. According to Strong-AI computationalists, the mind/brain is not merely like a computer; it is a computer (though an analog computer, not a digital one). Computationalists of all stripes also accept Church's thesis as true.[1] These two assumptions together imply that anything the mind/brain can do, a Turing-machine-equivalent computer can achieve. Computers can therefore be programmed to be mind/brains. Strong AI, the claim that some computers not only mimic minds, but are in fact minds, is a direct descendant of the information-processing perspective (weak AI merely claims that minds can be "modeled"). Strong AI computationalism, as the scientific expression of the information-processing model of the mind, is often expressed as an algorithmic model of mind. This has lead to confusion about what it means for the mind/brain to use algorithms, a confusion exploited masterfully by one of its principal proponents, Daniel Dennett.

2.2 Algorithms and the Computer Metaphor

2.2.1 Dennett's Mistaken Concept of 'Algorithm'

Cognitive scientists do not view the mind as merely a collection of representations. These representations must in some way produce meaningful behaviors. The processes by which representations are acquired, transformed, stored, and acted upon are algorithms that have somehow been coded in the mind. According to Dennett, algorithms have the following properties:

> (1) substrate neutrality:... The power of the procedure is due to its logical structure, not the causal powers of the materials used in the instantiation, just so long as those causal powers permit the prescribed steps to be followed exactly.

1. Church's thesis is that the classes of functions computed by recursive functions, Turing machines, Post production systems, etc., are the same. Each of these systems can compute any of the computable functions, the largest class of functions. Therefore, if something is computable, it is Turing-Machine computable.

> (2) underlying mindlessness: Although the overall design of the procedure may be brilliant, or yield brilliant results, each constituent step, as well as the transition between steps, is utterly simple. How simple? Simple enough for a dutiful idiot to perform-or for a straightforward mechanical device to perform.
> (3) guaranteed results: Whatever it is that an algorithm does, it always does it, if it is executed without misstep. An algorithm is a foolproof recipe (Dennett 1995, 50-51).

So an algorithm is a formal process for producing some result, or at least that tends to produce a particular result, regardless of how "interesting" that result is (Dennett 1995, 57).

There are a number of problems with Dennett's description of algorithms. First, it is unclear what is meant by the phrase "[t]he power of the procedure is due to its logical structure, not the causal powers of the materials used in the instantiation." Is Dennett suggesting the causal power of an algorithm is due to its logical structure and not the causal powers of what carries out the algorithm, or is there some other notion of power? When someone refers to the power of an algorithm, they generally are referring to its usefulness or efficiency for solving a problem relative to other algorithms for that problem. Algorithms are tools. To argue that an algorithm has the power to do something by virtue of its logical structure is akin to arguing that a hammer has the power to hammer nails by virtue of its physical structure. All algorithms are by definition useful for something, so either Dennett is merely defining power in terms of property (3), or as a subject-relative notion in the sense of one algorithm more efficiently achieving a person's desired results than another. The ambiguity in Dennett's use of the term 'power' is just what disturbs Searle about information-processing theories when he notes that syntax does not have causal powers (Searle 1992).

Second, what does it mean for there to be a misstep in an algorithmic process in nature? For example, consider what Dennett (1995) and Dawkins (1976) call "copying errors" in the replication of DNA. When DNA replicates, there is a high probability that the result is

identical to the original. Is replication an algorithm for producing identical copies, with nonidentical copies counting as missteps? Or is it an algorithm for introducing some variation into a population of genes? Or is it both, in which case, what counts as a misstep? What is the fact of the matter here? Evolution by natural selection requires some, but not too much, variation to act upon. Taken in the wider context of the process of natural selection, genetic changes introduced in replication are not necessarily errors. Yet Dennett constantly refers to them as "copying errors." Because he does not resolve this ambiguity, his definition of algorithm is not adequate for determining whether a natural process is an algorithm. Dennett would argue that because replication has been selected for, there is a sense in which it has a purpose, and therefore we can identify when it backfires. But Darwinian just-so stories that replication is introducing not error but needed variations can also be invoked. A process that is optimal when 70% of one outcome is produced and 30% of a different outcome is produced is not making a misstep when an outcome of 30% likelihood is produced.

Finally, just what in nature doesn't qualify as an algorithm? According to Dennett, processes need not have a guaranteed result to be counted as algorithms; rather, a guarantee of a tendency (the same thing as a tendency?) toward a particular result is all that is required to fulfill property (3). What apparently doesn't qualify as an algorithm is a purely random process, such as Brownian movement. Or does it? Might it be an algorithm that guarantees the random motion of molecules? Presumably, the causes of Brownian movement are not "for" the production of random motion in the sense that they were not selected for this effect. Yet, Dennett does not require that a process be selected for in order for it to be an algorithm. Further, a random number generator is an algorithm for producing random num-

bers (actually pseudo-random numbers). Why isn't the logical structure of the causal processes involved in Brownian movement an algorithm? Dennett sees no difficulty in the idea of an algorithm making use of randomness:

> Because most mathematical discussions of algorithms focus on their guaranteed or mathematically provable powers, people sometimes make the elementary mistake of thinking that a process that makes use of chance or randomness is not an algorithm. But even long division makes good use of randomness! (Dennett 1995, 52)

So it would seem that a process could make use of randomness as well as producing it, and still qualify as an algorithm. However, Dennett is wrong in his claim that long division makes good use of randomness. Because an algorithm is neutral toward the choice of inputs at a particular stage does not mean that it is exploiting randomness. A set rule for which number to choose when attempting to divide would work just as well for long division—it is neither required nor proscribed. One can choose randomly, but this random choice is not essential to the algorithm. It does not effect the algorithm's output. As we will see, Edelman and others argue that the stochastic nature of neural patterns is essential to their functioning. It is not merely that neurons make good use of randomness, but that this randomness is essential to their output behavior.

2.2.2 The Concept of 'Algorithm' as Found in Computer Science

It is not objectionable if Dennett simply wants to carve out his own notion of algorithms, but it needs to be pointed out that his definition does not correspond to that found in algorithmic information theory and computation theory. This is important to note when considering critiques of computationalism, such as Edelman's, when they make claims concerning the applicability of the notion of algorithm to the mind/brain. An example of what is considered an algorithm in algorithmic information theory is as follows:

> A Turing machine can carry out the most complicated calculations with the numerical information supplied, as long as these calculations consist of finite series of simple steps, of which each one follows on from the previous one in a purely mechanical way-that is, with no intellectual comprehension and *no random decisions*. A program that consists of instructions of this kind is called an algorithm. (Küppers 1990, 93) (emphasis added)

So when Edelman criticizes functionalism for its insistence on explaining the mind/brain in terms of algorithmic processes, when in fact neural functions are stochastic in nature, it is undoubtedly this notion of algorithm to which Edelman is referring. This is not how Dennett reads Edelman, however:

> Someone who does not understand this [that algorithms can involve randomness] is Gerald Edelman, whose "neural Darwinism" simulations are both parallel and heavily stochastic (involving randomness), a fact he often cites, mistakenly, as evidence that his models are not algorithms, and that he himself is not engaged in "strong AI" (e.g., Edelman 1992). He is; his protestations to the contrary betray an elementary misunderstanding of computers . . . (Dennett 1995, 444)

Dennett has expanded the definition of algorithm to a degree where it means little to say that a process falls under its rubric. Digestion is multiply realizable, guarantees a tendency to a result, and is simple in its mechanisms-and so is an algorithm in Dennett's sense. But computers can't digest. Edelman is at no risk in accepting Dennett's definition of algorithm and his contention that the brain employs algorithms. This does not imply that computers can realize the same algorithms; they cannot realize digestion, even though digestion is an algorithmic process in Dennett's sense of 'algorithmic'. Edelman argues that the 'algorithms' the brain employs are not information-processing algorithms. If they were, then computers could employ the same algorithms.

Dennett and the theorists of algorithmic information mean something radically different from one another when they use the term algorithm. A more common notion of algo-

rithm in computer science can be found in a popular instruction text on the subject. Cormen, Leiserson and Rivest (1990) offer the following definition:

> an algorithm is any well-defined computational procedure that takes some value, or set of values, as **input** and produces some value, or set of values, as **output**. An algorithm is thus a sequence of computational steps that transform the input into the output.
>
> We can also view an algorithm as a tool for solving a well-specified **computational problem**. The statement of the problem specifies in general terms the desired input/ output relationship. The algorithm describes a specific computational procedure for achieving that input/output relationship. (Cormen, Leiserson, and Rivest 1992, 1)

Here, an algorithm is not merely a substrate neutral method for doing something; rather, it is a procedure specifically for computing output values from input values. This is **always** what it means for a Turing Machine to be implementing an algorithm. A distinction therefore must be made between a computer's algorithm and, say, a bucket brigade. Organizing and running a bucket brigade is an algorithm, in Dennett's sense, for putting out fires. It doesn't matter if the brigade consists of people or robotic arms, if the buckets are made of plastic or metal, or if the substance smothering the fire is water or sand. It is a simple procedure, normally carried out by intelligent agents but not requiring them, and is guaranteed to tend to extinguish fires (it won't always succeed, but smothering a fire is going to lessen it). But the bucket brigade is not computing values, and as such is not an algorithm in the computer science sense. It is the computer science sense, especially the algorithmic information theory sense, that Edelman (1992) and Searle (1992) are attacking when they say that the brain does not use algorithms.

We could carry our abstraction of the bucket brigade system one step further and introduce *simulated bucket-brigading*. This might be a method of overwriting the contents of computer memory in one address with the contents at another address via copying through intermediate addresses. It might be used to wipe out sections of memory containing a virus,

thereby eliminating a spreading threat to the computer. While inefficient, it is nonetheless an effective algorithm, and it retains the logical structure of the real bucket brigade. Bucket-brigading is therefore a Dennett-style algorithm. Simulated bucket-brigading, however, does not produce the same result as its real world inspiration. It does not put out fires.

This distinction between algorithms generally and information processing algorithms is essential to the thesis proposed here. Dennett might tighten his definition of substrate neutrality to mean absolute substrate neutrality, but this would exclude informational processes from the category of algorithm. What logical structure is absolutely substrate neutral? So the path Dennett wishes to lead us down, that of

1. brain processes are algorithms,
2. computers can instantiate algorithms,
3. therefore computers can instantiate the same algorithms as brain processes,

is a mere equivocation. What Edelman has done is put his finger on what bothers Searle about cognitive science, a point Searle has been unable to make without sounding like he is proposing magic: brain processes involve causal powers not (currently) available in computers. The algorithms (Dennett's notion) the brain employs are substrate neutral, but not absolutely substrate neutral. Searle thinks they can never be implemented in a computer, Edelman is working to do so, while Dennett has skirted the issue by equivocating on the term algorithm. Dennett's path only works if you assume that the brain's algorithms are information-processing, with information processing understood as symbol processing. Since Edelman and Searle dispute the information processing nature of the brain's algo-

rithms, Dennett cannot win the argument by merely pointing out that the brain invokes algorithms.

2.3 An Information-Processing Model of Mind: Computationalism and its Varieties

Dennett's mistake is a common one. It is largely due to cognitive scientists harboring the unquestioned assumption that the mind/brain is an information processor. This assumption takes the form in cognitive science of computationalism, albeit a specific type of computationalism. In this section, we will examine the varieties of computationalism, especially information-processing computationalism.

While it is important to point out that computationalism-the theory that the states of mind/brains are correctly understood as computable functions-is not strictly equivalent to the information processing perspective, the vast majority of computationalist accounts are also information-processing accounts. Folk psychology, as an information-processing model, is a notable exception to the otherwise complete coextension between computationalism and the information processing view. Mainstream computationalism, however, is the scientific expression of the information-processing perspective.

2.3.1 Mainstream Computationalism

So what is computation, and what does it mean to have a computationalist perspective of cognition? As was noted above, computer scientists define algorithms in terms of computation. These algorithms are therefore computational algorithms, a subspecies of algorithm (they are also information-processing algorithms in the specialized sense of information introduced above). This definition seems to get muddied if we then view as 'computable'

those functions that can be evaluated by means of an algorithm, as Steven Horst does (1996, 30). A function is simply a mapping of inputs to outputs. When we talk about computable functions, we are talking about functions that take symbols ('values') as inputs and produce symbols as outputs. These functions are computable if there exists a formal procedure for carrying out the mapping or transformation. Computational algorithms are those algorithms that produce symbol mappings, as opposed to those algorithms that produce, say, fire-smotherings.

More precisely, a computational algorithm is a procedure to produce the ordered pairs that define a particular function. It takes a symbolic input and produces the symbolic second element of the ordered pair defining the transformation from the input domain to the output range for that input. A computational algorithm always produces the same output for identical inputs. Thus an algorithm that produces outputs according to a probability density function for identical inputs is not a computational algorithm.

A computational theory of mind is one that holds that the mind/brain produces representations as symbols, and transforms these symbols into other symbols (representations) and behavior by means of computational algorithms. As Horst points out, the computational theory of mind has two components: the thesis that intentional states are representations and the thesis that cognitive processes are computations over these symbols/ representations (Horst 1996, 37).

2.3.2 Weakened Computationalism: Computationalism as the Computability of Cognition

William Rapaport (1998) has developed a variation of computationalism that allows for the possibility that human minds are not engaging in computations. He draws a distinction between the computability of a process and that process's being a computation:

> One standard kind of example is illustrated by the solar system, which, arguably, computes Kepler's laws. However, it could also be said that it is Kepler's laws that are computable and that describe the behavior of the solar system, yet the solar system doesn't compute them, i.e., the behavior of the solar system isn't a computation, even though its behavior is computable. (Rapaport 1998, 404)

So too could the mind's functions be computable but not actual computations. If this is true, Rapaport argues, then any machine that actually computed the functions carried out by the mind would nonetheless be a mind.

This approach would conflate descriptions of behavior with the behavior being modeled, except that Rapaport makes one important distinction: although computing Kepler's laws does not result in a solar system's actual behavior (no stars as physical entities are born in computer simulations, only representations of stars), simulating mental functions does indeed produce true mental functions. Simulated thought is real thought. So if the human mind is not doing computation, as Rapaport allows, yet its behavior is describable in terms of computations, why should we equate the simulation with the actual activity in this case when we cannot with the prior two? Rapaport's answer is that simulated minds do the same things as human minds, transform symbolic input into symbolic output (where a symbol is understood to be a meaningful marker). The human mind might not use computation to get from input to output, but, according to Rapaport, the intermediate between input and output

is not essential. A computer that transforms the same input to the same output as the human mind does is thinking the same thought.

The two issues here are whether the intermediate process from input to output are irrelevant and whether computations can indeed produce the same output from the same input as humans do. To tackle the first, Rapaport examines the objection from Fetzer (1998) that computers cannot dream because dreams are not computational. For the sake of the argument, Rapaport assumes that dreams consist of random neuron firings and the interpretations that we place on these firings, though he allows that the true characterization needs to be worked out by neuroscientists. He concludes that "insofar as our ordinary interpretations of neuron firings in non-dreamlike situations are computable, so are dreams" (Rapaport 1998, 407). But if we suppose that dreams are what Rapaport suggests they are, then we must come to an opposite conclusion. A computational implementation of dreams would not only produce the interpretation, but also the material for interpretation: the random neuron firing. Random patterns, however, cannot be produced computationally, as was argued earlier concerning Dennett's claim that algorithms can "use" randomness. Rapaport would have to argue that the pattern of neuron firing is just input to what really constitutes the dream, the interpretation, just as a random pattern of lights can be the input to a computer vision system, but this would contradict psychophysiological research that uses the random or chaotic nature of the neuron firings to explain the "bizarreness" of dreams (Hobson and Stickgold 1995). The random neuron firings also differ from random light patterns in that they are internal and supposedly representational. The input to whatever does the interpretation is not a set of meaningless signals, but rather a set of associations. A large

amount of the neuronal activity that takes place during REM sleep is in the visual association cortex (Madsen et al. 1991).

Perhaps dreams are indeed a bad example. A computer that could think but not dream would still qualify as a cognitive agent. But the case of dreams raises our two important questions: Can a computation always get from the same input to the same output as human thought, and is the process by which output is produced from input essential to what thought is? Rapaport answers the first question in the affirmative; the second he must answer in the negative. So what examples other than dreams are there of thought processes that computation cannot reproduce? As mentioned earlier, some neurobiologists hold that randomness is essential to neuronal behavior (Edelman 1987, 1992). The theory of stochastic resonance has expanded the possible roles for randomness in neurons—in particular, recovering weak signals by adding an optimal amount of noise (see Gammaitoni et al. 1998 for an overview and application to neuroscience). Computation can approximate these behaviors, and is therefore useful for simulation, but it cannot faithfully reproduce them.

More perplexing is the notion that what gets you from input to output is irrelevant. Rapaport cites the analogy between getting from point A to point B and producing an output from an input in a cognitive system to illustrate how the intermediate process is irrelevant. The distance between A and B is both ‘drivable’ and ‘walkable’, but both driving and walking get you to your destination. The question, however, is not whether computers can get the same answers as human minds, but whether what they are doing is thinking. The analogy does nothing to sort out this question (it also raises the question of the qualitative experience of thinking versus computing). Walking is not driving, and so, similarly, computing may not be thinking. Rapaport, however, distinguishes between ‘thinking’ and

'Thinking', the former being what brains do, the latter what mind/brains and computers do. 'Thinking' is an abstraction of the processes of 'thinking', and what this abstraction captures can also be implemented in computers. Why does Rapaport hold this view?

Rapaport holds this view because he believes that the output of computation and the output of human thought are not only similar syntactically, but also similar semantically. For example, both a computer and a human can add 5 and 7 to yield 12. The two outputs of 12 are not the same merely as numerals, but they also bear meaning for the system that produced them. And if the structure of the computational system is right, then the meaning for the computer can be the same as the meaning for the human, though it need not be the case for the output to mean *something* to the computer. But before I examine his arguments for these claims, I will take a short detour that reveals the intuition behind them.

2.3.2.1 Why the Chinese Room Beats the Korean Professor

With his Chinese Room thought-experiment, John Searle (1980) laid bare the intuitions of both sides of the debate over the possibility of strong AI.[1] As an analogy to AI systems (in particular, natural language processing systems), Searle envisioned a room in which he was locked and given only a codebook to transform slips of paper with Chinese phrases on them, which are passed under the door to him, into other Chinese phrases that he passes back under the door. The instructions for doing this are given to him in English, but he himself does not understand Chinese (at least prior to entering the room). Searle argues that neither he, Searle-in-the-room, nor the system, Searle-in-the-room-plus-codebook, understands Chinese. If neither he nor the system understands Chinese, how can AI sys-

1. The notion that computers can be minds, and not merely simulate their behavior.

tems, which are doing essentially the same thing when running programs for natural language understanding, be considered to understand the languages they converse in? Searle contends that the Chinese room is engaged purely in syntactic operations, and that syntax is insufficient for semantics. Hence, the Chinese room, and AI systems in general, do not qualify as truly thinking agents.

To counter this intuition, Rapaport proposes the Korean-Room argument (Rapaport 1995).[1] It runs as follows:

> Imagine a Korean professor of English literature at the University of Seoul who does not understand spoken or written English but who is, nevertheless, a world authority on Shakespeare. He has established and maintains his reputation as follows: He has only read Shakespeare in excellent Korean translations. Based on his readings and, of course, his intellectual acumen, he has written, in Korean, several articles on Shakespeare's plays. These articles have been translated for him into English and published in numerous, well-regarded, English-language, scholarly journals, where they have met with great success. (Rapaport 1995, 254)

The analogy between the Chinese room and this Korean professor is not that the room understands Chinese just as the professor understands English. It is stipulated that the professor does not understand English; what he understands, Rapaport contends, is Shakespeare, and not just a Korean translation of Shakespeare. Of course, one might dispute this. Part of understanding Shakespeare is understanding the language he used, and this is at least partially lost in translation. Nonetheless, it can be granted that he does know something of Shakespeare that is not dependent on translation. Similarly, Rapaport argues, the Chinese room knows *something*, and even though this something may not be Chinese, it is natural language. If the Chinese room is capable of understanding a natural language, why can't a computer? In fact, why can't a computer understand any natural language?

1. This argument was suggested to him by Albert Hanyong Yuhan

With this thought-experiment, Rapaport has exploited a weakness in the Chinese room that needn't be there. By putting a thinking agent in the Chinese room, Searle has made it too easy for his opponents to subtly beg the question against him whether the room understands anything (not necessarily Chinese). Searle-in-the-room understands English, and he understands the instructions for composing new phrases. The latter is the proper analog to the Korean professor's understanding of Shakespeare. This does not imply, however, that a computer also understands a natural language, because these abilities were imputed to Searle *as a thinking agent*. And it certainly does not enable the leap that Rapaport makes from the professor's knowledge of Shakespeare to Searle-in-the-room's understanding of Chinese. The professor understands the meaning of 'To be or not to be,' regardless of which language he read it in, but the question is whether Searle-in-the-room understands the squiggles that he receives or returns. It was stipulated that Searle did not understand Chinese before entering the room. When does he start understanding it? When he reads and understands the English instructions for composing new squiggles? This is not even a proper analog to knowing the grammar of Chinese, since the rules are a set of look-up tables that he uses by comparing shapes of squiggles, not parts of speech. Does he start understanding Chinese when he receives the first slip of paper, or after he has composed a new set of squiggles in response to it? There is no reason for us to conclude that he ever starts understanding Chinese. Searle-in-the-room understands what he can read, namely the rules for composition, just as the Korean professor understands what he can read, Shakespeare.

What about Searle-in-the-room+book? Searle anticipates this by allowing that Searle-in-the-room has completely internalized the translation book (1980). Searle concludes that Searle-in-the-room still does not know Chinese. Rapaport's counter-intuition is that he

does. How to settle this? With one further intuition: Searle-in-the-room+book cannot translate from English to Chinese at all. Since it was stipulated that Searle-in-the-room knows English, it must be because he does not know Chinese. It is not that Searle can't translate very well; it is that he cannot translate at all. Rapaport, however, seems to believe the contrary:

> Were Searle-in-the-room, *with his book*, to be stranded on a desert island and forced to communicate with a Friday who only spoke Chinese, he—with the help of this book—would be able to do it. (Rapaport 2001, 18)

An anonymous reviewer of Rapaport's draft article (2001) pointed out that Searle would not know what he was saying at any point, and, therefore, would not be communicating. He could not generate sentences such as 'I am thirsty' to express his desires. Rapaport's reply is that Searle could make replies just as well as he did inside the Chinese Room. But this is to miss the point of the objection. Searle needs to make more than *replies* in order to communicate with Friday; he needs to make *requests*. Imagine the following scenario: Searle is thirsty and wishes to ask Friday to get him a drink. He cannot make this request because he cannot translate it from English (or his language of thought) into Chinese. The reason he cannot translate it is because he does not know Chinese, even with his book on hand. Furthermore, even the claim that he can reply to Friday as well as if he were in the Chinese Room is false. The replies inside the Chinese Room were to a story, and the answers Searle-in-the-room gives are correct replies to questions about the story. Searle's actual desires and needs do not come into play when he is in the Chinese Room. When Friday asks Searle if he wants a drink, however, there are two valid answers, only one of which actually reflects Searle's desire for a drink. Unfortunately, Searle cannot decide which answer to give, because he does not know what the answers mean (in fact, he does not know what the

question means, or if it is a question, only that given the utterance from Friday, he can make two replies).

The reason Rapaport believes Searle-on-the-island could communicate with his Friday is that he has added additional conditions to the scenario Searle describes. The instruction book that Searle-in-the-room and Searle-on-the-island must use is one that maps Chinese phrases to Searle-in-the-room's internal symbols. In other words, it provides the translation that I noted Searle-in-the-room was incapable of doing with the book Searle described. But this is to beg the question as to whether computers are capable of understanding, because it is *stipulated* that Searle-in-the-room understands something, whether it be English or his own internal language. Providing a translation between Chinese and this internal language makes the question whether Searle-in-the-room understands Chinese moot. Of course he does (assuming, as Searle granted, that he has internalized the book). Rapaport has turned the Chinese Room argument into a question of whether a true cognitive agent can learn Chinese. But we wanted to know whether a computer could be a true cognitive agent.

No doubt that given enough time and appropriate feedback (such as someone slipping translations under the door), Searle-in-the-room could learn Chinese. This is simply because Searle himself is capable of learning natural languages. The question is, instead, whether Searle-in-the-room would understand Chinese if he learned it as a computer would. Giving Searle-in-the-room feedback in English, or feedback that he can translate into English, is to bootstrap off of an already given knowledge of natural languages. A computer that learns a natural language for the first time cannot rely on such bootstrapping, although, as with humans when they learn natural languages for the first time, it can receive feedback on its utterances and hear examples of correct usage. To replicate a computer, we

would have to give Searle-in-the-room another language, Searlese, one which is not mapped to English, but is fully specified in its relations between its elements, i.e., one with a complete grammar and a specification of how each element relates to another (a semantic network, which will be more thoroughly detailed in the next section). Searle-in-the-room fully understands this language in the sense that he knows what relations obtain between each element, and how to put elements together into sentences. He knows, for example, that 'squigs' are a class of 'squogs'. He then learns what Chinese phrases map to Searlese. He can formulate any sentence of Chinese in Searlese, and vice versa. And he still doesn't understand a word of Chinese, because he doesn't understand Searlese except in terms of itself. He can ask questions in Chinese and Searlese, but only by accident or by the prompting of someone else who is trying to elicit a specific reaction from Searle-in-the-room and knows how to trigger it.

This is an imperfect example, because Searlese is a bit too weak. It should also map to basic operations that Searle-in-the-room can perform, although these operations are just simple computations. Searle-in-the-room would be given a set of registers in the form of holding bins in which chips can be stacked, and he would be able to add and substrate from these holding bins, as well as multiple and divide. Although Searlese is a bit weak, the alternative is to allow that Searle-in-the-room's knowledge of English is analogous to computer machine language. If English is allowed as Searlese, then we have already granted that the computer knows natural language. Searle's goal was to give the strongest possible situation and yet show that computationalism fails. In doing so, he offered too much, although it is still questionable whether Searle-in-the-room (or Searle-in-the-robot) of the original thought-experiment could learn Chinese through its experience.

Although Rapaport's thought-experiment fails to demonstrate that the Chinese room understands Chinese (or any other natural language other than what Searle-in-the-room is stipulated as knowing), it indicates the direction computationalism must take to establish that computers can think. Computationalists must establish that pure syntax gives rise to semantics. This is exactly what Rapaport attempts to show in his theory of syntactic semantics.

2.3.2.2 Taking Computationalism to its Logical Conclusion: Syntactic Semantics

In "Understanding understanding: syntactic semantics and computational cognition" (1995), Rapaport distinguishes between three types of semantics: internal, referential, and external semantics, the latter being what Harnad called grounding symbols. Unlike Harnad, however, Rapaport considers external semantics only necessary for mutual understanding between cognitive systems. The reason is that he is what he calls a "representative realist" (2001, 19), which is someone who denies that we have direct access to the external world because the external world is always mediated by our representations of it, but who accepts that these internal representations can be caused by external objects. So to qualify as a cognitive system, or at least as one capable of natural language understanding, one must have internal semantics.

So what fixes the internal meaning of a marker? Its location in the cognitive system's semantic network (Rapaport 1995). And what is a semantic network? This is a little more complicated. Rapaport contends that syntax is sufficient for semantics, and bolsters this claim with a pair of blueprints for how semantics can be converted into syntax, and so realized by syntactic relations.

The first approach to converting semantics into syntax takes as its starting point a definition of semantics as the relation between markers and what those markers mean. Meanings are then treated as markers themselves, and the set of markers and their meanings are unioned. Now, the relation between markers and their meanings is one simply between markers, and is therefore syntactic. But how do meanings become markers? The meanings are themselves internal representations, since Rapaport denies the possibility of direct access to objects outside of consciousness. The activated nerves that underlie the representation also serve as markers, so the relations between them are syntactic. What convinces Rapaport that there is no direct access to the world is a set of experiments. Looking at a light while closing one eye causes a representation slightly different from looking at the same light with the other eye closed. These representations are also different from that resulting from looking at the light with both eyes open. The distinctness of these representations leads Rapaport to conclude that it is always and only through representations that we come in contact with the world, and that therefore we have no direct access.

The second path to syntactic semantics is the need to terminate the infinite regress of semantic interpretation. The semantic domain is used to understand the syntactic, but how is it that we understand the semantic? We could formulate a third domain to interpret the semantic, and a fourth to interpret the third, and so on. At some point, this regress must cease for there to be understanding, and this final domain must be understood in terms of itself. According to Rapaport, the only way to understand a domain in terms of itself is syntactically.

2.3.2.3 Why the Syntactic Move Does Not Work

At first glance, Rapaport appears to be advocating a form of reductionism similar to the reduction of mental states to physical states. Semantics is simply the appropriate kind of syntactic operations. Rapaport's move, however, is more radical than this. He is not merely reducing semantics to syntax, but rather eliminating the former:

> I understand the internal symbols of my own Mentalese language of thought syntactically. One could say that "mental terms" don't *mean*; they just *are* (shades of Gertrude Stein?). More precisely, they *interact*: I manipulate them according to certain (no doubt unconscious) rules. (Rapaport 2001, 22)

This is precisely the move that methodological solipsists must make to address the question of what meaning is in the context of a computational system. The answer, that there is only syntax, provides a perfect description of computational agents. Unfortunately for the methodological solipsist, it does not describe true cognitive agents. Much of the remainder of this dissertation will address how systems built on solipsistic assumptions differ from what is known about human mentation and how these differences matter to being a cognitive agent.

One difficulty for methodological solipsists is how to explain what makes an internal representation a *representation* once semantics is eliminated. As we saw with the Chinese room, in order for markers to have meaning, there must be some way of translating them into one's already existent system of meaning. Rapaport offers one scenario on how this might happen given methodologically solipsistic assumptions. The nerve activation that is the visual perception of an object is bound or associated with the nerve activation that represents the word for that object. It is with these vague terms—*bound*, *associated*, *represents*—that methodological solipsists get away with murder. In what sense does a set of nerve activity represent the word 'tree'? How are sets of nerve activity bound?

This is not to be obtuse or ask the impossible. Granted that neuroscientists have yet to adequately answer these questions themselves, they have, nonetheless, provided outlines of the answers, and these outlines point away from methodological solipsism, as will be shown in later chapters. Furthermore, Rapaport cannot complain that he is not required to explain human cognition since he is only interested in the computability of cognition. His claim that semantics can be converted to syntax is dependent on his theory of representation. The applicability of this theory to computers is further dependent on the validity of his claim that however humans bind incoming signals to markers, so too can computers. Finally, the ability of computers to learn is also dependent on this theory of how representation arises, and the ability to learn is one of the qualities that Rapaport believes a cognitive agent must possess.

2.3.2.4 Damasio's Theory of Time-Locked Multiregional Retroactivation as a Biological Model for Syntactic Semantics

Rapaport (1996) offers Damasio's theory (1989) of time-locked multiregional retroactivation as a plausible explanation for how the binding that (syntactically) produces semantic relations occurs. Citing Damasio's theory as an example of how syntactic semantics might be realized in humans is an important step toward answering the question of whether the input/output relations found in humans can be realized in computers: if the brain implements a semantic network along computational lines, then a computer should be able to do the same. So is Damasio's theory a brain-analog of the computational semantic network that Rapaport envisions to be the foundation for understanding? Damasio (1989) has muddied the waters as to where his theory stands by claiming it to be compatible with both Edelman and Finkel's Theory of Neuronal Group Selection (Edelman and Finkel 1984) and

Fodor and Pylyshyn's (Fodor and Pylyshyn 1988) characterization of classical cognitive architecture, the latter being somewhat similar to Rapaport's. Edelman has since clarified how his theory stands radically opposed to the kind of cognitive architecture defended by Fodor and Pylyshyn, so it is not clear that Damasio accurately assessed the compatibility of his theory with others.

Damasio's theory is a challenge to the conventional view of the time that there is a uni-directional flow of information in the brain from sensory and motor systems to higher association cortices which progessively extracts features and produces finer and finer representations (Damasio 1989, 30). Damasio's theory also challenges the idea that there is a single site in the brain where sensory and motor information is integrated to produce the unity of consciousness. Instead, Damasio proposes a massively parallel architecture, distributed over the various regions of the brain, which feeds fragments of sensory and motor information into zones where feedback to the sensory and motor cortices is initiated according to the combinatorics of the incoming information. Recollection is reactivation of the regions of the brain associated with the early stages of experience.

The architecture proposed by Damasio consists of four layers:

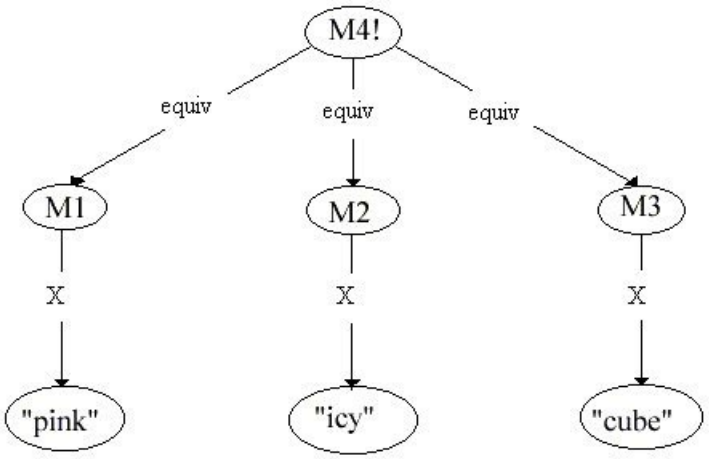
> 1.) neuron ensembles located in multiple and separate regions of primary and first-order sensory association cortices ("early cortices") and motor cortices; they contain representations of feature fragments inscribed as patterns of activity originally engaged by perceptuomotor interactions . . .
> 2.) neuron ensembles located downstream from the former throughout single modality cortices (local convergence zones); they inscribe amodal records of the combinatorial arrangement of feature fragments that occurred synchronously during the experience of entities or events in sector (1) . . .
> 3.) neuron ensembles located downstream from the former throughout higher-order association cortices (non-local convergence zones), which inscribe amodal records of the synchronous combinatorial arrangements of local convergence zones during the experience of entities and events in sector (1) . . .

> 4.) feed-forward and feedback projections interlocking reciprocally the neuron ensembles in (1) with those in (2) according to a many-to-one (feed-forward) and one-to-many (feedback) principle. (Damasio 1989, 25)

Experience is the time-locked activation of the fragmentary perceptuomotor records in (1), which feeds into (2), where records of the combinatorics of the perceptuomotor inputs are stored. Recollection is the time-locked activation of the records in (1) by feedback from the convergence zones. It is important to note that the convergence zones do not store representations, since convergence zones "serve as pivots for reciprocating feedback projections rather than as the recipients and accumulators of all the knowledge inscribed at earlier levels" (Damasio 1989, 38). Damasio introduces convergence zones to solve the binding problem, the question of how unified representations arise from fragmented sensory records. Damasio extends the binding problem to recollection as well as sensory experience, and seeks to cover both with his notion of convergence zones.

In what way does Rapaport see Damasio's theory as compatible with computationalism as he envisions the latter? Rapaport argues that Damasio's theory is an example of syntax producing semantics (or representations). He gives an example of a Damasio-style 'amodal' representation, shown in Figure 2-1, as it might be produced with the semantic network that is the focus of his work, SNePS, to illustrate this claim.

Figure 2-1. Rapaport's Damasio-like amodal SNePS representation of a pink ice cube. Node M4 is meant to represent the simultaneous experience of the fragments of the experience. (Rapaport 1996)



In a traditional SNePS-like network, M4 represents a pink ice cube by virtue of the connections it has to the features, i.e. its location in the network. In a Damasio-like SNePS network, on the other hand, it is not merely the connections but the co-temporality of activation that is essential for binding the features to produce a representation. The 'equiv' relation is meant to represent co-temporality in the SNePS network. The assumption is that a *representation* of co-temporality is sufficient to solve the binding problem. But the representation of co-temporality requires an additional mechanism, namely that which reads and interprets the 'equiv' relation. Co-temporality is not a symbol for binding, but rather a causal mechanism for bringing it about. Further, as noted earlier, and contrary to Rapaport's interpretation, convergence zones are not representations. Convergence zones are

"uninformed as to the content of the representation they assist in attempting to reconstruct" (Damasio 1989, 46) and "only the multiregional retroactivations of the fragment components become a content of consciousness" (1989, 28). The combinatorial codes stored at convergence zones are not themselves representations; they only facilitate representations. And representations are not amodal, because they are the result of the reactivation of perceptuomotor sites that produce the original experience.

There are a number of other ways in which Damasio's architecture differs from that proposed by Rapaport. The activation of a convergence zone does not merely depend on the simultaneity of incoming signals, and co-temporality does not suffice to unify features. If this were the case, then convergence zones would not be able to distinguish between different entities being experienced and feeding into the convergence zone at the same time. Other factors allow convergence zones distinguish between overlapping entities and events in experience, and to produce a gradation of responses. These factors are the differing size, locus, number, and location of the sites a convergence zone subtends, as well as the weighted potential trigger weights. The architecture Damasio describes resembles a connectionist network far more than a SNePS-like network. Convergence zones even allow for superposition of information, meaning that the combinatorial codes for different entities at a convergence zone can overlap, just as the weights of a connectionist network store information about a variety of inputs in the same set of weights.

What of the claim that what Damasio is describing is a form of syntax realizing semantics? Syntax is, in Rapaport's words, "pure symbol manipulation" (Rapaport 1996, 78). Is there really anything going on here that even resembles symbol manipulation? One might argue that the activity of the convergence zones, with their binding "codes," is a form of

symbol manipulation. The use of the term “code” is rather unfortunate. Convergence zones, and neuron ensembles in general, do not store codes. They are, instead, nonlinear dynamic systems whose parameters get tweaked by experience in such a way that they can bifurcate, or produce starkly different behavior from very similar sets of starting parameters, when different objects are encountered (these ideas will be explored more fully in chapters 6 and 7). The fact that convergence zones can produce pathological behavior, which is to say, get confused between inputs from different entities or events, strongly suggests that there are no discrete codes uniquely linked to particular entities. Even if we accept the notion of a stored code, it is clear that Damasio does not consider these codes to be representations.

Dynamic systems allow for combinatoric arrangements without the need for explicit codes describing those arrangements. Although Damasio writes, seemingly in the spirit of Fodor, that

> representations are interrelated by combinatorial arrangements so that their internal activation in recall and the order with which they are attended, permits them to unfold in a “sentential” manner. Such “sentences” embody semantic and syntactic principles,

he is neither endorsing the view that representations are combinatorial codes nor the view that representations are sentences in a language of thought. All that is addressed here is a timing issue, how the sequential occurrence of representations can be mapped to the sequential structure of human language.

Finally, there is Rapaport’s contention that Damasio’s theory fits well with his own Kantian “representative realism”: “Note, though, that we have here a neuroscientific analogue of a Kantian epistemology: Our conceptual schemes allow us to make sense of—to categorize—noumena . . .” (Rapaport 1996, 85). Damasio, however, sees categorization not as the imposition of our conceptual scheme on reality, but as a far more complex phenom-

enon. In criticizing the classical localization of function along traditional anatomical boundaries, Damasio argues that the structure, both the fragmentary records and the convergence zones, responsible for a particular function such as categorization are spread across anatomical boundaries and

> the region thus formed obeys anatomical criteria *dictated* by the nature of the entity represented, and by the interaction between perceiver and entity, and is *secondarily* constrained by the potential offerings of the anatomy. (Damasio 1989, 51) (emphasis added)

Here, we have the noumena dictating how they are represented, with neurobiological constraints being of only secondary importance. Neuroscientists tend to be full-blown realists rather than just "representative realists".

2.3.2.5 Computational 'Understanding' Is not Understanding

What implications does this analysis of the differences between Damasio's theory and Rapaport's computationalism have for our two questions? Even if we eliminate Damasio's theory as a candidate for how syntactic semantics is realized, we have not necessarily vanquished the latter's possibility. We may have simply weakened the case for the claim that computers cognize as humans do. The differences between the two, however, have revealed what computational systems as pure symbol manipulators currently can't do. If thought is merely the right input-output relations, then its realization is time independent (cf. van Gelder 1995). It does not matter if it takes a year or a nanosecond to produce the output for the process to qualify as a thought, nor is the timing of the process important to processes that use this output as input. One might attach a time-stamp to an outputted symbol, as is done in computational systems controlling time-critical operations, but this does not achieve what neuron ensembles are doing. The synchronization of neuron ensemble firing

is essential to the production of representations. Attaching time-stamps to computational markers and even synchronizing their production changes nothing essential about them.

Thus, we must answer the first question, whether computational systems can produce the same input/output relations, in the negative. But Rapaport is willing to accept that thoughts are not computations, so long as they are still computable. We have reason to believe from the Chinese Room argument that understanding is not computable. Rapaport's syntactic semantics is meant to counter that argument by showing how understanding can come from pure syntax. The key claim of Rapaport's theory is that self-understanding is necessarily syntactic; without establishing this, the base case of the recursive account of semantics would not be syntactic, implying that all of semantics is not syntactic. Self-understanding is syntactic, because it is a model of itself, a set of markers with connections back onto itself. Unless, of course, this is not what self-understanding is. The lack of evidence for the existence of such markers in mind/brains is one reason to reject this view of understanding. The error of identifying causal structure with syntax will be discussed in the next section. And finally, alternative notions of self-understanding will be presented in chapter 6. Absent these, there is a more powerful reason for rejecting syntactic semantics. Chapter 4 will show how systems using markers for representations and syntax to express the combinatorics amongst them always fall afoul of the frame problem.

The second question, whether the intermediary between input and output is moot, is itself almost moot at this point. More important is whether a computational system can do what a human can: understand its world. Still, we can't just dismiss this question. Rapaport argues that thoughts can be the results of an algorithm (Rapaport 1998). This could mean a variety of things. It could be that thought is the output of the computational process. This

would mean that thought is just a marker. Or it could be that a thought is the whole process: input, output, and how one gets there. Presumably, however, how one gets from input to output is immaterial for computationalists. Computationalists also hold that thoughts are individuated based on input/output relations, so where does this leave the intermediate process? Certainly, we don't want to leave it out. The idea that thoughts are just markers is ridiculous. Nor do we want to say that thought is just a set of markers in the right order. This would allow the possibility of a random generator of markers stumbling upon thoughts. So, it appears we are left with defining thought as any process that gets us from an input to the appropriate output, however that is defined. Computationalism doesn't explain consciousness; it only explains its effects.

2.3.3 Computation as an Abstract Description of Causal Relations

A third form of computationalism can be found in the work of David Chalmers. Rather than identify computation with the transformation of symbols, Chalmers views computation as an abstract description of the causal processes within a physical system: "A physical system implements a computation when the causal structure of the system mirrors the formal structure of the computation" (Chalmers 1996, 317-18). More specifically, there must exist a one-to-one mapping from the input/output of the physical system to the computation's input/output, as well as a one-to-one mapping from the physical system's internal states to the formal states of the computation that specify the transition between input and output. Furthermore, this mapping cannot be accidental. To avoid Putnam's (1988) criticism that any physical system could be seen to carry out any computation if it happens at any moment to display a series of internal states and input/output that map to a particular computation, Chalmers stipulates that the mapping must always hold for that particular process. A pro-

cess that does not map to the same computation each time it occurs is not truly a computation.

According to Chalmers's definition of a computational system, every physical system implements some computation. This means that rocks, mold, and the digestive tracts of animals are all computational systems. To distinguish cognition from digestion (or sitting around being a rock), Chalmers introduces the notion of an abstract causal topology. Something has an abstract causal topology if one can abstract its causal processes from their material instantiation in such a way that anything with the same input/output relationship as these processes can then replace them in the system. Digestion can be modeled as computation, but one cannot replace the parts of the digestive tract with silicon computational equivalents and still have a digestive system. Thus, digestion does not have an abstract causal topology. A thermostat, according to Chalmers, does have such a topology, and is therefore conscious. The difference between cognitive systems, like thermostats, and digestion is that the former is essentially computational; in other words, a system is cognitive by virtue of implementing the right kind of computations. Even this characterization does not fully capture Chalmers's view, because he believes that while rocks and stomachs are not conscious, they contain subsystems that are (Chalmers 1996, 297). A thermostat, unlike a rock, has "canonical experiences," experiences that "count" as the thermostat's (1996, 298). The rock has a conglomeration of experiences from its conscious subsystems.

Chalmers's account is computationalism's most severely flawed variety. First, it is plainly false that all physical systems implement computation as Chalmers defines it. A physical system that at the atomic level demonstrates Brownian motion is not implement-

ing computation, because there is no computable function that random motion can be mapped to. A random ‘function’ is not a computable function.

Second, Chalmers has failed to demonstrate that the mind/brain has an abstract causal topology. Until researchers can replace the brain's cognitive structures with silicon chips, this will remain an open empirical question. Further, because researchers have up until now failed to create artificial brains that clearly demonstrate cognition, the suspicion weighs heavily that cognition is like digestion, a process that requires certain physical substrates. Cognition would be the exception to the rule if it were otherwise. The burden of proof lies on Chalmers’s shoulders, and he can only write blank checks about what he thinks neuroscience will someday show.

Third, Chalmers's formulation of computationalism is inapplicable to plastic systems such as the human brain. Not only do neural structures carry out a variety of functions, they also change function in response to damage or learning (Kaas 2000). What Chalmers needs is a notion of the ‘normal’ function of a physical system as the evolved or designed function in order to constrain computational explanations to physical processes that don't just happen by accident to map to a computation. But normal function (Millikan 1984, 1993) does not apply to structures that are not evolved or designed, such as a rock. Thus, Chalmers must give up the strong modal implication between how a physical process maps to a computation and its being a computation, or his brand of computationalism does not apply to brains.

Fourth, the mapping of causal relations to formal states of a computation is itself a computable function, though not as Chalmers defines computation. Chalmers defines a computational system as a physical system that has a mapping of formal state-types that it is

mapped onto. To cognize the mapping between a causal and a formal system is to make a formal abstraction of a formal abstraction of a causal organization. Chalmers tries to hide this extra level by saying there *exists* a formal mapping that mirrors the causal structure. But mirroring is mapping. This mapping is itself a computation when carried out, but a computation in the usual sense of mapping one symbolic domain onto another. The mapping from causal relations to a formal system is itself not a formal abstraction of the causal organization. It is a function that takes as input the causal relations and that outputs the corresponding mapping between formal state-types. Thus, Chalmers's definition of a computational physical system is observer relative in the sense that it is a system that has a mapping to a mapping of formal states, and a system has a mapping only in virtue of possible cognitive agents being able to specify the mapping. A similar definition would be that physical systems are computational systems in virtue of there being scientific theories about them.

Finally, Chalmers's version of computationalism is highly susceptible to criticisms from proponents of dynamic systems theory (Globus 1992; van Gelder 1995). Dynamic systems theorists argue that the systems they study are not decomposable into transformations between state-types. Thus, there are a host of non-random physical systems that do not have mappings to mappings of formal states. An example we will see when we take up dynamic systems theory is the Watt governor, a system whose global function can be implemented in a computational manner, but whose own functioning cannot be decomposed into state-transitions.

2.3.4 Computationalism as Information Processing

Chalmers has tried to avoid the question of what computation contributes to a causal system by defining computation as an abstraction (of an abstraction?) of causal relations. Still, Chalmers has found it necessary to impute special status to the computation of cognitive systems over non-cognitive systems to avoid equating cognition with digestion. Replaceability of parts is this special distinction, though it is unclear why this would change the nature of a physical system. Suppose 1% of the stomach were replaceable with silicon: would this mean that digestion was now the same kind of special computation as cognition? Chalmers suggest that a special set of computations is needed for cognition, and presumably these computations are not found in the gut.

Mainstream computationalists have identified the type of these special computations as information-processing functions, though there is wide controversy on the specific functions themselves. Any device that does the same information processing as the mind/brain presumably can replace components of the mind/brain or be used to build a mind/brain, fulfilling Chalmers's replaceability requirement while not fulfilling the requirement that equivalent computational systems have equivalent formal mirrors of their causal organization.

Mainstream computationalism finds expression in both neurobiology and cognitive psychology. Since, as noted by Fodor (1975), cognitive scientists are at least token materialists, the prevailing view is that mental states are realized by computational brain states, and so at the very least are the products of computation. Whether attempting to explain mental states from the standpoint of cognitive psychology or of neurobiology, however, the computational nature of the mind/brain leads to the methodological assumption that these

states can be isolated. If a mental state is a computable function, then all that matters for explaining its nature is the mapping between its input and output relations and its relations to other mental states. The context of the function evaporates-it can be safely ignored. Philosophers express this as ‘meanings are in the head, not in the world’, meaning that mental states are individuated by examining their functional aspects: their input, output, and relations to other mental states.

2.4 A Step in the Right Direction: The Dynamical Systems Critique of Computationalism

There are several well-known challenges to computationalism, among them the recalcitrant strains of behaviorism and Hubert Dreyfus’s (1972) critique of AI. Recently, several philosophers and scientists have proposed a new challenge to computationalism in the form of a dynamical systems theory approach to cognition. Rather than a symbol-processing machine, philosophers such as Tim van Gelder view the mind/brain as a nonlinear dynamical system that is most appropriately understood in terms of state-space equations. Cognition is a result of the dynamics of the given system, and not due to any computational scheme. The dynamical systems approach, according to van Gelder, embraces the likes of connectionism, neurocomputational approaches, and artificial life, and provides a real alternative to the mainstream computationalism so dominant in cognitive science (van Gelder 1995).

2.4.1 Van Gelder’s Example of Dynamic vs. Computational Systems

To illustrate the difference between a computational system and a dynamical system, van Gelder describes two implementations of a governor, a device, in this instance, for auto-

matically adjusting the throttle valve on a steam engine in order to maintain its flywheel at a uniform speed. This can be achieved with the following algorithm:

1. Measure the speed of the flywheel.

2. Compare the actual speed against the desired speed.

3. If there is no discrepancy, return to step 1. Otherwise,

   A measure the current steam pressure;

   B calculate the desired alteration in steam pressure;

   C calculate the necessary throttle valve adjustment.

4 Make the throttle valve adjustment.

It can also be accomplished with James Watt's centrifugal governor, a device that does not make use of an algorithm. Instead, it solves the problem by the very nature of its construction. The Watt centrifugal governor's behavior is describable in terms of a second-order differential equation, but not, according to van Gelder, in terms of computation. He argues that computational systems have the properties of representation, computation, sequential and cyclical operation, and homuncularity that dynamical systems do not possess.

Representations in the case of the computational governor are the measurements of the steam pressure and speed of the flywheel, measurements that are stored and used as symbols. While the centrifugal governor behaves in a manner such that the angle of its swinging arms is related to the speed of the engine, the angles cannot be viewed as representations of the speed for four reasons. First, there is no utility in describing the relationship in representational terms, since the behavior of the centrifugal governor can be explained in

purely nonrepresentational terms. Second, while there is a statistical correlation between angle of the arms and the speed of the engine, this is not strong enough to establish a representational relationship. The governor and engine are co-determining one another, so there is not a uni-directional representation-represented relationship. Certainly, a discrete, symbolic representation of the engine speed cannot be given in terms of arm angle. Third, the relationship is not even mere correlation, because the correlation only obtains when the total system has settled into a stable equilibrium point, and no longer obtains when pushed out of this equilibrium. Finally, the appropriate “conceptual tool” for understanding the centrifugal governor is not the notion of representation, but the concepts of dynamics.

Because the centrifugal governor does not have manipulable representations, the hallmark of computation, according to van Gelder, it cannot be computational in nature. The sequential, cyclic nature of the computational governor, namely that it repeats a cycle of sequentially measuring angle and current speed and then adjusting the speed accordingly, cannot be found in the centrifugal governor. As long as the computational governor makes its computations within the appropriate amount of time, its operations otherwise have no time constraints. Measurement could take 90% or 10% of the time of the adjustment cycle, the exact amount being irrelevant. The centrifugal governor, however, has no such freedom from time constraints because its behaviors are all contemporaneous, not sequential.

The homuncularity constraint of a computational system is also often referred to as modularity. By 'homuncularity' van Gelder means that the overall functioning is decomposable into simpler functions carried out by simpler modules, whose interaction is the communication of results. Such is not the case with a dynamical system like the centrifugal governor.

2.4.2 Connecting Connectionism and the Dynamic Systems Approach

Van Gelder accepts connectionism to be a subset of the dynamical systems approach. In doing so, he undermines his critique of computationalism. According to van Gelder, connectionist systems differ from traditional computationalist systems in a variety of important ways. Computationalist systems manipulate symbols, and symbols are the basic elements of representations. While connectionist systems have representational interpretations, they also have numerical, not symbolic operations, and these analog numerical operations are usually nonlinear dynamical functions. This places connectionism within the dynamical systems model:

> The core dynamical hypothesis is that the best models of any given cognitive process will specify sequences, not of configurations of symbol types, but rather of numerical states goes hand in hand with a conception of cognitive systems not as devices that transform symbolic inputs into symbolic outputs but rather as complexes of continuous, simultaneous, and mutually determining change . . . (van Gelder 1995, 373)

However, this distinction is unprincipled. Numerical computation *is* symbolic computation. Indeed, it is almost humorous to think of standard treatments of the theory of computation, in which computation is defined in terms of transformations of input 0s and 1s to output 0s and 1s, in the context of van Gelder's critique. Van Gelder has apparently forgotten what connectionist systems are implemented on, namely symbolic processors (computers). While the connectionist paradigm, in its emphasis on analog computation, is closer to modeling the brain than traditional symbolic systems, the distinction does not imply that connectionist systems are not computational systems.

Perhaps van Gelder is suggesting that the nonlinear differential equations used in connectionist systems are not computable functions, and therefore connectionist systems are

not computational systems. This, of course, is false. Were it true, no one could implement a connectionist system on a computer.

The attraction of connectionist systems to van Gelder and others lies in their use of vectors and weights. Because it is implausible that there are 0s and 1s in the brain, and the networks of neurons in the brain are somewhat describable in terms of vectors of processing units, the superficial attraction of connectionism is obvious. But it is equally unlikely that the brain stores and manipulates vectors of floating point numbers (which computer implementations of neural networks do) as that it stores and manipulates 1s and 0s.

The homuncularity constraint on computational versus connectionist systems is another unprincipled distinction. At some level of computational systems, there are primitive functions that are no longer decomposable. Connectionist systems are more difficult to decompose, and a particular connectionist system might not be decomposable down to each value of its weights. The function a connectionist system computes is not, in its implementation, a primitive function. But then again, from the standpoint of the connectionist, no function is primitive, because all functions can be approximated to an arbitrary degree by using the Fourier approximations. This can be achieved in a connectionist system by using a feedforward network with one hidden layer of neurons computing cosine functions (Hecht-Nielsen 1989). What a connectionist system is doing is decomposable to a degree largely ignored by the proponents of the connectionist/symbolist distinction in cognitive science. To say that connectionist systems show no homuncularity seems to be nothing more than an effort to carve out a philosophical distinction where there is none.

As for van Gelder's arguments concerning the nature of representational versus nonrepresentational systems, they boil down to the following: if the behavior of a system is describable in terms of a differential equation, it is not a representational system.

2.4.3 Why van Gelder's Example is Not Enough

By seeking to place connectionism in the domain of dynamical systems approaches to the nature of cognition, van Gelder has blunted the legitimate criticisms of computationalism stemming from research into biological dynamical systems. His analysis of the differences between a computational governor and a centrifugal governor points to, though never explicitly names, the essential distinction between the two. The computational and the centrifugal governor both have as their gross function the maintenance of uniform engine speed. The possibility of a computational governor standing in for a centrifugal governor demonstrates that the centrifugal governor's gross function is computable, after Rapaport's usage, though not that it is indeed doing computation. Van Gelder attempts to draw this distinction by arguing that computational systems are symbolic, and therefore representational, whereas dynamical systems are numerical, and therefore nonrepresentational. As has already been argued, the distinction between symbolic and numerical systems is unprincipled, and so does not sustain van Gelder's representational/nonrepresentational dichotomy. Furthermore, if we consider cognitive systems at the ecological level of analysis, i.e., as persons functioning in an environment, dynamical cognitive systems are representational; this is their function in the sense of it being their 'purpose.' The purpose of a centrifugal governor is not to represent, but to control the speed of an engine. Since both the computational and the centrifugal governor achieve their purpose, that one is computational and another isn't doesn't matter to whether they are governors. Why should it matter

for cognitive systems whether they are computational or not? Rapaport is willing to concede that many systems are not computational, but what they do is achievable by a computational system. Although Rapaport contends that mind/brains are actually doing computation, it is enough for him that their functions are computable. Just as we can mimic a centrifugal governor as a computational system to a degree adequate to its gross function, so too, contends Rapaport, could we mimic a biological cognitive system with a computational system to a degree that it qualifies as a cognitive system. Van Gelder must argue that the implementational details of the biological cognitive system are more important than satisfying the gross, ecological level function of representing. Otherwise van Gelder has merely established that cognition doesn't happen to be computation, though it could be implemented as such.

In one sense, van Gelder's use of the two types of governors as a means of distinguishing representational from nonrepresentational was a poor choice. Both governors are engineered for the explicit purpose of controlling the speed of an engine, and therefore the implementational details matter only if they result in one governor performing better than another. Rapaport himself has provided a better example: the solar system. As has been argued above, even though the laws governing the solar system's evolution are computable, creating a computational model is not equivalent to making a solar system. A closer computational approximation would be to create large pseudo-planets and control their motions according to Kepler's laws with computers firing immense rockets. But if we could create these pseudo-planets, Kepler's laws would apply to them regardless of our computers being on board. Thus the implementational details in the case of the solar system are essential. There are, of course, numerous solar systems in the universe, and these solar systems have

different planets than our solar system. A solar system is a functional type at a certain level of analysis only. Structure does make a difference to a certain degree. Consider what level of structure it takes for something to count as a star. Computationalists argue, however, that structure makes no difference so long as you have a universal Turing machine available, as universal Turing machines can compute any function that a mind computes. Van Gelder has undermined his argument by relying on an example where the computationalist approach is as good as the noncomputational approach. What is the principled distinction between the two approaches that renders computationalism impotent to produce cognition?

That principled distinction is between information-processing and noninformation-processing systems as cognitive scientists understand information-processing. A solar system is an example of a noninformation-processing system, a system that can be modeled, but not replicated by computational systems. This is because solar system is not a pure functional type in the way that governor is. What counts as a governor is determined by the function it serves. What counts as a solar system depends also on the structure of its constituents and their causal interactions. For van Gelder's argument to work he would have to focus on the fact that a computational governor could never be a centrifugal governor, and, in fact, where he does make this argument, he is on solid ground. But he cannot make this argument with the distinction between symbolic and numerical systems, which becomes the main thrust of his argument, as there is no real distinction between the two. The argument to be made then is that information-processing systems, in the sense of information intended (vaguely) by cognitive scientists, cannot be true ecological-level representational systems, that computational systems cannot carry out the proper functions of cognitive systems. Such 'in principle' arguments about empirical matters are always dangerous, so we

will make the more tempered argument that there are strong reasons to believe that computational systems cannot replicate human cognition.

2.5 Refining the Dynamical Systems Critique

Although van Gelder has blunted the force of his critique of computationalism through his efforts to save connectionism, he has nonetheless laid the foundation for a true dynamical systems challenge. Van Gelder's contention that dynamical systems do not possess manipulable representations points to the need for new notions of information and information-processing. His discussion of reciprocal causation highlights an essential difference in mechanisms between computational and dynamical systems. Finally, his insistence that a computational governor cannot do what a centrifugal governor is capable of doing despite the equivalence in gross function suggests a stricter criterion for judging functional equivalence. We turn now to developing each of these aspects of van Gelder's critique.

If a cognitive system does not possess simple, manipulable representations, then its complex information-bearing states cannot be explained in terms of their decomposition into simpler representations. Just how radical this conclusion is has apparently not occurred to Van Gelder. One of the primary avenues for criticizing connectionism has been the failure of connectionist architectures to mimic the ability of humans to compose representations, and, in response, connectionists have focused on showing how their architectures can indeed compose representations as humans do. The dynamical systems perspective is that the (de)composability of representations is a myth, because mind/brains do not have representations as understood by cognitive scientists. Rather than arguing that neural nets can build representations too, dynamical systems theorists ought to provide an alternative

notion of information. J.A. Scott Kelso (1997) has presented one such notion. He argues that the order parameters that capture the relations between aspects of dynamical systems are themselves informational:

> order parameters in biological systems are functionally specific, context-sensitive informational variables . . . Order parameters are semantic, relational quantities that are intrinsically meaningful to system functioning. What could be more meaningful to an organism than information that specifies the coordinative relations among its parts of between itself and the environment? This view turns the mind-matter, information-dynamics interaction on its head. Instead of treating dynamics as ordinary physics and information as symbolic code acting in the way that a program relates to a computer, dynamics is cast in terms that are semantically meaningful. (Kelso 1997, 145)

Kelso's formulation is susceptible to a Searlean form of criticism: order parameters are not intrinsic to the physics of dynamical systems, so he is conflating the description of a system with its nature. But Kelso has not made the same error as those who impute syntax to the operations of the brain. An order parameter captures a known behavior of a system without requiring the assumption that the order parameter is somehow intrinsic to the system in some sense. What Kelso is really arguing is not so much that order parameters are semantic, but that the relations among parts of the system are intrinsically meaningful-they constitute meaning-and so the order parameters that describe these relations describe the semantics of the system.

This is all very vague, and will remain so until we attempt to describe some of the relations of the mind/brain dynamical system in chapters 6 and 7. However, what is clear is that the dynamical systems approach requires a radically different notion of information from that proposed by the proponents of symbolic models of mind. The beauty of such symbolic models is lost: the composability of symbols mirrors the apparent composability of representations. What the dynamical systems theorist gains, however, is an immediate link between brain and the environment, and thus is not faced with the monumental task of

explaining how symbols have meaning. As we will see shortly, some computationalists have recognized the benefits of the dynamical systems approach, and attempted to copy it by extending the notion of a computational cognitive state to include aspects of the environment.

Physicists have long marvelled at the astonishing capacity of mathematical/computational models to explain the workings of the universe. Why should the universe conform to mathematical description? Or thinking of the problem at hand, why should the brain conform to computational description? One answer is that perhaps it does not. Computational descriptions of brain processes have one glaring advantage over dynamical systems approaches: they are much easier to understand. This ease is due in large part to the decomposability of computational processes. Computers are designed as they are not only for the sake of efficiency in processing, but to facilitate analysis of their workings. Engineers design machines with the goals of getting the job done and understanding how this was accomplished. Accomplishing the latter goal enables the engineer to fix whatever goes wrong in the machines. A computational approach guarantees the engineer that for any given input, there will be one output that it is mapped to. But Nature does not need to satisfy this constraint, and this is where the all-too-often used analogies between Mother Nature and an engineer break down. For Mother Nature, transparency of function is unnecessary; what is important is that the 'job' gets done. If it doesn't, Mother Nature is capable of expending considerable resources trying other designs. This widely acknowledged fact leads to the suspicion that computational explanations of the mind/brain are so popular because they indulge our prejudice for transparent design. To recast the dynamical systems

critique in evolutionary terms, why would computational processes be selected for, or, even more perplexing, how could they self-organize?

Daniel Dennett and Richard Dawkins have each made this question seem unnecessary by characterizing mental development and genetic evolution as processes of incremental building. A computational design seems to be the most advantageous option for an organism if at every stage of its development/evolution, its behavior must give it some benefit. Being decomposable also means being composable, so computational processes can be slowly added with each prior stage still providing an advantage. For example, an organism that has only simple edge detectors in its visual system might later 'add' line detectors, which take as input the output of the edge detectors. Dawkins (1976) has offered an evolutionary/genetic account of this process, and even theorized how the genotype of an organism is decomposable in a way similar to its phenotype. Genes 'for' edge detectors provide a selective advantage, and genes 'for' line detectors provide a greater advantage, albeit in the context of there already being edge detectors. But Dawkins's Selfish Gene metaphor does not imply that Mother Nature has a predilection for decomposable, let alone, computational processes. Furthermore, it would be incorrect to think that a dynamical system cannot be incrementally built. Rather than composing computable functions, Mother Nature might be tweaking the parameters of various dynamical systems, resulting in better performance or even novel functioning. Evolution may add components to the dynamical system in question, but this is not the same as composing computable functions. A whole new dynamics arises from such an addition.

So is there any reason to believe that natural selection favors one approach over another? Yes, because computational systems have an inherent drawback not found in

every dynamical system: their brittleness. The nature of a computable function is to map a unique output to each input. An error in output at one level means an error in input to the next level of the computation. This problem can be corrected by routines that check the output, but this increases computational cost. But such routines are necessary to prevent catastrophic breakdown. Examples are: Intel's FDIV bug, a design error in their Pentium processor that resulted in floating point division errors, errors that compounded if one computed a series of divisions; spray-paint robots on assembly lines turning and spraying one another; and robots that drive autonomously suddenly spinning out of control. This is not just an artifact of human error in designing computers and robots, although it is best highlighted by such cases. Error will inevitably creep into an organic system that is in the process of dying, as we are. Hysteresis, a lag in effect when forces change, is just one property of dynamical systems that may act to dampen error. All computational systems are threatened with catastrophic failure due to compounding error, not all dynamical systems are. This seems to be a clear advantage for the latter and reason to suspect that Nature has selected for a dynamical rather than a computational cognitive system.

The differences in mechanism between computational and dynamical systems responsible for this difference in the effects of error is paramount. Computational systems are mapping systems: their causal processes are constructed so as to conform to a mapping between abstract domains. In the case of the computational governor, an error in the measurement of the fly-wheel results in an error in the adjustment to the throttle valve. In the case of the centrifugal governor, the relation between the speed of the fly-wheel and the throttle valve is not a mapping between states. Instantaneous measurements of the states of both reveal only a statistical correlation. The mechanism for adjusting the throttle valve is

built into the system; no measurement is required. Nor do systems that dynamically adjust in this manner employ mechanisms to check for error in their components. Rather, the global error of this system can be measured, and parameters tweaked to reduce the error. A functional decomposition is not available, because such systems are not designed by composing functions. What this means is that if the mind/brain is such a dynamical system, functionalist philosophy of mind is false. At best, the global functions of brain might be replicated in silicon. The question is now whether cognitive functions are replicable, or simply the behavioral outputs of these processes.

The proponents of the dynamical systems approach to cognition must admit that the teleonomic or purposive function of a given dynamical system can be replicated in computational systems. Both the computational and centrifugal governors serve the purpose of maintaining appropriate steam pressure by adjusting the throttle valve. Similarly, robotic arms and grippers can pick up garbage and throw it in the trash just as humans do. But this does not imply that robotic arms are replicating the mediating processes that result in human arm motion. Cognition is itself a mediating process for advantageous behavior. That a computer can add 2 and 2 to get 4 does not imply that its mediating processes are forms of cognition as the mediating processes in humans are when they do addition. What the dynamical systems proponents have demonstrated is that the mediating processes in dynamical systems are different from and not replicable by computational systems. The functions of the two forms of governors are the same, but the mediating processes are radically different. Similarly, the type of information-processing found in a computer is radically different from that in a putative dynamical cognitive system, and cannot replicate the

latter's processes. This is why functionalist answers to the dynamical systems challenge are inherently inadequate.

2.6 Problems with Clark's Strategy of Partial Programs

In his recent book *Being There: Putting Brain, Body and World Together Again* (1997), Andy Clark concedes that cognitive science has focused too much on disembodied computation, and that the criticisms of ecological psychologists, situated roboticists, and dynamical systems theorists must be incorporated into cognitive science if it is to make headway. In particular, computationalists have ignored the extent to which the environment is used by sentient beings to constrain and simplify their cognitive tasks and discounted the idea that the environment ought to be considered part of the cognitive system itself. Clark makes two important suggestions as to how computationalists can incorporate these criticisms without surrendering the basic computational account of cognition.

First, Clark suggests that more emphasis be placed on what he calls action-oriented representation. This is representation that is tailored to specific, local tasks rather than to generic problem solving. Much of the representational burden is "offloaded" to external representations or constraints. Thus humans and other animals tailor their environments to simplify tasks, whether it be by leaving post-it notes, as in the case of humans, or chemical trails, as in the case of some insects. Language is the most powerful tool at our disposal for offloading computational burden and restructuring problems to make them easier. Action-oriented representation is a fusion of environmental cues or external representation systems and internal representations. In Putnamian terms, representations must be individuated in terms of wide content, and Clark seems convinced that a computational scheme is up to this

task. Dynamical systems theorists might be right that certain behaviors are best understood by the tools of dynamical systems theory and not in terms of representations, but these phenomena were not 'representation-hungry' in the first place. By this, Clark means cases in which it is hard to imagine how the phenomena could occur without internal representations. These are cases where a system must use an inner code or pattern to "stand in for" states of affairs that are not immediately present to the system, and these codes are "usefully individuable" (1997, 168). "Usefully individuable" denotes "fine-grained assignments of inner vehicles to information-carrying adaptive roles" (Ibid).

The second way in which computationalists can refine their theories is by de-emphasizing the picture of the mind/brain as a central executor carrying out a stored program for cognition. Daniel Dennett has made a similar point with his multiple-drafts model of consciousness. The simplest and most compelling argument in its favor is that of the poverty of the genetic code for storing the program(s) implementing consciousness (Eigen 1992). The brain has on order of 10^{10} neurons and 10^{14} connections between neurons, well beyond the storage capacity of the genetic code to even set up the architecture for running the program. Instead, Clark advocates limiting the explanatory role of computation to localized, partial programs. Clark likens the difference between the two approaches to two methods for doing one's taxes: the first has an elaborate LISP program to do the step-by-step calculations, whereas the second is merely the stored command Do Tax that runs on hardware specifically for calculating taxes. The dynamical systems theorist can explain how the latter method's hardware comes about, but ultimate explanation lies in the simple stored program.

In making these concessions, Clark has bitten off more than computationalism can chew. Rather than simplifying the explanation of cognition, Clark complicates it by requiring an interface between the dynamical systems and the partial programs he views as running the subsystems of the mind. Clark suggests that the oscillation of a set of neurons is not only a dynamical system, it is also an information-bearing system. For the partial programs to use this information, they must be able to decode the information from the dynamical system, a herculean task for neuroscientists though not necessarily for neurons. Such a decoding scheme would presumably be part of the stored partial program, increasing the complexity of what is to be stored. Just as with Daniel Dennett's multiple-drafts model, these partial programs require integration amongst one another for coherent mental states to arise. Is this integration done by another program, or are the partial programs aspects of a global dynamical system which itself is not computational? As with Dennett, Clark does not answer this question.

Clark's response to the dynamical systems theorist is weakest when he considers the constraint of contemporaneous, reciprocal causation that dynamical systems exhibit, and so Clark tries to restrict the instances of pure dynamical systems to non-representation-hungry cases. The causal organization of dynamical systems is generally not decomposable into a step-by-step causal chain similar to what we find when we describe the motion of a billiard ball. Like those of the Watt governor, the components of the dynamical system are in a reciprocal causal relationship. Further, the behavior does not occur in a series of discrete time steps. Initial conditions of the system also factor heavily into the behavior of the system, so that two seemingly identical dynamical systems that are started with different initial conditions can exhibit startlingly different behavior. There are two important consequences

of these aspects of dynamical systems. First, such a dynamical system is not a computational system, either in the mainstream or in Chalmers's sense. Second, partial programs that interface with them must not only be able to 'start' them, but also be cognizant of the proper initial conditions for starting them. This increases the computational load on the partial program to such a degree that it loses its explanatory value. Remember that partial programs were part of a system that offloaded computational burden. So even if the dynamical systems approach only applies to non-representational-hungry phenomena, the partial programs for representational hungry problems, by Clark's own admission, still make contact with these basic means of cognition, and thus require an interface between them. In many cases, it doesn't take much to kick-start a dynamical system, unless, of course, you want it to behave in a certain way. Then it takes substantial computational work to make sure the initial conditions are right. In the end, fully computational system and fully non-computational systems are less complex than hybrid systems.

2.7 The Uses and Limits of Computationalism

It is impossible to limit the dynamical systems critique of computationalism to symbolic AI and language-of-thought models of mind, as van Gelder has attempted. Connectionist and neurocomputational models of the mind/brain employ the very same notion of computation as that found in symbolic AI, and later I will demonstrate just how little connectionism departs from traditional AI. Yet, neurocomputational research appears to be yielding significant results, allowing scientists to predict and affect how the brain operates (Churchland and Sejnowski 1992; Gazzaniga 1995). If computationalism is inherently flawed, how do we explain these results?

I do not endeavor to solve this apparent paradox at this juncture, leaving it for the final chapter on an alternative approach to the mind/brain. Instead, I will merely hint at its solution. Neurocomputational models of the brain restrict themselves to small subsets of brain activity. They primarily describe the behavior of single cells and maps of cells taken in isolation. It is not surprising that a computational model can simulate the behavior of a cell, any more than it is surprising that a computational governor can simulate the behavior, grossly understood, of a centrifugal governor. The key is that this is only simulation, and the computationalists generally acknowledge this. Thus, cells in the visual cortex are thought to respond to certain stimuli 'as if' they were computing the Laplacean of a Gaussian (LOG) function. But the behavior of the cells do not perfectly map to a LOG function, or any of the other similar proposed functions. Computationalists think that they have just not found the correct function. An alternative theory is that the cells are not computing a function, though their behavior can be grossly described by a particular function, but contributing to a dynamic that we do not yet understand because we do not even remotely understand the global dynamics of the brain. The components of a centrifugal governor could be modeled computationally, and such a model could yield useful predictions, but they would not enable us to explain the global dynamics of the system.

The successful modeling of single cells and maps of cells is supposed to tell us the place of these components in the informational flow of the system. There can be little dispute that they do tell us something of this nature: when a subject sees an edge in his visual field, and a simple cell fires consistently when, and only when, an edge of that orientation is present, there can be little doubt that this cell is at least partly responsible for the subject 'seeing' the edge. What is in dispute is whether this cell computes a function indicating the existence of

the edge from the information flowing to it from the visual pathway, and whether it passes this information on, encoded in its firing pattern. Only if the nature of information passed in the brain is equivalent to that passed in a computer can we conclude that cognition is computable. The dynamical systems theorists have given strong reason to doubt this is true. If they are right, then the methodology of computationalism is a dead-end street.

2.8 Computationalism's Failure to Explain the Nature of Representation

According to computationalists, minds carry out information-processing operations by carrying out computations. The abstract tokens that form the input and output of computations are not merely syntactic markers in the view of computationalists: they are symbols standing for objects and events in the world. Thought is the act of transforming symbolic input into symbolic output: the production of semantic information. Thus, the task of explaining what semantic information *is* is transformed into that of explaining how abstract tokens become 'grounded' symbols.

Just as semantic information could not be understood in terms of syntactic information, so too the representational capacity of minds cannot be explained in terms of formal rules over abstract tokens. Representations track the world; they are plastic and adaptive. The environment of a representational system is essential to the formation of its representations. The environment of a computational system is just its input space. Computational systems are incapable of understanding, whether Chinese or how to bake a cake, because they do not adapt themselves to the world. Which symbol stands for 'cake' is entirely arbitrary, and

by virtue of what it stands for cake is indeterminate. In fact, what relation ties a symbol to its object is also arbitrary (causal connection is not a necessary part of a symbol system).

Dynamic systems theory has emerged as an alternative to computationalism, and not merely a complementary mechanism. It offers an account of how representational systems attune themselves to the world that eschews the need for symbol-grounding explanations. Connectionism, on the other hand, is a compromise between computationalism and dynamic systems theory. While computational in its implementation, connectionism is often regarded as an exception to computationalism (van Gelder 1995), possessing computationalism's abstract realizability, while discarding its implausible aspects (such as symbol manipulation and seriality). In the next chapter, I will address the theory of connectionist exceptionalism. What will emerge is that connectionism bears the worst faults of computationalism, while possessing little of the aspects of minds as they are realized in real brains.

Chapter 3 Connectionism: Computationalism's Prodigal Son

The reemergence of artificial neural network research during the 1980s gave rise to a new philosophy of mind known as connectionism. Artificial neural networks (ANNs) provided connectionist researchers with a tool for modeling cognitive behavior which resembled the structure of the brain as it was being uncovered by neuroscience. While connectionists parenthetically noted the important differences between biological neural nets (BNNs) and their artificial cousins, the similarities were considered to be an important gain over traditional AI tools. The structure and functioning of ANNs also suggested a radical view of representation, one that contradicted the long-standing notion that representations were implemented in minds via atomic symbol tokens. This encouraged the view that a connectionist system is a different kind of computational system, because its primitive computations are not over representations. The computation that connectionist systems carry out is termed subsymbolic, because it is computation over symbol tokens, which do not have a representational interpretation. Representations are held to emerge at the network or system level, rather than at the level of symbol tokens. This distinction is held to provide a variety of benefits, such as enabling subsymbolic systems to avoid Searle's Chinese Room argument (Chalmers 1992).

In this chapter, I will examine claims that connectionism is a significant departure from traditional computational approaches in symbolic AI, one that holds the keys to solving AI's problems. Researchers such as Churchland, Smolensky, and van Gelder have argued

that connectionism possesses a form of computational exceptionalism, that its computational methods do not suffer the same drawbacks as traditional symbolic systems. Contrary to this claim, I will demonstrate that connectionism still harbors the flaws that prevent traditional AI from achieving the goal of implementing thinking agents, that the exceptionalism of its representational interpretation is in some cases irrelevant, in others non-existent. In addition to detailing the general shortcomings of connectionism, I will examine several examples of connectionist systems: the neural network for vehicle navigation known as ALVINN, Phillipe Schyns' work on categorization using self-organizing neural networks, the state of the art in connectionist natural language processing (NLP), and a neurocomputational model of mammalian navigation. Regarding the first two, I will show that the claims concerning their ability to think or adequately model human cognition are false. Connectionist NLP will return us to the problem of symbol grounding, demonstrating that connectionist systems, just as symbolic systems, require the intervention of a cognitive agent for symbols to be grounded; hence, they do not truly ground their symbols. The last example will show that ANNs cannot provide a true dynamical systems account of neural performance, countering van Gelder's (1995) claims that connectionist and neurocomputational models represent dynamical systems rather than computational approaches.

3.1 What is Connectionism?

Connectionist networks are collections of simple processing units working in parallel to transform inputs to outputs. This transformation is generally achieved by passing the inputs through weighted connections and applying a transfer function to the sum of the product of weights and inputs. The simple processing units, often referred to as neurons, are usually

organized in layers. Connections between neurons can be inter- or intra-layer, inhibitory or excitatory, integer- or real-valued. The transfer function applied to the sum of the weighted inputs can be either linear or nonlinear, e.g., the commonly used nonlinear transfer function tanh. The network can have a single layer or multiple layers, although the latter is far more common. Layers between the input and output layers are referred to as hidden layers. The activity of the hidden layers of a network is generally the focus of any analysis of the network's performance, because the hidden layers are the ones that form representations of the input-output relationships.

Hidden-layered networks have important properties that make them desirable for use in connectionist modeling. Hidden layers enable the input to be cast nonlinearly into a higher-dimensional space. According to Cover's theorem (1965), a problem cast nonlinearly into higher-dimensional space is more likely to be linearly separable[1] than one cast in low-dimensional space. Therefore, using a hidden layer with greater dimensionality than the input space increases the chances that a solution to the problem at hand will be found. Furthermore, hidden-layered neural networks such as backpropagation nets and Boltzmann machines can serve as universal function approximators, capable of approximating any continuous function to an arbitrary degree (Haykin 1994).

Not all networks, whether multi-layered or not, count as connectionist networks. An important constraint on connectionist networks is that the computations are all local. What this means is that each neuron only knows about its own inputs and outputs. There is no global information directed to a neuron, such as a rule to the effect: If 5 other neurons

1. Linear separability refers to the ability to separate a problem space into the desired classes by partitioning with a hyperplane. The hyperplane separating points in 2-dimensional space would be a line, in 3-dimensional a plane.

output 1, also output 1. This constraint disqualifies certain activation spreading models from connectionism, and is based on the assumption that a similar constraint holds in BNNs. This constraint figures importantly into connectionist explanations of cognition, as connectionists claim it enables their systems to discharge any homunculi by building intelligent systems out of dumb components.

Connectionism can be defined in part by the departures it takes from symbolic AI. The following is a brief list of connectionist criticisms of symbolic AI and the connectionist alternative, loosely following Churchland and Sejnowski (1989) and Smolensky (1988):

1. The brain is a parallel, rather than a sequential, processor. While symbolic AI tends to view conscious thought as sequential processing of information, although not necessarily sequential, connectionists argue that the parallel nature of the brain is essential to how conscious thought occurs. Connectionist models therefore consist of massively parallel nets of simple neuron-like processing units, each neuron computing in itself a simple function, which contributes to the global function computed by the net.
2. Symbolic AI models cannot match the speed at which the brain processes, despite the relatively slow rates at which neurons fire. Connectionist models achieve high rates of processing by means of the same mechanism as the brain- namely parallel processing.
3. Connectionists contend that the memory of their models is content addressable, rather than location addressable. This means that instead of retrieving a memory by providing the location or address of that item in the brain/computer, memories are retrieved in connectionist models by matching or completing a key based on the

content of that key. Many symbolic AI models store specific memory items in specific locations, to be retrieved by specifying the location. Since representations in connectionist models are massively distributed, being stored as weights between neurons, no specific location can be given for a memory item. Retrieval is achieved by presenting the network with input in some way associated with the memory item. This objection to symbolic AI holds only for those symbolic systems that do not use content addressable memory.

4. Connectionists complain of traditional AI's dependence on Von Neumann style architecture. Again, this is a complaint about the sequential nature of the latter's models, this time with regard to the hardware used.

5. Connectionists such as Churchland and Sejnowski reject the hardware/software distinction regarding the mind/brain made by functionalists. The behavior of a neuron is not due to a program running on it; rather it is tied to its physical structure.

6. Symbolic AI models only a small fraction of mental life, namely higher cognitive functioning. In doing so, its practitioners have elaborated a series of theories that have no application to animal cognition, and as such, posit an enormous leap between the cognitive capacities of humans and animals. Such a gap is not warranted from an evolutionary perspective, since the development of conscious through evolution is undoubtedly a continuum (saltational accounts of evolution being generally rejected). Do we posit a more primitive language of thought for animals? Even a primitive language of thought could serve as the basis for a linguistic capability, albeit primitive, lacking in the vast majority of animals.

7. The sentence/logic model of cognition employed by symbolic AI is clearly not a true theory of how humans think. Research by Nisbett and Ross (1980) has shown that humans generally do not follow logical patterns of thinking, instead employing all sorts of shortcuts, hedges, and even fallacies to draw inferences. Human thought that does not conform to the structure of formal logic, i.e., intuition, is better captured by probabilistic or constraint-satisfaction models. Both types can be implemented as connectionist networks.

Thus, connectionist criticism of symbolic AI focuses on three areas: the neural plausibility of symbolic models, the form of cognitive processing implied by these models, and the nature of the representations processed by symbolic systems. But before we assess each of these claims, we will first take a closer look at the structure of the ANNs that are most often cited as examples of connectionist systems.

3.2 A Brief Overview of ANNs

A host of neural network paradigms have sprung up since the work of Kohonen, Werbos, Rumelhart, Smolensky, McClelland, and Hinton in the early- to mid-eighties that brought neural network research back to life. As we have seen, an ANN is a collection of simple processing units connected together to compute a more complex function. The function to be computed is generally not given explicitly. Instead, the ANN learns the function implicitly through the presentation of input patterns. The method of training separates ANNs into two classes: unsupervised, or those that learn without feedback from a trainer; and supervised, those that require a trainer to specify the correct output patterns.

3.2.1 Unsupervised Learning: The Kohonen Neural Network

The Kohonen neural network, also known as the Self-Organizing Feature Map, is a common example of an unsupervised ANN. The architecture of the self-organizing feature map consists of an input vector completely connected to a layer of neurons that form the map. The connections between the input and the map are weighted, randomly at first, and the changes in these weights are what constitute learning in the feature map. Usually the neurons in the map receive lateral connections from one another, and these connections are either excitatory or inhibitory according to the distance in the map between the connected neurons. A variety of functions are available for specifying the relationship between the distance of neurons and the degree of excitatory/inhibitory connections, though the Mexican hat function is the common choice. The learning procedure consists of presenting each input pattern to the network and determining a winning neuron for the input pattern at hand. One method for determining the winner is to choose the neuron with weights closest to the input vector as measured by Euclidean distance. The weights are updated according to either a winner-takes-all procedure in which only the winning neuron's weights are changed by moving the weight vector closer to the input in Euclidean space, or with a neighborhood function. A neighborhood function makes use of the excitatory-inhibitory lateral connections, so neurons within the excitatory neighborhood of the winning neuron have their weights moved closer to the input vector, while those neurons in the inhibitory neighborhood have their weights pushed farther from the input vector. The result of this process is to produce a topographic map of the input space.

The Kohonen self-organizing feature map has been used for a wide variety of applications, from robot arm manipulation to pattern classification (Ritter, Martinetz, and Schulten

1992; Kohonen 1995). It is the most extensively used ANN by cognitive psychologists and linguists for modeling semantic and episodic memory. Because it develops prototypes of the input classes by forming a topographic map of the input, it is thought to be both psychologically and neurally plausible.

3.2.2 Supervised Learning: The Multi-layer Perceptron

The most popular supervised learning technique, in fact the most widely used ANN technique, is learning by backpropagation of error. Backpropagation networks are Multi-Layered Perceptrons, that is, ANNs with weighted forward connections from an input layer to at least one hidden layer of neurons, as well as connections from the last hidden layer to the output layer. Layers can be fully or partially connected, though neurons in the first hidden layer usually receive the full input from the input layer. Training patterns are presented to the network, and each hidden neuron in the first hidden layer computes the sum of its weights times the inputs, and then a nonlinear squashing function is applied to this sum. This nonlinearity is important, as it allows the network to classify the input space with nonlinear decision boundaries. The output of each layer is then propagated to the next layer in the same manner as the input. Once the final output is computed, a trainer supplies the network with the desired output for that input. The error between the actual output and desired output is computed, usually by simply taking the Euclidean distance between the two, although sometimes a cross-entropy measure is used. The gradient of this error is then propagated back to the prior layer, and its weights are updated by adding a fraction of the gradient to them. The error for this layer is propagated back to its prior layer, and the procedure for updating the weights is carried out again. Perceptrons thus have the property of moving

their weights toward local minima in error space; in practice, they often find the global error minimum, although there is no guarantee that this will happen.

Backpropagation of error forms the basis of most of connectionism's successes, and backprop nets have seen commercial application. Yet, backprop nets are among the least neurally plausible ANNs, and their apparent ability to mimic human behavior is merely a corollary of their ability to mimic ANY continuous function. While admitting these facts, connectionists are generally not troubled by them; as we shall see, connectionism "properly understood" is more an abstract thesis about the nature of representation than a theory about the specific architecture for simulating minds.

In evaluating the criticisms against symbolic AI, as well as the positive case for connectionist systems based on these types of neural nets, we will use the "Proper Treatment of Connectionism" formulation of connectionism as advocated by Paul Smolensky (1988). Although a decade old now, Smolensky's position still serves as the clearest argument for how connectionism differs from symbolic AI, and why that difference is important.

3.3 The Proper Treatment of Connectionism

According to Smolensky's view, connectionism is a radical departure from symbolic AI by virtue of its novel conception of representations, and not because of a rejection of rule-following procedures. Smolensky acknowledges that the rules in a symbolic system might be implicit in that they are hard-wired, and that such implicit rules do not diminish the power of the symbolic systems. This acknowledgment disarms the common objection that symbolic AI is unrealistic because it is unreasonable to suppose that humans have explicit representations of all the rules that they make use of. Smolensky also acknowledges the

usefulness of symbolic systems for mimicking human rule-following behavior, such as scientific research, game-playing, or logical reasoning.

While symbolic systems are useful for modeling such explicit rule-following behavior, the connectionist argues that they are entirely inadequate for modeling intuitive processes, processes in which humans not only do not follow explicit rules, but also do not behave as if they are following implicit rules. Proponents of symbolic AI generally conceive of cognition as the unconscious following of implicit rules that have a syntax and semantics similar to that of explicit rules. Connectionists reject this claim for three reasons: symbolic AI systems built on these assumptions perform poorly; it is impractical to articulate rules for domains such as common sense; and a symbolic system built on this principle would contribute little to our understanding of how the brain might accomplish the task.

Smolensky argues that the reason symbolic AI systems perform poorly as models of intuitive processes is because intuitive information is not stored as symbols bearing semantic content. Symbolic systems require that representations be mapped to symbol tokens in order for rules to operate over them. Representations in connectionist systems, on the other hand, are formed as patterns of activations over sets of neurons. Representations are distributed across the weight vector of a connectionist system, and the weight vector serves to store more than one representation. Therefore, individual neurons cannot be identified with individual concepts. Their activity is subconceptual, or subsymbolic. That the two paradigms are incompatible is demonstrated by attempts to realize one type of system by modifying the other: symbolic systems used to implement subsymbolic systems (which is how most ANNs are implemented) no longer allow for a conceptual level analysis of their symbols (or so connectionists argue); subsymbolic systems used to implement symbolic sys-

tems do not suddenly gain a conceptual level analysis, as their units are still subconceptual in nature.

Connectionists also make much of the similarity between their systems and neural architecture, in spite of the significant difference between true neural systems and artificial ones. In the face of criticism from neurobiologists (see Edelman 1987), connectionists point out that their systems are higher-level approximations of neural processes, and therefore are not required to be perfect replicas of biological systems (Smolensky 1988). Presumably, connectionist systems also provide the neurobiologist with general principles of how subsymbolic systems might behave, a contribution neurobiology is in desperate need of. At the very least, connectionist systems are closer to neural architecture than symbolic systems are.

3.4 Why Connectionism Fails to Deliver: Neural Implausibility and Its Relevance

While vaguely resembling biological neural networks (BNNS), ANNs are still wildly implausible as models of real biological brains. Yet, one of their selling points has been that they are more realistic models of actual cognition than traditional symbolic systems. As will become apparent, this is true only in the most superficial of senses. The gain in neural plausibility is minuscule compared to the loss in high-level cognitive capability when moving from symbolic to connectionist systems. There is an inverse relationship between the neural plausibility of a connectionist system and the capability of the system to model high-level human cognition (e.g. natural language processing). The further one goes toward mimicking the brain, the further one moves away from mimicking human cognition.

3.4.1 Artificial Neural Networks Don’t Really Resemble Biological Neural Networks

An accurate model of a single neuron and its communication with other cells would require a “vast number of very complicated nonlinear partial differential equations” (Kohonen 1995, 53). While some of the equations governing neuronal behavior appear to be linear, the generation of nerve impulses at the axon hillock[1] of each neuron is clearly nonlinear (Pribram 1991). The neuron-like units in an artificial neural network generally compute only one equation, one that is not always nonlinear in nature. Therefore, connectionists have opted for a black box model of neuronal behavior as subsymbol production in the same way that symbolic AI has adopted a black box view of neural networks as symbol producers. Is this relevant to whether connectionism can achieve AI's goals of creating thinking machines? If Roger Penrose is correct in his assertion that thought is the result of quantum level effects, say, in the microtubules of cells, then it surely is (Penrose 1989). But one need not descend to the level of the microtubule to identify possible problems with connectionism's oversimplification of neuronal activity. An individual neuron is itself capable of as many operations as a full-blown connectionist network (Sun 1995). Much of this processing is done in parallel, although the propagation of the activation potential is sequential. Does this mean we ought to develop a sub-subsymbolic understanding of cognitive ability? This is not simply to extend the connectionist move to the neuronal level (contrary to Chalmers 1992), for here the notion of weighted connections has no obvious correlate. One might even go the opposite direction, and give single neurons representational level inter-

1. The axon hillock is the portion of the neuron where the axon first protrudes from the soma, or cell body.

pretation. If a single neuron is able to do the work of a connectionist network, then it is plausible that it can, by itself, produce representations.

But the only appropriate answer to these questions at this point in time is that no one knows for certain, although the representational interpretation of neurons is likely to be false. Connectionists argue that the most important aspect of their models is the distributed nature of the representations they form, so while the behavior of artificial neurons may diverge from that of true neurons, the global similarity between artificial and true neural systems in dynamics of processing distributed representations is nonetheless maintained (Smolensky 1988). But there is reason to believe that this claim can no longer be sustained.

For a time, neuroscientists considered the simple and complex cells in the visual cortex to be feature detectors for various geometric primitives (such as lines), and that the recognition of objects is a result of composing the outputs of these detectors along a hierarchy of geometrical complexity (Pribram 1991). As one ascends the hierarchy, neurons are responsible for higher and higher level representations by composing the inputs of the lower levels. This view has since been rejected with the discovery that the behavior of these neurons is far more complex, corresponding more closely to a Fourier analysis of the input than to simple line detection, and that visual processing is not only bottom-up, but top-down as well, with higher levels feeding back to the simple cells (Pribram 1991, 13). Little of this discovery has filtered into the connectionist program, although it undermines the neural plausibility of the connectionist methodology of using the output neurons of their systems as feature detectors, and composing primitive networks for feature detection into more complicated pattern recognizers.

The form of communication between neurons in connectionist systems diverges significantly from the neuronal communication found in biological networks. Connectionist neurons compute a transfer function of the sum of the products of their inputs and weights, and then pass the output value to the next level. Biological neurons release what were previously called 'neurotransmitters,' and are now more accurately called ‘neuromodulators’ (Pribram 1991), at the synaptic gap. These neuromodulators change the behavior of the postsynaptic neuron, in essence changing the nature of the function describing the neuron's behavior. From a connectionist point of view, this would imply that the transfer function itself has changed. But transfer functions do not change in connectionist networks. Thus the dynamics of BNNs do not resemble the dynamics of ANNs at the level of inter-neuron interaction.

Nor is there significant resemblance at the network level. There is no known analog in BNNs to the stored weights used in ANNs. The putative similarity is due to the vague notion of connection strength found in Hebb's formulation of a possible learning rule (1949). According to the Hebb rule, biological neurons are thought to have connection strengths that increase when activity in connected neurons is correlated. While the strength of excitation or inhibition between neurons might seem to fit the bill, these excitation and inhibition strengths are not fixed weights, as they are in most trained-up connectionist systems. Connectionist systems learn by adjusting weights, and store representations in their weights. Strip the weights out of a connectionist network and place them in another network with the same architecture, and the result is the same formation of representations. What might we strip out of the brain to achieve a similar result? No one knows if there even is an analog to connectionist weights, let alone what it might be. Thus, the plausibility of

ANNs lies in our ignorance of the brain. Explaining behavior by reference to demonic possession was also once plausible in this sense of 'plausible.'

Even if such an analog were to exist, the dynamics of modifying biological neural weights would differ vastly from the methods used to train the most successful ANNs. While connectionists can lay claim to a number of unsupervised learning algorithms, backpropagation networks are the most successful of the neural network architectures. This is unlikely to change for the simple reason that backprop nets are universal function approximators that can be trained in a manageable period of time. Yet, the backpropagation learning algorithm in no way resembles what is known about learning in the brain, nor is any future discovery likely to change this fact. The reason is that backpropagation learning requires that each output neuron be supplied with the correct value of its output for each member of the training set. This would mean neurons in the brain would have to know what they are supposed to learn before they learn it. Some researchers have suggested that examples of human supervised learning circumvent this difficulty. But in these cases, humans are given the true classification of the examples at the level of inter-personal communication: a teacher says 'bear' when showing a picture of a bear. Backpropagation learning requires the true output to be given at the neuronal level, which means biological neural nets must have some way of breaking up the correct output into the correct signal for each output neuron. Lacking a shred of evidence, this suggestion is sheer speculation. In any case, there is no known neural mechanism for backpropagating errors. The very notion of backpropagating errors suggests neural information transmission rather than the neuromodulation that forms the basis for inter-neuron interaction in BNNs.

The dynamics of connectionist systems therefore show no significant similarity to the dynamics of biological neural networks. What remains are the possible similarities in gross functionality and the nature of representations. If we view connectionist systems at the level of gross functionality, however, they lose their uniqueness. It is not necessary to implement a connectionist system to achieve similar functioning. One method of mimicking the behavior of ANNs is the use of probability density functions. In fact, Richard and Lippmann (1991) have proven the multi-layer perceptron, the most widely used ANN, to be an a posteriori probability approximator. MLPs trained as pattern classifiers compute the probability of an input's class given the input itself. The MLP only has the training set by which to estimate this probability. Direct methods of estimating the probability can be used in place of an MLP. Viewing the MLP as an a posteriori probability estimator, we find it is easily replaced by direct probabilistic methods. The parallelism of the MLP's computation adds nothing significant to the task of pattern recognition other than a speed-up in computation time if implemented on a machine capable of parallel processing; the MLP that does perform better than current statistical techniques on some pattern recognition tasks, the Time Delay Neural Network, is even one step further than the general MLP from neural plausibility due to its shared weights. All that is left to the connectionist is the claim that the distributed representations in connectionist networks mimic the distributed representations in BNNs.

In addition to the biological implausibility of connectionism, there is the underlying assumption in this movement that cognition is information processing that must be confronted. Although in disagreement with LOT theorists about how mental cognition occurs, they are in general agreement with their antagonists concerning what cognition is. The con-

nectionist believes that the mind/brain constructs representations of the world, stores this information, and retrieves it to process it for the sake of producing desired behavior. Of course, no connectionist network has ever done this without some form of supervision or intervention, so it is unclear why we are to believe that they are adequate models of the mind/brain. A few examples of the failure of connectionist systems to act as mind/brains, that is, as stand-alone systems creating and interpreting their own representations, will be provided to substantiate this claim. Specifically, we will investigate Herbert Simon's (Vera and Simon 1993) claim that Navlab and the neural network ALVINN that was the 'brain' of this system think, as well examine the use ANNs have been put to in cognitive psychology. But before taking on these connectionist applications, we take a closer look at what it means for a connectionist system to have distributed, 'nonsymbolic' representations.

3.5 Symbols and Representations: The Subsymbolic Fallacy

Whereas Smolensky's connectionism identifies the distributed nature of representations in connectionist systems to be what distinguishes itself from symbolic AI, some connectionists (Churchland and Sejnowski 1989) have pointed to the analog nature of the computation carried out in the connectionist processing units as indicative of these networks greater plausibility over symbolic systems. Notions such as 'nonsymbolic computation' have been invented to distinguish the operations of artificial neural networks from the symbolic computation of traditional AI systems (Stufflebeam 1998). Here, 'symbolic' is meant in the syntactic sense (Smolensky 1988). Yet, when connectionists attempt to defend this position, they revert to the semantic sense of symbol, defining a symbol as a token that can arbitrarily designate, whereas analog signals lack this property (Harnad 1990; Touretzky and

Pomerleau 1994). A second, syntactic property is also referenced when distinguishing between analog signals and symbols, namely that the latter are recursively composable via rules.

These distinctions are meant to answer a simple objection to the symbol-nonsymbol distinction. The floating-point operations found in artificial neurons can be, and often are (for the sake of speed), translated into integer operations. In fact, all neural networks implemented on digital computers are digital operations; vectors of 1s and 0s represent all of the floating-point operations in such neural networks. Thus, this supposed distinction appears to hold only if connectionist systems are implemented on analog devices. However, if the composability constraint applies to connectionist systems implemented on digital computers as well, then these systems are also nonsymbolic.

The argument for the noncomposability of analog signals seems to rest, however, on the assumption that composition requires explicit rules. Thus Touretzky and Pomerleau argue:

> Analog numerical representations violate both of the above requirements for a physical symbol system. First, they lack the capability of arbitrary designation, because they are constrained to maintain an analogical relationship to the thing they represent. Consider, for example, the homomorphism between voltage levels in a thermostat and temperature values in a room. Each voltage level is a distinct analog pattern, but not an arbitrary one. And relationships between analog patterns (e.g., that voltage level x denotes a warmer room state than voltage level y) are not defined explicitly by symbol structures or symbol manipulation rules. Rather, they are predetermined by the causal structure of the thermostat sensor-the source of the analogy between voltage and temperature. (Touretzky and Pomerleau 1994, 346)

This contradicts Smolensky's view that implicit rules may be hardwired into a system (be part of its causal structure) that does not possess explicit rules, yet the system still be considered a symbol system. Furthermore, Touretzky and Pomerleau argue specifically that it is analog systems that lack the ability to combine or compose signals as if they were symbols:

> Analog representations also appear to lack combinatorial power. One may speculate about uses for fractal structures or chaotic attractors in knowledge representation schemes, but it is not evident to us how to achieve useful compositionality in a low dimensional, analog numerical system. (Touretzky and Pomerleau 1994, 346)

What is missing from connectionist accounts is a clear statement whether analog systems are truly special, and therefore a renunciation of all digital implementations of connectionist systems as instances of a new paradigm in cognitive science.

The symbol-subsymbol distinction is not the same as the symbol-nonsymbol distinction, relying on different notions of 'symbol'. The latter is a restatement of the analog-digital distinction, viewing computation over analog signals to be somehow different in kind from computation over symbols. The former view is that symbols are not merely syntactic items, but embody representations, and therefore already bear semantics. On the former view, connectionist systems differ substantively from symbolic systems because the representations are distributed, hence the operations of individual units cannot themselves admit of a conceptual level analysis. Thus binary neural networks count as connectionist systems, whereas according to the analog-digital distinction they do not.

So why does it matter whether representations arise from subsymbolic processes rather than being directly mapped to/from symbols? There are a number of reasons. Presumably, this is how it happens in real brains (Smolensky 1988; Churchland 1995). More specifically, the connectionist argues that brains store and retrieve representations in ways incompatible with symbolic AI systems, but quite like connectionist systems. To establish this point, connectionists must both provide evidence that brains store and retrieve representations as connectionist models describe and construct working connectionist systems according to this model. I have already presented evidence against the first of these efforts.

I will now show that connectionist models themselves do not live up to the connectionist paradigm.

The subsymbolic hypothesis, as stated by Smolensky, is that connectionist systems do not "admit a complete, formal and precise conceptual-level description" (Smolensky 1988, 12). Even symbol systems used to implement subsymbolic systems no longer admit of a conceptual-level description. But this is to overstate the conceptual opaqueness of connectionist systems; not all connectionist systems are opaque. For example, an ANN that uses a hidden layer of neurons equal in size to the data set it is meant to classify would end up training one neuron to react strongly for each member of the data set—in essence, creating a look-up table. The hidden neurons in such a network admit of a clear conceptual level description: they are representing exemplars. Furthermore, the output neurons then admit of a conceptual level analysis: each output neuron, representing a class of outputs, merely classifies a portion of the look-up table.

This kind of network is an extreme example, one that would never be used even if a look-up table were desirable. It is equivalent to a localist network, the type of network that uses neurons to represent concepts and produce output that consists of weighted concepts. In distributed networks, the sizes of the hidden layers of ANNs are always smaller than the size of the training set, for it is desired that the hidden layers generalize on the data set and not merely encode for particular instances. Hidden neurons become feature detectors. When the network is simple enough, the particular feature being detected by a given neuron can often be discovered, as has been done with the neural network ALVINN's[1] hidden neurons (Pomerleau 1992). An individual weight might not be able to given a conceptual-level description, but there is no reason that feature detecting hidden neurons cannot. And when

the hidden neurons of a network cannot be analyzed, this is largely due to the complexity of the network, and not any magical opacity that necessarily accompanies distributed representations. Better analytical tools may eventually clear away this opacity; connectionists certainly have not given any reason to believe it impossible. Even the contribution of a single weight to the feature detection carried out by a neuron may yield to analysis-in fact, there are instances where it does. Simple character recognition networks, such as a single-layer perceptron trained via the Least Means Squares algorithm, yield weights that act as a fuzzy template to be matched against. A weight is used to detect a particular feature in its corresponding area in the input space. Analysis of weight contributions to more complex networks is not yet possible.

If the function of a weight or hidden neuron can be discovered, then this weight or neuron can serve as a symbol for a system that reads off its value and draws conclusions as to what features are present. This is already done with the output neurons of connectionist systems. In fact, the outputs of connectionist systems are generally treated as symbols. One method for output-to-symbol translation is simply assigning a label/symbol to each output neuron, and an input receives the label of the output neuron with the strongest activation on that input. Therefore, such connectionist systems always admit of a conceptual level analysis at some point in their operation. In fact, a general rule seems to be that the more symbolic post-processing, the better the performance. Examples of this abound in handwriting

1. ALVINN, the Autonomous Land Vehicle In a Neural Network, is the creation of Dean Pomerleau and Todd Jochem, researchers at Carnegie Mellon University. It is a multi-layer perceptron that was trained by means of backpropagation to autonomously drive a vehicle on a variety of road types, including both single and two-laned highways. ALVINN's success at this task is considered a major achievement for neural networks, having driven a vehicle 96% of a 70 mile road trip, with the human supervisor taking over only for passing and on/off ramp access (Pomerleau 1992). Recent developments have even allowed for passing (Jochem, Pomerleau, and Touretzky 1995).

and speech recognition systems (see Waibel et al. 1989). Working connectionist systems are hybrid systems, combining a parallel processor used to approximate functions and symbolic processors used to apply the outputs to the domain at hand.

So there are two problems with Smolensky's account. Smolensky acknowledges that a complete, formal and precise description can be given of the elements of connectionist systems, those components that operate at the subconceptual level. However, as has been shown with ALVINN, these elements also permit of a conceptual level analysis. It would seem then that connectionist systems are not subconceptual processors. A similar confusion arises in Touretzky and Pomerleau's account. On the one hand, Touretzky and Pomerleau argue that the hidden neuron activity patterns have a conceptual level interpretation as meaningful as that found in a symbol system:

> [ALVINN's] patterns are not arbitrarily-shaped symbols, and they are not combinatorial. Its hidden unit feature detectors are tuned filters. Their activation denotes the presence of visual features in the view ahead of the vehicle, and analysis of their input weights and response properties reveals that they are responding to objects such as road edges and lane stripes... Greeno and Moore (1993, 54) claim there is no semantic interpretation of these hidden unit activity patterns. But their response, albeit analog in nature, is just as meaningful as the discrete output LANE-STRIPE-DETECTED. (Touretzky and Pomerleau 1994, 349)

Then Touretzky and Pomerleau go on to endorse the PTC view of hidden unit activity:

> We cannot assign concise meanings to elements at this lower level. Whatever phrase we choose would correspond to a concept structure, not a symbol. Except for those symbols that serve as names for familiar concepts, primitive symbol tokens would appear to be indescribable. A comparable notion exists in the connectionist literature, where "subsymbolic representations" are postulated whose components are "microfeatures." The subsymbolic level is really a subconceptual level . . . (Touretzky and Pomerleau 1994, 349)

Either we must conclude that symbols such as LANE-STRIPE-DETECTED are really subconceptual, or that the hidden layers of connectionist systems are capable of a conceptual

interpretation similar to that of symbols. It is not clear from (Touretzky and Pomerleau 1994) which it is to be.

Second, the claim that the network level does not admit of a conceptual level analysis is questionable given that the elements of the network may be assigned conceptual-level interpretations, and that the symbolic interpretation of network outputs is standard practice among connectionist researchers. The elementary concepts only seem to disappear in the network because of the nonlinearities of typical connectionist systems. Even if the hidden neurons act as feature detectors, this precise interpretation disappears in their interaction with the output neurons. Or at least it seems to. The unspoken assumption present in such claims is that a rule cannot be a nonlinear function on concepts; otherwise, connectionist systems would simply be rule-based systems. But once we have a conceptual level analysis of the hidden neurons, their connection to the output neurons is primarily linear. The usual squashing function (tanh or the logistic function) does nothing to change the ordering of the output neurons' strengths: the neuron with the highest activation before the squashing function is still has the highest activation after application of the squashing function. So the introduction of this nonlinearity does not prevent a rule-based interpretation of the hidden-to-output neuron relationship, and therefore does not prevent a conceptual level analysis of network behavior. Finally, if the network's behavior is too imprecise or informal, how is it that the network can fit in a with a standard symbol system? Is something lost when we assign a symbolic interpretation to the network's behavior?

3.6 Why the Causal Efficacy of Representational Structure Is Not a Form of Exceptionalism

David Chalmers has argued that connectionist systems possess an internalist semantics lacking in traditional symbol systems (Chalmers 1992). The internalist view of semantics is that meaning is determined by the internal properties of mental events and not by their extension to the external world. In Chalmers's view (a view he has somewhat backed away from), this exceptional aspect of connectionist systems insulates them from Searle's contention that programs cannot have semantics.

According to Chalmers, connectionist systems possess internalist semantics because the structure of their representations is in itself causally efficacious. The structure of a connectionist representation is the pattern of activity over the nodes of the network. Each of these nodes plays a causal role in producing the representation. Unlike in symbol systems, where any arbitrary configuration of bits could be used for a particular atomic symbol, changing the pattern of activity over the nodes in a connectionist system changes the representation. In Chalmers's view, the pattern of activity in a connectionist system causally determines the nature of the representation. A similar contention is explicit in the sections from (Touretzky and Pomerleau 1994) quote above.

Searle (1980) contends in his Chinese Room argument that syntactic elements do not bear semantics. This suits Chalmers's reading of connectionism, because the level of syntax in a connectionist system, the subconceptual, is not meant to bear semantics. It is at the network or representational level that semantics emerges. At this level, representations have an intrinsic structure built up out of the subconceptual level, and it is this intrinsic structure which determines their meanings. It is for this reason that Chalmers concluded

that connectionist systems are not vulnerable to the Chinese Room argument as symbolic systems might be (if we assume that Searle's intuitions are correct). The vulnerability of symbolic systems is that their symbol tokens bear no intrinsic meaning:

> For all intents and purposes, a computational token is a featureless chunk, coming only with an arbitrary label that serves only to distinguish it from other computational tokens. Nothing intrinsic to the ELEPHANT token makes it any more closely related to the elephant concept than to the apple concept. (Chalmers 1992, 13)

Being embedded in a system, that is, having relations to other tokens, does not suddenly imbue an otherwise meaningless token with semantic content. A Searlean would view this as mere syntactic linking of meaningless tokens.

If we accept Chalmers's account of connectionism, namely that subconceptual elements do not bear semantic content, and that the semantics of connectionist representations derive, if at all, from their intrinsic structure, we must conclude that connectionism is at least as vulnerable to the Chinese Room argument as traditional symbolic systems. Chalmers's characterizations of connectionist systems mean that the derivation of their representations' semantics cannot be from any semantic content at the subconceptual level. The 'causal efficacy' of the subconceptual level must do all the work. But Chalmers has mischaracterized the nature of connectionist representations.

First, it is not entirely true that the activation patterns of a trained-up network cannot be altered without altering the representation. This is assuming a 1-to-1 correspondence between patterns and representations, a correspondence that does not always hold. For example, Kohonen networks yield neighborhoods of neurons whose activation yield the same representation.

Second, it is deceptive to characterize the relation between gross network behavior and neuron behavior as a relation between representations and subconcepts. The relation is one of output to the function that yields the output. It is a purely syntactic relationship. The output of the network is as meaningless as any symbol token in a traditional symbol system. Its rich intrinsic structure is similar to the rich intrinsic structure of a sentence: if you change the syntactic elements making up the sentence, you change the syntax of the sentence itself. Furthermore, if the syntactic elements making up the sentence do not have semantic content, the sentence itself does not.

So even if the output of a connectionist system is not arbitrary because it is determined by the nature of the network, its semantic interpretation is arbitrary, assuming we are considering internalist semantics only. A different configuration of activation patterns could have been formed on the same input if we were to train up a second network. What is thought to give connectionist networks their semantic content is not merely the constitution of the network, but also the semantic content of the input. If we were to train a network on images of animals, and output neuron 1 fired more strongly than others for images of elephants, then output neuron 1's activity (and its underlying cause within the network) might be thought of as a representation of elephants. Take away this external semantic interpretation of the input, and there is no reason to consider the activity of the network to also have a semantic interpretation. The network is merely computing a function on the input. This function could be computed with or without the underlying neural network; the network is useful primarily in cases where we cannot analytically discover the function to be computed.

Thus, connectionist systems do not have a special internalist semanticity that symbol systems do not bear. In some cases, they have a richer syntactic structure. Merely complicating the syntactic structure of a representation does not thwart Searle's Chinese Room argument, if the argument has any bite at all.

3.7 Connectionist Systems Don't Really Possess External Semantics

I have previously argued that, contrary to connectionist claims, the 'subconceptual' aspects of connectionist systems can be given external semantic interpretations, that is, an external observer can identify correspondences between neuron activity and features in the network's input. I have also argued that the activity patterns of connectionist networks do not bear semantic content by virtue of an internal semantics. Thus, if a connectionist system does have semantics, it is due to its external connections to the world. In this section, I argue that connectionist systems are not exceptional by virtue of external connections: If symbolic systems do not have semantics by virtue of external connections, then connectionist systems do not either. This will be a continuing theme throughout this chapter, one which I will return to when examining actual connectionist networks.

The case for connectionist exceptionalism with regard to externalist semantics derives from the ability of connectionist systems to automatically detect features and generalize from their given data sets. As has been noted, the hidden neurons of ALVINN were discovered to be detecting sets of parallel lines of varying orientations in images. Having been trained on a representative sample (more on this in later sections) of roads, ALVINN was able to generalize from the relation between sets of parallel lines seen to the appropriate

steering curvature when presented with novel road scenes. Thus, ALVINN, and connectionist systems in general, form their representations by discovering in the input from the external world actual relations between properties of the input. While a supervisor may provide a connectionist system with the correct value for its output during training, the supervisor does not tell the system what features it is to discover in the input, nor how to generalize from these features. It would seem that the homunculus fallacy does not plague connectionist systems, at least regarding the representations formed by the hidden neurons.

The ability of connectionist systems to automatically form representations of data and to generalize to novel instances, however, does not imply anything about the semanticity of these representations. A connectionist system will do this regardless of whether the input signal bears semantics. The representations formed in such a case do not suddenly gain semantics; they are generalizations about meaningless signals.

Even if we were to concede semanticity to connectionist representations, this would in no way save the connectionist enterprise. As is conceded by Touretzky and Pomerleau (1994), the connectionist system is able to do little with the representations it forms: it simply is mapping input to output. Touretzky and Pomerleau theorize that a symbol system would have to be added on top of any connectionist system to get the full range of representational behavior found in humans, because true (subsymbolic) connectionist systems are incapable of the manipulation of (representational) symbols that humans do with ease.

3.8 Why Connectionism Is Not an Implementation of Heideggerian Principles

Hubert and Stuart Dreyfus have argued that connectionism is a novel development in cognitive science, because it represents a shift away from the Cartesian, analytical AI that sought to reverse engineer organisms in order to replicate their functioning, and is a move toward a "synthetic" or holistic approach to understanding consciousness. By focusing on building physical symbol systems, early AI researchers were trying to isolate the atomic elements of thought so that they could then build a mind piece-by-piece. As Dreyfus and Dreyfus state it:

> AI can be thought of as the attempt to find the primitive elements and logical relations in the subject (man or computer) that mirror the primitive objects and their relations that make up the world. Newell and Simon's physical-symbol system hypothesis in effect turns the Wittgensteinian vision (which is itself the culmination of the classical rationalist philosophical tradition) into an empirical claim and bases a research-programme on it. (Dreyfus and Dreyfus 1988, 311)

At the same time, a movement to model the brain's functioning, led by Frank Rosenblatt, had begun the first perceptron research. Dreyfus and Dreyfus see this as a branchpoint in the history of AI, when the choice was between an analytic approach (what is now referred to as traditional or symbolic AI) and a synthetic or holistic approach. Due partly to the accidents of governmental funding, and partly to the devastating critique of perceptrons written by Minsky and Papert (1969), the former won the day and held sway for decades since.

For Dreyfus and Dreyfus, this was not merely a clash between computer scientists trying to get funds for their research. It was a battle between heirs of disparate philosophical traditions. On the one side was symbolic AI as heir to the early Wittgenstein and rationalists, on the other was the nascent form of connectionism as heir to the later Wittgenstein

and early Heidegger. According to Dreyfus and Dreyfus, the later Wittgenstein and the early Heidegger had severe conclusions for the plausibility of symbolic AI:

> Both these thinkers had called into question the very tradition on which symbolic information-processing was based. Both were holists, both were struck by the importance of everyday practices, and both held that one could not have a theory of the everyday world. (Dreyfus and Dreyfus 1988, 320)

For the Dreyfuses, the foundering of symbolic AI on the difficulties faced in constructing adequate representations is evidence that the later Wittgenstein and early Heidegger were correct. Similarly, they consider the move in cognitive science toward connectionism to be another proof that these two philosophers were correct, because connectionism offers a holistic account of representation and thinking. Specifically, connectionism views information as stored ubiquitously in the network rather than at specific nodes, meaning information must be 'evoked' rather than retrieved. Connectionism does not require that the constituent parts of the network be analyzable in terms of features recognizable to a human, and it does not require that the researcher specify the rules by which the network may generalize about inputs it has not been trained on.

Still, Dreyfus and Dreyfus view connectionism as not quite holistic enough. Connectionists, in their opinion, still retain the deplorable habit of trying to analyze what their network is doing, not realizing what Wittgenstein and Heidegger had to say about such a practice. Nonetheless, connectionism represents to them a substantial move toward the holistic approaches of these two philosophers, and a repudiation of the analytic approaches of the symbolic AI researchers.

Dreyfus and Dreyfus have misconstrued the implications of connectionism. If a move toward holism would validate the later Wittgenstein and early Heidegger, and a move in the

opposite direction undermine their philosophies, then connectionism represents the most promising avenue for refuting the later Wittgenstein and early Heidegger. Connectionist approaches are indeed often used in domains where symbolic AI has had little success, but merely to achieve the success symbolic AI sought. One such area is pattern recognition. The best function for separating patterns in a way similar to how humans categorize is generally unknown. If an ANN can learn the relationship between inputs and outputs, an analysis of what the ANN is doing often yields the function being sought. The ANN can then be scrapped if it is not the most efficient method for applying this now known function. One example of this process of learning opaque functions through neural nets is the Navlab system at Carnegie Mellon. Navlab is an autonomous driving system, with platforms of a van and military HumVees. Originally, the neural network ALVINN learned the functional relationship between road inputs and driver reactions. Once it had done this, Navlab researchers analyzed ALVINN's hidden nodes and discovered the features it was keying on. RALPH, a non-neural network system that applies the function discovered by ALVINN, has since been able to replace ALVINN (Pomerleau 1995).

Dreyfus and Dreyfus are correct that ANNs do not require an analysis of their hidden nodes, but analyzing hidden nodes is the direction connectionism is heading. While researchers generally do not use backpropagation networks to model the brain, the neural implausibility of these networks having been evident since their inception. Instead, these ANNs yield solutions to problems that otherwise cannot be discovered by means of the analytic tools at hand, and these solutions are themselves analyzed. The point of a hidden layer of neurons is not to reduce the analyzability of a solution. Quite the contrary, it is to increase its analyzability. The hidden layers of MLPs and RBFs serve the purpose of casting the

input space into a higher-dimensional space to improve the separability of the patterns. Analyzing the activity of hidden neurons then allows the researcher to determine what features of the input the network is keying on in making its classification. So while a back-propagation network might not be neurally plausible, it can nonetheless provide researchers with insights into what BNNs are doing to the inputs they receive. So if connectionism does not require that the hidden nodes be analyzed, it does encourage this. In any case, holistic accounts tend to claim that the processes underlying, say, commonsense cannot be analyzed, not that if we close our eyes to the possibility of such analysis, we then have achieved holism.

An example of the analytical usefulness of ANNs is the work on object position location in monkey brains accomplished by Richard Anderson's lab at MIT in conjunction with the ANN researcher David Zipser (Barinaga 1990). By training an ANN to locate an object on the basis of retina and eye position, and then analyzing the hidden nodes of the network, these researchers were able to make predictions about how the monkey's biological nets would respond to tasks similar to those put to the ANN. When the two showed similar activity, the researchers could then formulate theories as to what neurons were involved in the location task in the monkeys' nets. The ANN does not prove anything about how the brain functions, but does suggest possible routes for analyzing its activity. Rather than providing a holistic approach to modeling cognition, ANNs are tools for breaking through the apparent holism of the mind.

3.9 Why Connectionism Doesn't Really Aid Cognitive Psychology

Over the past decade, many cognitive psychologists have turned to connectionism to provide the computational tools for understanding the mind. Because cognitive psychologists regard the neural plausibility of these models as secondary to whether the models capture aspects of overt cognitive behavior, the backpropagation network finds widespread use alongside more "plausible" nets such as the SOFM. The connectionist models devised by cognitive psychologists are part of an effort to model cognitive behavior through piecewise simulation of mental computation: the modeler will pick out a specific cognitive function, such as categorization, and devise a system specifically to simulate human performance on this function. The drawback of these models is that they capture only a small piece of cognitive behavior without developing how these processes would fit into a complete system. In what follows, we will show that these piecewise models cannot be fit into a complete cognitive system, because they rely on human interaction for their behavior to be meaningful to such a system. As a demonstration of this point, we will examine one such model, namely Philippe Schyns concept-acquisition network (Schyns 1991), a model that makes use of a neural net that is often, and erroneously, thought to avoid the need for a human interpreter: the SOFM.

3.9.1 Schyns' Work on the Nature of Concept Acquisition as Prototype Formation

Cognitive psychologists have recently debated the merits of two models of concept acquisition and categorization: prototype theory and exemplar theory. The former holds that categories are stored as statistical means, or prototypes, of the range of inputs falling within

each category (Anderson 1991). New inputs are categorized according to their distance in feature space from the stored prototypes, and the new input is assigned the class of the closest prototype. Exemplar theory holds that the mind stores a multitude of instances or exemplars, rather than one prototype, for each class, and categorization is a competition between sets of exemplars (Nosofsky 1986). The newly categorized input can then itself serve as an additional exemplar, with the number of exemplars being constrained by memory limitations. To further establish the plausibility of prototype theory, Philippe Schyns has presented an ANN model of concept acquisition that is compatible with this theory, using a SOFM neural network and a self-supervised backpropagation (SSB) network.

Schyns separates the concept-acquisition problem into two tasks: categorization and naming. The SOFM is used to categorize the inputs, and then the SSB-net assigns the inputs their appropriate labels. Inputs to the SOFM are idealized digital representations of real-world objects. These representations are black-white drawings in 10x10 arrays of pixels, with each one derived from a prototype drawing. Schyns uses prototypes for the categories 'dog', 'cat', and 'bird.'

Inputs are generated by randomly altering 1-10 of the pixels lying around the contour of the shape, either turning them "on" (coloring them black) or "off" (coloring them white) depending on their current state. To prevent any category from being defined by singly necessary and jointly sufficient features, the decision to turn a pixel "on" or "off" was made probabilistically (changing the state of a selected pixel occurred with a probability of 0.75).

The network of 10x10 neurons was then trained on a large sample of these individual instances. As is the case with SOFMs, the network formed a topographical map of the input space. The output of this SOFM was then sent to the SSB net that had been trained as a pat-

tern associator. A pattern associator is simply a network that associates an input pattern with an output pattern. If the associated output is the same as the input, then the network is an autoassociator, the type of associator used by Schyns. The SSB net is therefore simply trained to take as input the category indicated by the SOFM and return the label for that category. Schyns calls this the naming of the category.

Schyns's network shows a variety of the characteristics of categorization predicted by prototype theory. Prototype theory predicts that the closer the input is to the stored prototype, the faster the categorization process will be. Schyns's network demonstrates just such a prototype effect. Also, when Schyns introduced a fourth category, that of 'wolf,' and had the network relearn the categories, the network classified wolf-inputs as closer to 'dog' than to 'bird' or 'cat.' This is to be expected given that the prototypes of 'wolf' and 'dog' share more features than 'wolf' and 'bird,' or 'wolf' and 'cat.' Furthermore, Schyns argues:

> Much like children's overextension errors and their eventual corrections, in this model, when the new category wolf is learned, conceptual interpretations of its exemplars are labeled as "dog." At this stage, the lexical item "dog" is overloaded. Its set of referents is overgeneralized because it comprehends not only the dogs, but also the wolves. However, the conceptual map interprets wolves distinctly from the dogs, so a new label can be associated to the wolf area of the map. When this new category term is learned, the initially overextended category name "dog" narrows down to the correct set of referents, whereas "wolf" refers to the remaining referents. (Schyns 1991, 488)

A further experiment by Schyns was designed to show how hierarchical networks can handle subcategorization in a way consistent with prototype theory and the psychological data concerning humans' ability to subcategorize. Schyns trained two separate SOFMs, one presented with more instances from the set of dog, the other with more instances from the set of bird. The naming network for the first SOFM is trained on all the possible names (within a restricted set for the experiment) of types of dogs, as well all the possible names (again, a restricted set) for types of birds. The naming network for the second SOFM only

is trained on a few of the bird subcategory names while still receiving full training for dog type names. Each main category has four subcategories in this experiment. Early in the training, each SOFM organized in such a manner as to discriminate between categories, with the majority of the neurons being assigned to each SOFM's dominant training class, but showed little discrimination among subcategories. According to Schyns, this corresponds to data indicating that medium-level concepts emerge before basic-level concepts in infants (Schyns 1991). At the end of the training, the subcategories had their own weight vectors. The naming network for the SOFM that was expert in the bird category had only one subcategory label, and so showed a quicker response time for naming the output of the SOFM than the naming network for the dog expert. Each expert showed more discrimination within the category it was expert in than in its novice category. Schyns argues that this is consistent with Rosch's hypothesis that experts have more low-level concepts than novices do (Rosch et al. 1976).

Schyns has developed a model which, he argues, correlates well with data about concept-acquisition in humans. This model, however, bears the weaknesses of all connectionist models for cognitive psychology, weaknesses I will now examine.

3.9.2 The Invisible Hand of the Researcher: A Critique of Schyns' Model

That a SOFM would bear resemblances to the categorization behavior predicted by prototype theory should surprise no one. The SOFM acts as a clusterer, with neurons in its map corresponding to cluster centers. A similar result could be achieved with a non-neural technique, such as k-means clustering. Adding new categories can also be handled by variations of k-means that split clusters according to a metric such as intra-cluster scatter. So the use

of an ANN has not increased the explanatory value of Schyns's model. Furthermore, a similar SOFM could be used to validate exemplar theory. Simply increase the number of neurons in the map to at least the size of the input set, and the SOFM is likely to assign one neuron to each member of the input. If it does not, simply increase the number of neurons. Clustering might still occur, but then again, it might not. So how does the SOFM validate prototype theory over exemplar theory?

The use of a SSB network for naming is similarly spurious. The SSB autoassociator learns through supervision-so why not just hand label the output of the SOFM? The name 'self- supervised backpropagation' is deceptive, in that it might lead one to think that a form of self-organization is occurring during training. However, the network is merely hard-wired to supervise itself. In other words, the designer of the net predetermines the true label of the input to be supplied to the output neurons. (But what if our genes do a similar predetermination? Yes, what if, but first give us reason to believe this.)

Schyns's accomplishment is to use a prototype-extracting neural network to extract prototypes. The psychological data that supports prototype theory will undoubtedly resemble the behavior of the network in some sense. But even if his network perfectly mimics the human ability to categorize, there is no reason to believe that he has discovered the mechanism underlying the human activity. The reason is that such a network could never do what it does without human intervention. Therefore, Schyns, like all other computationalists, has not given an account of how such a system might fit into a complete cognitive system without a human puppeteer still pulling its strings.

The human puppeteer emerges first in the presentation of the input. Humans are capable of developing ad hoc categories: e.g., all chess pieces except the kings and all oranges

together form the ad hoc category 'phlegs.' Schyns's input, however, is structured such that there are inherent similarities within classes of inputs, and these similarities are what the SOFM is keying on. The SOFM is incapable of forming clusters of data that lack an inherent similarity. So at best, Schyns is modeling categorization of natural kinds. As we shall show later regarding ALVINN, special preparation of the input is a hallmark of connectionism (and often the essential part of getting a connectionist net to do what one wants).

Human intervention continues with the interpretation of the SOFM's responses to each input. Once the SOFM has concluded its training phase, the network can then be used to classify data. When presented with an input sample, the network produces a pattern of activation across its map, with some neurons responding strongly, some weakly and some not at all. The researcher must then decide upon a threshold for what degree of neuronal activity counts as a category response by the network. An alternative is to count all activity as a desired response, and to consider the degrees of response to be degrees of certainty or degrees of membership in a class. While desirable for capturing certain 'fuzzy' aspects of human thinking, it is not a general solution: a poodle does not belong to the class of dogs with certainty 0.9 and to the class of birds with certainty 0.1. What cognitive modelers need is a mechanism for autonomously deciding what class an input belongs to-the very thing they are supposed to be modeling!

3.9.3 How Connectionism Hinders Cognitive Psychology

Connectionism provides the cognitive modeler with powerful tools (too powerful compared to actual neurons) for mimicking cognitive functions, yet connectionist systems show little of the human capacity for perception and cognition. No connectionist system even

approximates human visual capabilities, let alone reasoning skills. How do we account for this disparity?

One reason that a connectionist might give is that we do not fully know what functions the human mind uses, and so we cannot construct adequate models, connectionist or otherwise, of these unknown functions. But a function can be defined by its input-output mappings, and although cognitive psychologists do not have a complete set of such mappings for any cognitive function, they do have extensive data sets. MLPs trained via backpropagation should be able to approximate cognitive functions given these data sets. Yet they cannot.

Two possible reasons for this failure are that the brain might not be executing computational functions in the first place and that the functions that any particular biological neural network carries out might be dynamically changing. The first possibility rules out connectionist systems in principle. The second rules out connectionism in its current incarnation, and likely in any incarnation, though any such bet is risky.

To model dynamically changing functions requires a system to be capable of both executing each of the functions the system oscillates between and also executing the function that switches between these functions. This presents several obstacles to an ANN. First, the typical ANN is incapable of learning more than one distinct function, one reason being interference effects (Ratcliff 1990). Once a function is learned, such as by backpropagation, it is embedded in the system in the form of the adjusted weights. Second, collapsing the two types of function into one function means creating a new, discontinuous function. ANNs that act as universal function approximators are only able to approximate continuous functions to an arbitrary degree. Switching between functions would require an explicit rule,

which would violate a connectionist constraint. Third, there is little by the way of connectionist theory on how feedback from other systems should be incorporated in a given connectionist system. Presumably, volition is a form of feedback that changes what the BNN is currently computing (for example, the decision to create the ad hoc category *oranges and chess pieces* when there is no already existing category). There is no plausible connectionist model of a feedback system like volition.

While connectionists eschew most extreme forms of modular models, the systems they build are nonetheless specialized, in that they learn one specific function, and isolated from other systems, as they cannot handle dynamic feedback; in other words, highly modular. In their current form, these systems are a dead-end for cognitive psychologists looking developing global theories of cognition. Connectionists have as yet provided little reason to believe that they can be modified to suit the purposes of a comprehensive cognitive psychology.

3.9.4 Closing a Gap: The Status of Connectionist NLP

If connectionist systems are to be considered the cure for the computationalist's problems, they must not only be capable of developing representations bearing semantics, but must perform at least as well as symbolic systems in manipulating these representations. Natural language processing (NLP) is perhaps the paradigm of the human capacity for manipulating representations. And over the past decade, the state of the art in connectionist NLP has progressed in nearly every major domain of natural language processing: dynamic binding of variables to values (binding names to content), functional bindings and structured pattern matching (e.g., unification), encoding and applying recursive structures, the formation of

lexical, semantic and episodic memories, and symbol grounding (Dyer 1995). Yet, in every one of these domains, connectionist systems lag behind the best symbolic AI systems. What connectionists have achieved is to prove that their systems are capable of basic symbol processing (see Touretzky and Hinton 1988). The price of this achievement has been to further sacrifice neural plausibility, as well as adherence to connectionist principles such as exclusive reliance on local computations (no global rules) and distributed representations (no symbols in the guts of the system). Thus, the price of approximating the performance of symbolic systems in NLP has been to surrender some of the purported exceptionalism of connectionism.

Efforts at symbol grounding illustrate connectionist NLP's reliance on traditional AI techniques, and thus the inheritance of the latter's limitations. Connectionist symbol grounding is restricted to toy environments, where the complexity of the environment is several orders of magnitude less than that of real-world environments. Although symbolic AI shows little promise in the realm of perception, symbolic inputs to the connectionist symbol grounder are used to represent the inputs from perceptual systems. This simplifying assumption has the effect of pre-grounding the symbols: the question at hand is how symbols are grounded, but the inputs from the simulated perceptual system are both symbolic and assigned a meaningful interpretation by the system designer. For example, in the DETE system developed by Dyer and his colleagues, a circular blob is meant to be a ball. The system is merely taught to associate the verbal input (also symbolic) with the extracted invariant shape of the blob, as well as with its color and the simple action/events in which it participates (Dyer 1995). The DETE system is a symbol associator, not a symbol grounder; developing a symbol associator that "like a child, . . . must be taught incremen-

tally" (Dyer 1995, 414) does not solve the symbol grounding problem. Unlike a child, the system cannot develop prelinguistic categorizations of objects. According to computationalism, these prelinguistic categorizations are also symbols in the head that are somehow grounded, though clearly not by binding an explicit name to a 'representation.' Thus the symbol grounding problem is not the problem of how meaning is associated with a symbol, but what meaning is such that it can be associated with a symbol, if indeed that is what takes place. Assuming that symbol grounding is an association between symbols merely begs the question concerning what it is to ground a symbol.

Another connectionist method of NLP that does not simply mimic symbolic AI approaches is Riisto Miikkulainen's system for reading and answering questions about scripts, the DIstributed SCript processing and Episodic memoRy Network (DISCERN) (Miikkulainen 1993, 2000). The system consists of 4 subsystems, including a parser, generator, question answerer and memory subsystem. Each subsystem has two modules. The parser consists of a sentence and story parser, the generator of a sentence and story generator, the question answerer of a cue former and answer producer, and the memory of a lexicon and episodic memory. The sentence parser reads the input sentences of a script one word at a time, and the story parser combines sequences of sentences into internal representations of the story to be stored in the episodic memory. The sentence generator generates words for individual sentences, and the story generator generates the sentences of a paraphrase of an internally represented story. The cue former creates cue patterns for memory from questions parsed by the sentence parser, and the answer producer generates answers given the question and the story. The lexicon stores representations of words as vectors of continuous values from 0.0 to 1.0. The episodic memory is a hierarchical struc-

ture of ANN feature maps, similar to the SOFM. The parsing and generating is achieved by means of a Forming Global Representations with Extended backPropagation network (FGREP), which is a variation of Elman's (1991) Simple Recurrent Network, a powerful but neurally implausible ANN.

Miikkulainen (1997) emphasizes the point that his system is not merely a connectionist implementation of symbolic NLP. He points to 4 differences between the symbolic and subsymbolic (connectionist) approaches, several of which we have seen before. These are:

1. Subsymbolic representations are continuously valued.
2. Concepts have similar representations in subsymbolic systems; the 'marker' for a concept is not arbitrary.
3. Subsymbolic representations are holographic: any part of a representation can be used to reconstruct the whole representation.
4. Separate pieces of information are superimposed on the same finite hardware.

The final distinction, known as the superposition of information, is not to be confused with the distributive nature of representations. It will be of prominence in later chapters.

Although Miikkulainen's system has shown significant success in processing scripts, he admits that it has a long way to go before it approaches the NLP capabilities of symbolic systems (2000). One difficulty facing subsymbolic approaches is how to learn abstractions, i.e., correlations that do not appear in the raw data.

Advances in connectionist NLP, including Miikkulainen's, serve to further establish the (empirical) rule of the inverse relationship between the neural and cognitive plausibility of connectionist systems. Until this rule is broken, there is little reason to believe that con-

nectionism sheds any light on how cognitive agents learn and process natural language. As we will argue in the next section, this conclusion extends to connectionist learning in general.

3.10 Connectionist Learning in ALVINN: An Illustration of the Gap between Human and Artificial Neural Network Capabilities

What is most revealing about ALVINN is not that it can mimic the human behavior of driving a car, but how it manages to do this. The neural network used for ALVINN is not exceptional in itself-it is like any other MLP trained via backpropagation. Its input is a digitized image of the road scaled down to 30x32 pixels. It outputs a steering direction, which is tied into the actual steering apparatus of the vehicle. ALVINN learns how to steer by receiving both inputs from the cameras on the vehicle and the steering angle the supervisor chooses as he drives as the correct output, this being fed from the steering apparatus to ALVINN. But first the input data must be transformed for this form of learning to work. The input image of the road is shifted both to the right and to the left of the center. In this way, ALVINN is exposed to other possible road images that the driver does not. This is done to feed ALVINN a representative sample of possible situations. Because the supervisor driving the vehicle is unlikely to move away from the center of his lane (training on real roads prevents the supervisor from swerving out of his lane), ALVINN will not be fed an example of this unless the road image is translated right and left. Furthermore, it is essential that ALVINN be fed both right and left road curves, and that neither type predominate. Finally, for ALVINN to perform well on both single and two-lane roads, separate neural nets must be trained on each type.

These limitations are clearly not present in humans. A human does not need to have trained on swerving out of his lane in order to be able to compensate when this does happen. Connectionists, as well as computationalists in general, often argue that such training techniques are not unheard of in humans: humans often prepare themselves for situations by imagining them beforehand. Even if this analogy is valid in some cases, no such analogy holds for simple road-following behavior in humans. Humans learn how to drive on what would be a drastically impoverished signal if it were all that were available to ALVINN. It is therefore implausible that a connectionist system such as ALVINN even remotely captures important aspects of human neural and cognitive structure if its learning procedures must deviate so greatly from human learning.

ALVINN is not alone in having to learn from input prepared in an extraordinary manner. All computational systems require this. What the example of ALVINN reveals is a flaw in computational learning that is extensive and likely insurmountable. Computational systems cannot adequately learn from what is for them an impoverished signal, let alone learn from signals that are impoverished for human learners.

3.11 Computationalism Revisited: Impoverished Learning

One of the most astonishing feats of human cognition is the ability of children to learn language from impoverished stimuli (Gleason 1993). The stimuli are impoverished in the sense that they only contain positive evidence: children hear only a limited number of correct utterances, without being taught what not to do, and yet are able to learn the full range of a language. This singular ability has led a number of researchers to postulate a language

organ in the brain that comes preprogrammed with a universal grammar (Pinker 1989). Although connectionist NLP researchers have reported limited success duplicating this ability with a modified form of backprop that draws implicit negative information from positive examples, connectionist inferencing is entirely bounded by how representative the data set is (Regier 1992). As we have seen, this was such a severe obstacle for ALVINN that a method for extending the data had to be developed. While it is debatable just how much learning is involved in language acquisition, and thus how much humans are themselves limited by the representativeness of the instructive stimuli, there is another cognitive domain for which the stimuli is impoverished in relation to the behavioral output: human creativity. Computational systems in general are incapable of mimicking human creativity, for they can 'only connect' by mapping inputs to outputs.

Artists do not merely mirror the natural world, nor simply transform their subject according to preestablished conceptions of it, rather they create new ways of considering their subject matter. But ANNs trained on a given function simply reproduce that function, generalizing it to novel instances, but never changing the function. One could introduce some randomness into the ANN, but this would not model directed creativity. Art does not happen by accident. This criticism extends to all computational systems. For a computational system to model artistic creation, this activity must consist of a computable function from input stimuli to output creation. Computational art has been attempted, but it is, like all computation, a syntactic process. But the production of works of art is largely a semantic operation: It is transformation of the natural world according to an idea rather than according to the structural features of the environment. Computationalists argue that the syntactic underlies the semantic. Yet, art often breaks the already given connections between syntax

and semantics, creating new connections. James Joyce's Ulysses is one such example where the traditional English syntax-semantics correlation yields to a new conception. Computational systems might be able to create pretty pictures, but they cannot yield ones imbued with a new meaning, if any meaning at all.

3.12 Why Connectionist Systems Are the Wrong Kind of Dynamic Systems

In chapter 2, I noted the incongruity of van Gelder advocating a dynamic systems alternative to computational theories of mind and his assertion that artificial neural networks are a subset of the dynamic systems approach. Like van Gelder, Smolensky sees the connectionist approach to modeling intuition as falling under the rubric of dynamic systems, and thus bearing "special" properties. He offers the connectionist dynamical system hypothesis:

> The state of the intuitive processor at any moment is precisely defined by a vector of numerical values (one for each unit). The dynamics of the intuitive processor are governed by a differential equation. The numerical parameters in this equation constitute the processor's program or knowledge. In learning systems, these parameters change according to another differential equation. (Smolensky 1988, 780)

What distinguishes a connectionist system as a dynamic system is that the state of the system can be described as "a numerical vector evolving in time according to differential evolution equations" (Smolensky 1988, 780). Van Gelder and Smolensky both have in mind a restricted sense of 'dynamical system,' because the symbol systems they criticize are themselves dynamical systems (Strogatz 1994). Symbol systems just aren't the right kind of dynamical system.

Again, no distinction is made, either by van Gelder or by Smolensky, between connectionist systems implemented on analog devices and those implemented on digital comput-

ers. Yet, according to van Gelder, one of the aspects of dynamic systems that distinguishes them from computational systems is their relation to time. A dynamic system evolves in continuous time, a computational system in discrete time steps. Connectionist nets run on digital computers lack this supposedly important property. Connectionists must make a choice: either give up the “specialness” of temporally continuous processes, or give up the results obtained on digital computers.

Similarly, connectionist nets on digital computers do not evolve continuous variables, but discrete approximations of such variables. Either there is something special about the analog nature of signals, and thus digital connectionist nets lack this special property, or there is nothing special about it. Connectionists again seem unwilling to make a choice.

Of course, digital approximations of continuous systems are of extreme value to engineers and scientists. Digital simulations of dynamic neural processes are now widely used by researchers to gain insights into the brain's processing power. An example of such methodology can be found in Touretzky and Redish's work on rat head-direction cells (Redish, Elga, and Touretzky 1996; Touretzky, Redish, and Wan 1993). Touretzky and Redish developed a neural network consisting of arrays of neurons serving as attractors for specific directions (their preferred direction). Their responses change over time according to discrete realizations of differential equations incorporating such parameters as angular velocity (for when the rat turns its head) and the neuron's preferred firing rate. Touretzky and Redish contend that their model is compatible with the response patterns of cells in the rat's postsubiculum and anterior thalamic nuclei.

The success of the Touretzky-Redish model does little, however, to bolster van Gelder's case that ANNs are a branch of dynamic systems as he envision this field, despite the fact

that this model is a much closer approximation of neural behavior than the typical back-propagation net. The Touretzky-Redish model is computational in nature (as its developers point out) and discrete in implementation (both its temporal and numerical aspects). Thus, even the best of ANN approaches does not fall within dynamic systems as van Gelder imagines them, and so is not an exception to symbolic AI in this regard.

3.13 Connectionism's Contribution: Superposition of Information

Much has been made by connectionists of the distributed nature of representations in connectionist systems, despite the fact that localist networks used for natural language processing are not distributed in any sense that distinguishes them from non-connectionist systems. When connectionists speak of the importance of the distributed nature of representations in the brain, however, they are not merely pointing to the fact that representations are spread out over a massive number of neurons and neuronal connections, but that these sets of neurons and their connections subserve numerous representations. For ANNs, this means that the same set of connection weights can produce a variety of representations. This is the superposition of information: unique sets of information collapsed into the same structure. It is the superposition of information that makes neural networks difficult to analyze, because the activities of neurons cannot be associated with discrete actions. And it is the superposition of information that enables neural networks to embed relations between representations within their representational framework without the need for explicit axioms defining the relations.

The superposition of information over distributed structures distinguishes ANNs from traditional symbolic AI systems, as well as psychologically-inspired enhancements such as spreading activation networks. Superposition of information in neural networks generally occurs not only in connection weights, but in the time-locked activations of neuronal ensembles, whereas in spreading activation networks, the nodes are generally used to each represent a single concept. The importance of the superposition of information for cognitive systems will be addressed in chapters 4, 5, and 6.

3.14 Connectionism Does Not Remedy Computationalism's Problems

Connectionism is often represented as a radical break from computationalism, one that incorporates important properties of biological neural systems, properties that are ignored by computationalists as irrelevant to information processing. These properties include the distributed nature of biological neural nets, their ability to superpose information, the subconceptual nature of connectionist processing, the graceful degradation of performance in the presence of damage to the system, the ability to generalize from training data, the nonlinear nature of processing units in the network, and neurally plausible learning algorithms. Some subset of these properties (which subset depends on which connectionist is making the argument) is presumed to enable connectionist systems to escape the failure of computationalist systems developed around the paradigms found in traditional AI to reproduce cognition. This is little more than wishful thinking.

Connectionist systems are only slightly more neurally plausible than traditional AI systems. This difference in plausibility does not amount to connectionist systems being faith-

ful reproductions of the apparent secrets that Nature has discovered in developing cognitive systems. The primary indictment against connectionist systems in this regard is that there exists an inverse relation between the neural plausibility of a connectionist system and its ability to mimic human cognitive capabilities. Backpropagation nets are the most widely used form of artificial neural networks, but are also among the least neurally plausible. For speech and handwriting recognition, the time delay neural network, a variant of the back-propagation network, has produced the best results, but is a step further from the generic backpropagation network in terms of neural plausibility (its shared weights do not have a known analog in neural systems). The lack of neural plausibility extends to the ability of artificial neural networks to continue to function despite being damaged: biological neural networks rewire themselves in response to damage, and so don't just degrade gracefully, but repair damage.

What remains of the analogical argument of connectionists is the distributed nature of neural networks, biological and artificial, and their capacity to superpose information. But the distributed nature of neural networks does not imply that their subunits are subsymbolic in nature, and are, therefore, not computing over symbols. The flawed assumption is that if a neuron is a feature detector for a set of features for which we have no concept, the neuron's activity cannot be regarded as symbolic. In itself, the distributed nature of representations in neural networks is just a fact; it does not imply that minds form concepts through subconceptual operations. It does, however, enable a greater superposition of information. And superposition of information allows for a reduction of what needs to be explicitly represented, such as relations between representations.

In itself, the capacity for superposition of information, however, does not indicate how semantic information *arises*. Instead, it points to how certain practical problems surrounding how semantic information is manipulated, problems that arise when considering the frame problem, may be solved. But by telling us how information may be stored, it circumscribes what semantic information can and can't be. Whether the frame problem requires that minds make use of such techniques as superposing information, or whether it can be solved within the framework of traditional AI, is examined in the next chapter.

Chapter 4 Semantics and Robotics: How the Frame Problem Continues to Plague Computationalism

Searle's Chinese Room argument has exposed AI's inability to produce semantics from syntax. And while it may be true that "just about everyone who knows anything about the field dismissed it long ago" (Dennett 1997), it continues to generate replies, as well as discomfort among computer scientists who come in contact with it for the first time. But even if we assume that those who dismiss it are in the right, AI still faces a host of severe obstacles. An AI system that has symbols bearing semantics is not frozen in time and space. It exists in a changing world, and these changes must be reflected in the system if it is truly 'grounded'. This task, traditionally considered to be independent of the semantics of the system, is a daunting one, because the system must choose which of its rules are relevant to the next moment in time and its novelties, and which are not. It must also keep its beliefs up-to-date as the world changes, and these beliefs must represent all of the conditions that need to be satisfied for the system to successfully carry out a task.

4.1 What Is the Original Frame Problem?

Each of these problems has been confused with what McCarthy and Hayes (1969) identified as the frame problem. The original frame problem is not one of how to adapt representations in real time to a changing world, nor how to make representations reflect the complexity of that world. It is a problem that arises when one reasons *about* a changing

world. It afflicts both real-time systems and systems that need not keep up with the world, although not all AI systems are susceptible to it. In particular, systems which do not represent time-related actions are not affected by the frame problem.

The frame problem arises when one tries to represent change with a time-related logic. The situation calculus is one example of such a logic. It is an extension of first-order logic for the purpose of expressing time relations, such as the outcome of action from one instance in time to the next. To be able to reason about actions and their outcomes in time, one must be able to determine what changes an action causes as well as the changes it does not cause. As Hayes expresses it:

> One feels that there should be some economical and principled way of succinctly saying what changes an action makes, without having to explicitly list all the things it doesn't change as well; yet there doesn't seem to be any other way to do it. That is the frame problem. (Hayes 1987, 125)

Hayes has attempted to guard the notion of the frame problem from efforts to expand it to the problems listed in the introduction to this chapter. Nonetheless, he gives some indication that he believes that there is something deeper involved with the frame problem than just properly augmenting logic-based knowledge representation systems: "we aren't carving nature at the right ontological joints, if you ask me" (Hayes 1987, 130).

4.2 How the Frame Problem Has Evolved: The General Frame Problem

Much to chagrin of its original framers (Hayes 1987), the frame problem has since metastasized into a number of problems concerning a representational system's relation to a changing world. We have seen a few of these already. The control problem is the problem of making sure the system makes the right deductions without wasting time on irrelevant

ones (Hayes 1987). The update problem is that of making sure the system's beliefs match the world as it currently is, i.e., to make sure they are up-to-date (Hayes 1987). The qualification problem is that of making sure one has represented everything that needs to be satisfied in order to succeed in a task (Janlert 1987). A few of the problems identified with the frame problem can rightly be dismissed as not really being problems, but rather areas of study (Hayes 1987), such as Fodor's notion of Hamlet's problem (Fodor 1987), which is the problem of when to cease thinking and just act. Nevertheless, the frame problem does extend beyond axiomatic systems, because it forces researchers to develop a representational scheme that avoids it, i.e., finding a replacement for logic as the representational medium. The problem of "finding a representational form permitting a changing world to be efficiently and adequately represented" (Janlert 1987, 7-8) is the general frame problem, and it afflicts AI as a whole.

The expanded frame problem is generally conceived to be a question of the form, not content, of representations:

> The frame problem is an indication that informationally equivalent systems, systems conveying the same information about the represented world, may yet differ drastically in efficiency. The situation is somewhat analogous to choosing between different programming languages; although they are computationally equivalent in the mathematical sense, we recognize that there are important differences in the ease and efficiency with which they are applied to different types of tasks. (Janlert 1987, 8)

Janlert paradoxically argues that the frame problem is a problem not of what data-structures and algorithms to use, but of how to take into account the way the world is. He goes so far as to contend that a solution to the frame problem will have to get the metaphysics of the world to be modeled correct. This would seem to contradict the proposition that the frame problem is not concerned with the content of representations.

It is Janlert's quoted formulation, and not his tying of the form of representations to the metaphysics of the world, to which computationalists must adhere. If AI systems are capable of representing what humans represent, but nevertheless fall victim to the frame problem (whereas humans do not), then the fault must be in the form of the representation. But this is to assume that 1) AI systems can represent what we represent and 2) the content of a representation can truly be separated from its form. Janlert's stipulation that the suitable form of representations depends on the metaphysics of the world indicates that he doubts 2).

If we are to assume Church's thesis, it is unclear how the form of representations can be the key to solving the frame problem. AI systems with different forms of representation may still be computationally equivalent: each capable of computing the same functions. Thus, a computationalist who accepts both this formulation of the frame problem and Church's thesis must reject functionalism, because he must accept the possibility that between two functionally equivalent systems there is a world of difference, namely the ability to model the world. Difference in form of representation can account for difference in efficiency of computation, but this is only one half of a solution to the general frame problem. The other half, finding a representational scheme that "adequately" models the world, is left unsolved. Unless, of course, the form and content of a representation are not arbitrarily related as the Physical Symbol System Hypothesis would have it.

4.3 Attempted Solutions to the Original Frame Problem

4.3.1 The ‘Sleeping Dog’ Approach to the Frame Problem

One approach to solving the frame problem is to declare that it doesn’t exist, an approach taken by Drew McDermott in his essay “We've Been Framed: Or, Why AI Is Innocent of the Frame Problem” (McDermott 1987). Dividing the frame problem into its logical, computational, and metaphysical aspects, McDermott declares the former two solved and the last aspect irrelevant.

According to McDermott, the logical aspect of the frame problem, the question faced by axiomatic systems as to which axioms to apply and which not to, is solved by jettisoning monotonic logic and using nonmonotonic logic. Nonmonotonic logic systems allow for default beliefs that take precedence in an absence of contradicting evidence. A nonmonotonic logic system that saw a person seated behind a desk but was unable to see the person’s legs would not deduce that the person had no legs. The default belief that people have two legs would be preferred over a deduction with insufficient evidence. Only evidence that could demonstrate the original belief to be false would be sufficient to revise this belief. One axiom replaces the countless axioms specifying the relations between all possible events:

> If a fact is true in a situation, and it cannot be proven that it is untrue in the situation resulting from an event, then it is still true. (McDermott 1987, 115)

The computational problem is solved by ‘letting sleeping dogs lie’, or just not computing anything beyond the effects of a given situation. This assumes that we know what to change in our database for each situation, or if a separate database is used for each situation, what to put in those separate databases.

As for the metaphysical problem, McDermott sees this as the problem of noticing "all and only relevant inferences about change" (1987, 121). He argues that because humans cannot accomplish this feat, we should not expect AI systems to, nor need they, in any case. But this argument whitewashes important differences between what humans and AI systems can do by lumping both into the irrelevant category of 'incapable of knowing all' and overestimating the degree to which AI systems can mimic humans. The limitations AI systems have in accomplishing what humans do indicates that this aspect of the frame problem is very real.

4.3.2 Why the 'Sleeping Dog' Approach Does Not Address the Frame Problem

It is not necessary to address the frame problem's logical and computational aspects, as McDermott sees them, in order to demonstrate that the general frame problem still exists. But if McDermott's claims about these two aspects are themselves incorrect, then the general frame problem must exist, as no one has solved its most concrete obstacles.

The 'sleeping dog' approach does not, in fact, address the computational facet of the frame problem. The reason is that the method for making the problem computationally tractable is itself computationally intractable. McDermott's proposed solution, having separate databases for each situation and only updating information in the database for the situation at hand relevant to the occurring event, assumes an exhaustive solution to the frame problem. If we could only create databases of all the knowledge that enables us to act in all situations, as well as of what is relevant to each event occurring in each situation, then we could create the solution McDermott envisions. McDermott is right that humans do not know all relevant inferences for every situation, but this does not imply that encoding

human ability to make relevant inferences is a tractable problem. Humans are capable of making the correct inference in countless novel cases. The novelty of the situation may be irrelevant, but the irrelevance of it must also be accounted for in the databases. And in many cases, one cannot imagine a solution to a problem until it is experienced. How is this to be encoded in an AI system?

The restrictions McDermott places on what needs to be encoded are not sufficient to make this a tractable problem. The solution McDermott imagines assumes a massive collection of databases. In order for this implementation to be plausible, that is, within reach of a system that does not have a brain the size of a room, there must be superposition of information in the system's brain. But the superposition of information excludes McDermott's proposed solution of separate databases for situations. McDermott's claim that "no working AI program has ever been bothered at all by the frame problem" (1987, 116) reduces to: No AI program works in such a way as to need to solve the frame problem. This is just to say that no AI program tackles the hard problems of cognitive beings. McDermott's supposed solutions are no more than blank checks.

4.3.3 Circumscription as a Possible Solution to the Original Frame Problem

The 'sleeping dogs' strategy is an attempt to formalize what has been termed as the 'common sense law of inertia' (Shanahan 1997), which holds that inertia is the norm and that change is exceptional. As noted above, accommodating this law in one's representational system generally means moving away from monotonic to nonmonotonic logic. One such effort to formalize the law of inertia by means of nonmonotonic logic is McCarthy's circumscriptive logic (1980, 1986). It is an extension to the situation calculus which adds a

predicate, *Abnormal*, and attempts to minimize it, i.e., to limit the set of objects about which is true.

The situation calculus is a many-sorted first-order predicate calculus, where many-sorted means that the variables, constants and functions of the calculus can be assigned to different sorts. The sorts generally consist of *situations*, *fluents*, and *actions*. Situations are snapshots of the world at a particular moment. Fluents can be functions with the space of situations as domains, or they can be reified to be objects representing a limited situation. Actions are what the name suggests. The situation calculus also includes a function *Result* from actions and situations to situations, and a predicate *Holds* that takes a fluent and a situation as arguments. To express that it is raining in situation S0, we would write: Holds(Raining, S0) (Shanahan 1997). *Effect axioms* are formulae that express what holds as the result of a particular action.

A circumscriptive logic adds the additional predicate Abnormal, applying it as follows:

Equation 4-1. Application of Abnormal in Circumscriptive Logic

$$\neg \text{Abnormal}(a, f, s) \rightarrow [\text{Holds}(f, \text{Result}(a, s)) \leftrightarrow \text{Holds}(f, s)]$$

This states that if an action on a fluent in a situation is not abnormal, then the result of the action on the fluent does nothing. Minimizing the predicate Abnormal means adding a second-order formula to limit its applicability while allowing other predicates to vary (they can apply or not apply as their arguments are abnormal or not). In the situation calculus defined above, it would be the predicate Holds that is allowed to vary.

Although promising, circumscription has been demonstrated to lead to what is referred to as the “Yale Shooting Scenario” or the Hanks-McDermott Problem (Hanks and McDermott 1986). The Yale Shooting Scenario attempts to formalize a situation in which some-

one gets killed by a gunshot using three actions (Load, Wait, and Shoot), two fluents (Alive and Loaded), and two effect axioms:

Equation 4-2. Effect Axiom 1: Load puts a bullet in the gun

$$Holds(Loaded, Result(Load, s))$$

Equation 4-3. Effect Axiom 2: Shoot action kills victim

$$Holds(Loaded, s) \rightarrow \neg Holds(Alive, Result(Shoot, s))$$

There are two observation sentences about the initial situation, namely that the victim is alive and the gun is not loaded. We can produce a model with minimal abnormalities that produces the intended result, but it is also possible to produce an anomalous minimal model in which the Wait action unloads the gun. Thus, minimizing a predicate does not guarantee us a result free of nonsensical consequences. This has been demonstrated for a number of nonmonotonic logics (Hanks and McDermott 1986).

A number of other criticisms have been directed against the nonmonotonic logics that have been suggested as solutions to the frame problem. These criticisms include that these logics cannot represent concurrent action, continuous change, domain constraints, and actions with non-deterministic effects (Shanahan 1997). Shanahan claims to have overcome all of these defects and to be on the verge of a complete solution to the frame problem.

4.3.4 Recent Developments: Shanahan's Circumscriptive Event Calculus

Shanahan (1997) proposes three criteria that a solution to the frame problem ought to meet: representational parsimony, expressive flexibility, and elaboration tolerance. A solution that is representationally parsimonious allows for the construction of representations that are compact, meaning that the size of representations is approximately equal to the sum of the number of fluents and actions involved in the representations. Expressive flexibility

means that a solution for a simple domain should be applicable to a complex domain. In other words, adding arbitrary domain constraints, non-deterministic events, and continuous change to a domain should not result in the reintroduction of the frame problem. A solution is elaboration tolerant if adding new information does not demand additional effort beyond the complexity of that new information; the ideal is that one could append new sentences directly to a theory and produce the new, desired theory. These criteria were introduced by Shanahan to avoid solutions that require the theorist to produce domain-specific logics, which in turn increases the computational burden of a cognitive system employing such a scheme. A solution that meets these three criteria would be more easily implemented. In fact, it is conceivable that solutions to the frame problem that do not meet these criteria are not implementable.

Shanahan claims that his near solution to the frame problem, the circumscriptive event calculus, meets all three criteria. The circumscriptive event calculus departs from the situation calculus by adapting a narrative time line and associating a situation with each point in the line. A narrative is "a distinguished course of events about which we may have incomplete information" (Shanahan 1997, 155). The calculus introduces the functions *Initial(), Start(), Happens(), Actual(),* and *HoldsAt(),* the purpose of which is evident from their names. Happens() is minimized, while a further function *State()* (State(t,s) means that time t is associated with state s) is allowed to vary. A state is simply a set of fluents. The purpose of introducing State() is so that

> each time point is associated with a single, characterizing state s, such that the fluent f is in s if and only if f was initiated by some event before t and still holds at t, and Not(f) is in s if and only if f was terminated by some event before t and still doesn't hold at t. (Shanahan 1997, 274)

Abstate() is the event calculus version of Abnormal as it is applied to states. Shanahan's proposed solution to the frame problem consists of minimizing Abstate() at a high priority, minimizing Happens(), Initiates() and Terminates() at a lower priority, and allowing HoldsAt() and State() to vary.

Events with non-deterministic effects are handled by assuming that the non-determinism really doesn't exist. Instead, an event with non-deterministic effects is considered to be a disjunction of events with deterministic effects, and the non-determinism appears as such only because we don't know which events occurred. It may be noted that this approach contradicts the reigning interpretation of quantum mechanics, which holds that events are in fact truly non-deterministic, but not being able to represent quantum mechanics is hardly confined to just AI systems.

Representing concurrent events requires the introduction of an axiom defining a variation of Happens() that allows us to express that two events are concurrent, and a predicate *Cancels(a1, a2)* that means that events a1 and a2 cancel one another. Cancels() is minimized with the same priority as Happens().

To represent continuous change, Shanahan uses the predicates *Triggers()* and *Trajectory()* and distinguishes between discrete and continuous fluents. Trajectory(f1, s, f2, d) denotes that after a period of time d, the continuous fluent f2 holds if the discrete fluent f1 is initiated in state s. Triggers(s, a) indicates that an event of type a occurs in state s. Shanahan illustrates this formalism with the example of how one would represent the event of a ball moving horizontally along a surface toward a vertical shaft, falling into which it bounces back and forth until it reaches the bottom. He introduces four event types: *Propel(v)*, *Drop*, *Bounce*, and *Stop*. Propel(v) means that the ball is set in motion with velocity

v. Four fluents are introduced as well: *Distance(x)*, *Height(x)*, *Moving(v)*, and *Falling*. Distance(x) indicates that the distance to the hole is x, Height(x) means that the height of the ball is x, Moving(v) indicates that the ball is moving with velocity v, and Falling means that the ball is falling. The quantities A, B and C represent the distance from the starting point to the near wall of the shaft, the distance from the starting point to the far wall and the height of the shaft, respectively. A set of axioms formalize notions about horizontal movement:

Equation 4-4. Axioms concerning horizontal movement

$$\text{Initiates}(\text{Propel}(v), \text{Moving}(v), s)$$

$$\text{Releases}(\text{Propel}(v), \text{Distance}(x), s)$$

$$\text{HoldsIn}(\text{Distance}(y), s) \wedge x = (y + v \times d) \rightarrow \text{Trajectory}(\text{Moving}(v), s, \text{Distance}(x), d)$$

The next set formalizes what happens when the ball reaches the shaft.

Equation 4-5. Axioms representing the ball beginning to fall

$$\text{HoldsIn}(\text{Distance}(x), s) \wedge x = A \wedge \text{HoldsIn}(\text{Moving}(v), s) \wedge v > 0 \rightarrow \text{Triggers}(s, \text{Drop})$$

$$\text{Initiates}(\text{Drop}, \text{Falling}, s)$$

$$\text{Releases}(\text{Drop}, \text{Height}(x), s)$$

$$\text{HoldsIn}(\text{Height}(y), s) \wedge x = (y - 4.9 \times d^2) \rightarrow \text{Trajectory}(\text{Falling}, s, \text{Height}(x), d)$$

The next set represents when the ball reaches the far wall of the shaft and a sequence of bounce events are triggered.

Equation 4-6. Axioms for bounce events

$$\text{HoldsIn}(\text{Distance}(x), s) \wedge \text{HoldsIn}(\text{Moving}(v), s) \wedge [x = B \wedge v > 0] \vee [x = A \wedge v < 0] \rightarrow \text{Triggers}(s, \text{Bounce})$$

$$\text{HoldsIn}(\text{Moving}(v1), s) \wedge v1 = -v2 \rightarrow \text{Initiates}(\text{Bounce}, \text{Moving}(v2), s)$$

$$\text{HoldsIn}(\text{Moving}(v), s) \rightarrow \text{Terminates}(\text{Bounce}, \text{Moving}(v), s)$$

The last set represents what happens when the ball reaches the bottom of the shaft.

Equation 4-7. Axioms for halting

$$\text{HoldsIn}(\text{Height}(x), s) \wedge x = C \wedge \text{HoldsIn}(\text{Falling}, s) \rightarrow \text{Triggers}(s, \text{Stop})$$

$$\text{Terminates}(\text{Stop}, \text{Moving}(v), s)$$

Terminates(Stop, Falling, s)

HoldsIn(Distance(x), s) → Initiates(Stop, Distance(x), s)

HoldsIn(Height(x), s) → Initiates(Stop, Height(x), s)

Along with the appropriate domain constraints guaranteeing the uniqueness of the ball's height and distance from the starting point, as well as the proper initial conditions, these axioms produce only conclusions which would be expected without any bizarre consequences as found in the Yale Shooting Scenario. The work of Shanahan and those upon whom he builds appears to have produced real progress toward a solution for the specific frame problem.

4.3.5 What Is Missing from the Circumscriptive Event Calculus?

Shanahan is leery of declaring the frame problem solved, so what then remains unsolved? Shanahan gives only a few details of what might be considered missing from his circumscriptive event calculus. It is apparently unable to deal with hypothetical sequences of events, which deductive approaches are able to handle. It is also unsuited to representing complex actions, although Shanahan contends that they can be accommodated with only minor modifications to the calculus. So why not declare the frame problem solved?

The primary reason for not declaring the circumscriptive calculus of events to be a solution is that there is as yet no general proof that it will not produce anomalous models. Another reason, though denied by Shanahan, is that it is not really representationally parsimonious. Shanahan's own example of the ball and shaft situation requires on the order of (f x a) axioms rather than the (f + a) that Shanahan defines as the approximate measure of parsimony. And there is reason to believe that this problem is only exacerbated when we consider complex actions in addition to continuous change and concurrent events.

Unfortunately, the world does not resemble the ball and shaft example, with simple actions and linear change. Some researchers (Lespérance et al. 1994; Levesque et al. 1997) have attempted to accommodate complex actions in their solutions to the frame problem, but it is important to note what they conceive to be complex actions:

> the behavioral repertoire of a robot must include *complex* actions, for example the action of clearing off a table, defined as something like "While there is an object on the table, pick it up, then place it on the floor". This action is defined in terms of primitive actions "pick up an object" and "put an object on the floor", using iteration and sequence. (Lespérance et al. 1994)

A basic tenet of this approach is that complex actions can be broken down into simple, primitive actions, yet primitive actions are only consistently primitive—involving only one action—in laboratory or strictly controlled settings. Consider the real-world complex action of a soccer goalie attempting to catch a swerving ball crossed to the center of the goal mouth with opposing players attempting to head the ball into the net. What are the primitive actions that make up this complex action? If they exist, they are extremely context dependent. A different set of primitive actions are required when slightly different conditions prevail, such as the direction of the wind.

When complex, concurrent, and continuously changing events occur, we must either provide the frame axioms describing the interactions of the primitive actions and fluents, or we must substitute equations describing the physics of the situation. An example of this latter tactic can be seen in Shanahan's ball and shaft example. Presumably with our soccer goalie example, an equation for the ball's trajectory must be supplied, otherwise a step-by-step description of its flight pattern would be necessary. Yet the situation and event calculi are attempts to avoid having to know the physics of the world in order to represent it and act within it. One alternative to either knowing the exact physics of the world or formalizing

common-sense in event or situation calculus is pattern recognition, an alternative that has been suggested by a number of authors (Dreyfus 1972; Margolis 1987). One aspect of certain pattern recognition systems, the superposition of information of neural networks, will be examined in the context of its usefulness in solving the general frame problem.

4.4 How the Specific Frame Problem Leads to the General Frame Problem

Both Shanahan (1997) and Hayes (1987) contend that the specific frame problem can be divorced from the general frame problem. The general frame problem spans a host of specific problems—the update problem, the prediction problem, the control problem—all of which relate to a system's ability to represent and act in a changing world. While it is true, as Shanahan argues, that one can solve the specific frame problem without discussing the general frame problem, it is not necessarily true that each component of the general frame problem can be solved independently to form a final, general solution. A solution to the specific frame problem may worsen other aspects of the general frame problem. For example, a solution to the specific frame problem that is not elaboration tolerant and representationally parsimonious makes the update and control problems much worse by expanding the number of axioms that must be considered in order to properly update and control the system. And using event or situation calculus to tackle the qualification problem leads to a solution to the specific frame problem that is not representationally parsimonious, because it attempts to represent complex events in terms of primitive actions and their combinatoric relations.

What this means is that even if cognition is computable in Rapaport's sense, it might not be realizable as a computation because of the demands it would place on its realizing hardware. Hayes (1977, 1987) and others (Janlert 1987; Shanahan 1997) have speculated that the origin of the frame problem may lie in attempting to represent that world as a collection of individuals and the relations between them. This is exactly the position argued for here. Symbol systems are inappropriate means for representing the world. This is due to two factors. The first is that attempting to build representations from primitive actions and entities and their relations leads to representational schemes that are too complex. Even if such a system were to solve the specific frame problem, it would be faced with the problem of updating the system's beliefs in real-time. This leads to the second factor. The representations of symbol systems are not 'live' in the sense that they adapt with the changing world. Symbols and computations over them are timeless in that the algorithms in which they figure produce the same result regardless of how long they take. If something is an algorithm for thinking thought X, it is such whether finishing it takes a nanosecond or a century. They are also contextless in that if a system's connections to the world are cut, the algorithms in which they figure produce the same content. It is because of these aspects that the update problem occurs.

A working alternative to the symbolic approach can be found in the human brain. Although we do not know exactly why it works, there are aspects of human neural functioning that are known and offer a possible course for solving the general frame problem. One such aspect is the superposition of information in neural structures, an aspect that, as was noted in the previous chapter, also appears in artificial neural networks.

4.5 Superposition of Information as a Step Toward Solving the Frame Problem

In the previous section, I argued that attempting to pick out and tag primitives in the world with symbols and attempting to enumerate their relations is what leads to the frame problem. Connectionists have proposed that distributed representations rather than discrete symbols provide a better way of representing the world. The distributed nature of connectionist representations does indeed assist in solving one aspect of the general frame problem, namely the need for representations that are plastic and gracefully degrade so that there is not a catastrophic failure to match representation and world when the world (or representational system) changes. Superposition of information in ANNs has also received some attention, but not in the context of the frame problem. Connectionists have pointed out that relations between representations are captured implicitly through the superposition of information (Smolensky 1988; van Gelder 1991; O'Brien and Opie 1999).

The relation between representations that are superposed is generally considered to be that of similarity, either membership in the same category or similarity across categories. Even if it merely establishes similarity, superposition gains the cognitive system a considerable reduction in complexity when combined with spreading activation. If one representation triggers activation of other neural structures representing a particular relation, a similar representation may also trigger this relation. A computationalist might object that something similar can be achieved with the generalization of frame axioms; however, if there are any qualifications on these generalizations, they will have to be formalized, whereas these qualifications could be superposed in a neural network.

Superposed information embodying similar representations need not trigger only similar relations. Chaotic systems, which, according to Skarda and Freeman (1987), at least some of the brain's structures qualify as, are capable of bifurcations—large-scale changes in the dynamics of system—with only small parameter changes. Two superposed representations might trigger radically different behaviors depending on slightly different outputs produced in their formation, which is the activity of the net given a trigger for each representation. Bifurcation will be discussed more thoroughly in the following chapter.

Superposition of information may also help alleviate the update problem. Updating the relations pertaining to a specific representation can also result in the representations superposed with it also having their relations updated.

In the final analysis, if the conclusion that the frame problem, both general and specific, is tied to symbols systems is correct, we would see it manifested in actual AI systems. Much is made by proponents of symbolic AI that the specific frame problem never really troubles any AI system (Hayes 1987; McDermott 1987). This is certainly true for chess-playing systems. It is certainly not true if we look at robotic systems that are meant for operating in domains that are not simply toy (laboratory) worlds, and especially if we consider the general frame problem. So how might robotics provide a test for solutions to the frame problem, and, consequently, for whether we are, as Hayes (1987, 130) puts it, "carving nature at the right ontological joints"?

4.6 A Test of Semanticity: Robotics and its Failures

As McCarthy (1996) has pointed out, Dreyfus, in his critiques of symbolic AI (1972, 1992), has given little consideration to the wealth of research programs that fall under that rubric.

Dreyfus focused primarily on chess playing programs and the knowledge representation system Cyc (Lenat and Guha 1990), an attempt at creating a cognitive system by representing enough information. Deep Blue's triumph over Kasparov does damage to Dreyfus's critique, because Dreyfus attempted to apply that critique to a domain which does not suffer from the frame problem, specific or general. The Cyc project, on the other hand, continues to miss the deadlines set by the researchers working on it, and Dreyfus is right to characterize it as a "degenerating research programme" (Dreyfus 1996). However, neither Cyc nor Deep Blue assists us in assessing whether symbolic AI has overcome the general frame problem and, therefore, the representational challenge underlying it. An indication of whether symbolic AI can overcome the general frame problem must be sought in the realm of robotics, for only there is the AI researcher required to solve it. Furthermore, the field of cognitive robotics allows us to judge whether the efforts at solving the specific frame problem yield representational systems that can be applied effectively to real, dynamic environments.

The task at hand is not merely to find robots that successfully carry out the tasks assigned to them. Most of the successes in the field of robotics can be ascribed to efforts to constrain working environments so that a robot can operate in them. Highly specialized sensors that reduce the need for reasoning about perceptions are also essential for robot navigation and localization. Whereas humans use stereo vision along with monocular depth cues to perceive 3D environments, structured light techniques are the most widely used methods for achieving this goal in general computer vision, and, by extension, in robots that depend on computer vision. Structured light techniques involve casting a light source with known characteristics (usually a laser line) onto a scene, and using the knowledge of the

position of the light source relative to the camera to determine the depth of the detected light points in a captured image. When cost is not an issue, laser range finders might also be used to produce even denser range data. We might characterize this as a perceptual approach to robotics, as opposed to a cognitive approach, which would attempt to use high-level reasoning abilities on less rich data to 'figure out' how to achieve a goal. The perceptual approach seeks to find a unique attribute of a goal or obstacle and a sensor that picks out that attribute with the least possible noise, and then often attempts to constrain the environment so that there is no interference in sensing this unique attribute.

The perceptual approach is more cost-effective and easier to implement than a solution to the general frame problem. It is questionable, however, whether it can be extended beyond the simplest of environments. The difficulties lurking in very constrained environments for robots based on perceptual approaches is starkly illustrated by the Demeter robot for automated harvesting (Ollis 1997; Pilarksi et al. 1998).

4.6.1 Percepts without Concepts: The Perceptual (Sensor-based) Approach to Robotics

Demeter is a New Holland harvester that has been automated by the National Robotics Engineering Consortium to cut alfalfa and sudan fields without the intervention of human operators. Alfalfa fields in the Imperial Valley, CA are exceptionally suited to robotic operations, with few obstacles, lush crop, soil color quite distinct from crop color, and rectangular fields. The original Demeter harvester was guided by a computer vision system that attempted to detect the boundary line between cut and uncut crop. A variety of bandpass filters were tested to find whether unique portions of the light spectrum corresponded to cut and uncut crop, but this methodology yielded little information to aid segmentation of

images. The final algorithm segmented cut from uncut crop in images taken by cameras mounted on either side of the harvester cab by searching for a step in color space and adapting weights applied to the RGB values in the images according to the separation found in prior images. To deal with shadows, the shadowed pixels, determined by applying a threshold, were adjusted according to the difference in spectral power distribution between shadowed and unshadowed regions. In a number of settings, these techniques worked admirably, enabling the harvester to cut with accuracy and speed approaching human competency.

Even in an environment as constrained as an alfalfa field and as ideal as the Imperial Valley, Demeter is unable to consistently track the boundary between cut and uncut crop under normal working conditions. These conditions include cutting throughout the day and without having to stop for patches of sparse crop. The position and cause of shadows change as the day progresses, making a poor shadow compensation algorithm even worse. The problem of shadows obscuring important features is ubiquitous in computer vision; a survey of the relevant literature reveals that no general technique for adequately compensating for shadows exists. Yet human drivers are able to follow a boundary line obscured by shadows even when there is no appreciable color difference to the two sides of the line. A number of cues are used, including depth perception, the position of cut crop piles (known as windrows), where the crop hits the cutter bar, and knowledge of the general curvature of the boundary line. Sections of sparse crop require similar reasoning about the situation for the driver to make the correct cut. Demeter has since been augmented with GPS to constrain the output of vision, and subsequently to replace the vision system. Demeter no longer perceives, let alone reasons—it is guided through a combination of knowledge of

a field's layout and godlike omniscience of where it is provided by GPS. Demeter succeeds because the environment it faces has been made static. Its frame problem has not been solved so much as eliminated.

4.6.2 Concepts without Percepts: The Cognitive Approach to Robotics

At the other extreme is the approach taken by researchers in Cognitive Robotics, which is an effort to develop high-level robotic control, enabling a robot to reason about its environment in addition to perceiving it (Levesque and Reiter 1998). This stands in contrast to 'automated planning', a methodology that seeks to determine how a robot can achieve its goal a priori and then provides the robot with the solution (this describes Demeter's current state). Levesque and Reiter (1998) identify four problems with the automated planning approach, problems that describe the general frame problem:

> *no sensing*: the planning system is expected to generate a sequence of actions without considering the results of sensing;
> *lack of reactivity*: exceptional situations might arise during execution: high-priority interrupts, failures of execution modules, unanticipated situations;
> *computational intractability*: for all but very simple domains, automated planning appears to be infeasible; at its very best, planning seems ill-suited to generating very long sequences of actions;
> *incompatibility with conventional robotics*: conventional robotics deals with micro-actions where decision are made many times per second in worlds characterized by noise and uncertainty. (Levesque and Reiter 1998, 1)

Their proposed method for tackling high-level robotic control incorporates a version of the situation calculus developed by Reiter (1991) to solve the specific frame problem and cast into a programming language, GOLOG, based on Algol. The input to robotic systems based on this approach is not a goal, but a program to be executed, and a program interpreter generates the primitive actions needed to execute it.

A pair of robots have been implemented using GOLOG, including a mail delivery robot at the University of Toronto and a museum guide at the University of Bonn (Levesque and Reiter 1998). According to Tam (1998), the mail delivery robot dealt with obstacles by stopping for 15 seconds. If the obstacle remained, it was considered permanent and mail could not be delivered to the intended person. Each person was given 2 credits, and the credit total for a person was reduced by 1 each time the robot could not deliver mail to that person. At 0 credits, the person would no longer receive mail from the robot. Tam points out that, due in part to hardware limitations, the methodology could not be tested "with a real robot and environment" (Tam 1998, 124). An interesting aspect of the experiments with the mail robot is that the term used by the robot to describe the current situation grew over time such that the robot took longer and longer to reason. A technique referred to as 'rolling forward' or 'progressing a database' (Lin and Reiter 1997), which treats the current situation as an initial situation, was used to limit the length of the term. No information was given regarding the computational burden of such a procedure or its possible side-effects. Intuitively, removing the robot's memory of a situation's history would seem to greatly limit its ability to adapt to its environment. All persons with 0 credit would have that as an initial condition rather than as a condition caused by the environment and therefore amenable according to changes in the environment.

The cognitive robotics research presented above represents only minor progress toward a solution of the general frame problem, and is hardly a renaissance for the symbolic approach as McCarthy (1996) makes it out to be. And it gives little indication that further, substantial progress will be made, given its limited success with a toy environment and task.

4.6.3 How to Apply Robotics: A Robot Turing Test

Although it has been argued that for a computer to successfully pass the Turing Test, it must have solved the frame problem (Crockett 1994), a more direct test of an AI system's capacity to overcome the general frame problem, and, therefore, to represent the world as humans do, at least in instrumental terms, can be easily developed. Embodying an AI system in a robot that must navigate in real-world settings while carrying out complicated tasks would automatically eliminate any charge of rigging results or lack of adaptability. The test would not rely on how well the system can deceive observers, but how well it achieved its explicit goal. The test would not necessarily require human-level intelligence, although raising the threshold to that level would eliminate any further doubts whether that additional step could be attained, nor would the robot have to fool people into thinking that it was actually a human. Such an exercise undoubtedly tests more than the minimal requirements necessary for establishing a system to be cognitive, but this would answer critics of the Turing Test who argue that it does not set the threshold high enough.

What then are the general criteria for such a test? The following list is not meant to be exhaustive, nor do I claim that it is minimal. It is simply meant to expose whether a robot has solved the general frame problem, which I have argued indicates whether a system can adequately represent the world.

1. *Goal Identification and Generalization*: The goal should not be too specific or too uniquely identified. The test should not be one of, for example, taking uniquely identified object X to uniquely identified place Y. Generalizability means that it should be extendable to similar situations. An example of a goal that is not too unique, and that is generalizable, is to stock products on a shelf in a busy store. If

the robot can stock items in aisle 1, it should be able to stock them in aisle 2, even if aisle 2 has differently positioned shelves. Nor should the robot require a predetermined map and plan of the shelves.

2. *Complex Goals*: The robot's goals should not require merely simple actions to achieve, such as: drive straight, drop item, drive back.

3. *Dynamic Obstacle Detection and Avoidance*: The robot should be able to detect obstacles, whether permanent or temporary, moving or static, and either maneuver around them or move them out of the way. A test in which the robot need only give up after detecting a permanent obstacle tests very little. This criterion alone tests a wide array of abilities, such as the capacity to plan a new path and the ability to determine the presence and type of obstacle.

4. *Goal Reformulation*: If achieving the goal is not possible, can the robot reformulate the task to achieve an approximation to it. One might argue that this requires human-level intelligence, but a simpler version would simply require that the robot do as much as is possible.

5. *Tolerance to own failures*: The robot should be able to recover from its own mistakes as well as tolerate minor failures in its system. An example of the former is picking up a product if the robot dropped it in its attempt to stock it on a shelf. An example of the latter is the robot continuing to complete its task even if one of its wheels breaks and it is still able to drive; the robot may have to adjust how it completes the task, but if it is physically within its ability, it should attempt it. Even better would be to dynamically determine whether the task is still within its ability.

Requirements such as these will produce howls among researchers who concentrate on the theoretical or higher-level aspects of cognition. They will object that these criteria represent 'hardware problems', and that they should not be required to solve them. This is utter nonsense. Although there is an underlying assumption that certain hardware solutions are found, such as a sensor to detect whether a wheel is broken, the criteria test the robot's ability to handle the update, prediction, and control problems in addition to reasoning about a changing environment without deducing anomalous occurrences (such as stocking an item on a shelf causing the floor to disappear), i.e., the general as well as the specific frame problems. Requiring this to be done in hardware does more than make more work for researchers: it eliminates the wild implausibilities and beneficial constraints of simulation. Everything works in simulation.

I have already given a hint as to what an example of such a test might look like. A robot that must stock shelves in busy grocery stores would face all five tests. The robot could be taken to any store that stocked products on shelves, and it would successfully stock them as customers are walking through the store. The reason robotics is so far off from achieving this is not hardware limitations. CCD cameras as sensitive as the human eye have long been available. What is lagging behind is the state of AI.

4.7 The Frame Problem Persists

Symbolic AI faces severe theoretical and implementational obstacles, not least of which is the specific frame problem. Although progress has been made toward a solution of the specific frame problem, a solution to the general frame problem continues to elude AI researchers. It has been argued here that the proposed solutions to the specific frame prob-

lem actually exacerbate aspects of the general frame problem. The source of this is the nature of symbolic AI, namely its efforts to carve the world into individuals represented by discrete tokens and their relations. Thus, the problem facing symbolic AI is not merely which algorithms to use, but how it conceives the relation between representation and world.

A general test of symbolic AI's ability to tackle these difficulties can be found in the progress or lack of such in the field of cognitive robotics. Cognitive robotics has yet to produce a system that can reason about, react within, and plan for dynamic environments. Approaches in robotics that depend on reliable sensors in effect reduce their environment to that of their sensor domains. The unexpected need not be anticipated or adapted to, because the state of the sensor in question reliably correlates with expected states of the environment. The environment is static in the sense that all of its relevant states are known and correspond to particular sensor states. Neither one of these approaches gives hope as yet that the general frame problem will be solved—in the case of cognitive robotics, the reasoning abilities of the robots are not up to the task, in the case of perceptual approaches, we do not know if all of the relevant uncertainties of dynamic environments can be eliminated by sensing. In the end, it is an empirical question whether symbolic AI will succeed in its ultimate goal of creating cognitive agents, but there are strong reasons to suspect that it will not.

Chapter 5 Ecological and Evolutionary Alternatives to Explaining Semantic Information

Computational approaches generally rely on methodologically solipsistic notions about the nature of content. ‘Meaning is in the head’ is a slogan that is often used to sumup the general thrust of methodologically solipsistic presuppositions about the nature of content. Rapaport’s syntactic semantics is paradigmatic of such approaches: semantics is derived from a syntactic system within the cognitive agent and does not require links to the outside world. Even attempts at symbol grounding, such as Harnad’s, understand the meaning of symbols in terms of icons within the cognitive system and not by reference to the environment. And computationalists such as Fodor have argued that cognitive psychology requires narrow content, or content that depends only on functional role (understand in terms of input-output relations) and relations to other mental states (although Fodor seems to have changed his tune a bit, see Fodor 1994).

Although methodologically solipsistic approaches to explaining mental content continue to be applied by cognitive scientists, many philosophers have come to believe that they are fatally flawed. This intuition was given expression in Putnam’s Twin Earth thought-experiment (1975), and Burge’s subsequent critiques (1979, 1982, 1986) of individualistic accounts of mental content. These philosophers argued that mental content is wide rather than narrow in the sense that more than just the individual’s cognitive state determines content; facts about the environment of the individual are also relevant in setting the content of mental states. Theories of wide content are ‘externalist’ in nature, which

is to say that external facts factor into the content of mental states, and therefore that mental states 'are not in the head', to use another slogan. There is an entire spectrum of externalist theories, ranging from locating semanticity exclusively in the environment (such as a more radical reading of Gibson's direct realism) to various mixes of external and internal factors.

Externalist theories also run the gambit with regard to the degree to which they rely upon the notion of natural selection to explain how mental states relate to the environment. Twin Earth thought-experiments do not rely on the historical or evolutionary considerations for their force. But such purely ecological or relational approaches leave open the question of by virtue of what do mental states relate to the environment. One answer is that they covary reliably with environmental features, i.e., they track the environment. Many externalists, however, do not believe that this is a sufficient answer, for it leads to panpsychism. The natural next step is to say that they do so by virtue of their having been selected for how they track environmental features. Thus the prominence of evolutionary considerations in externalist accounts.

It is important to note that externalism is not inconsistent with computationalist explanations of cognitive function. Markers in the head can conceivably track the environment, in the sense of being reliably tokened when and only when the corresponding environment feature is present, and transitions between markers that track changes in the environment can conceivably be brought about by computational processes. It is just that the content of these markers derives not merely from the internal processing, but also from their relation to the environment and the historical processes through which they were selected for this job. Hence, an externalist like Tye can imagine that what is going on in the head is computation without being committed to methodological solipsism. Nevertheless, according to

externalism, computationalism cannot explain mental content on its own, whereas externalism is not dependent on the mechanisms of the mind being computational.

Externalism is, therefore, a theory of semantic information, the truth of which is independent of the truth of computationalism. Thus, if externalism is true, this fact would seem to only imply that methodological solipsism is false, not that computationalism is false. However, the truth of externalism does affect the nature of what computationalism can be. A computational system must, in the light of the truth of some forms of externalism, be a naturally selected system. This would restrict the class of computational cognitive agents, but not prohibitively so. Presumably, computers that were designed through a process of natural selection, say by use of genetic algorithms to select successful information-processing structures, might fit the bill, depending on whose form of evolutionary externalism one accepts.

As I will demonstrate, it is just these sorts of qualifications that indicate that externalism as formulated by Millikan and Dretske is false. In the hands of externalists, natural selection becomes a magic formula for producing contentful states. According to some externalists (Dretske), it even offers something over and above artificial selection. Yet, there is no important difference between artificial and natural selection. Natural selection is equivalent to a search algorithm through genotype space. If a creature created through this search technique has contentful mental states, an equivalent creature created through another search technique must also. Anything else is just magic. To illustrate this point, I will reformulate the Swampman thought-experiment (Davidson 1986) so that in illustrating that an entity's history is irrelevant to its capacity to produce mental states bearing content, it does not stretch one's imagination or require that the entity be a copy of a naturally selected being

(Dretske's objections, 1995). In doing so, I am agreeing with another externalist, Michael Tye, who accepts that a being without a historical past can have contentful states (1995). Tye's externalism relies on the causal covariance model of relation without necessary reference to how that causal covariance came about.

On the face of it, the causal covariance model of representation leads to panpsychism. The eroded surface of a rock tracks the dripping of water onto it. Tye attempts to avoid this by referring to the informational function of representational structures in the mind, but does not provide an account of what information is. The useful portion of Tye's externalism is merely an outline, with the details needing to be filled in. Tye tries to fill them in with traditional information-processing, computational processes, but this will not do. Computational systems suffer from a host of problems, detailed in the preceding chapter, that indicate that they do not, and quite possibly cannot, track the environment. The appropriate mechanism for Tye's representationalism is outlined in the following chapter. Only self-organizing dynamic systems with the capacity for *re*-presenting internal behavior correlated with the environment are capable of properly tracking it. The final chapter will, among other things, examine whether such a form of externalism can adequately explain phenomenal qualities as Tye purports to have done with his representational framework.

This chapter proceeds by first outlining the motivating intuitions of externalism. I then turn to considering both evolutionary and ecological variations of externalism, principally Dretske, Millikan, and Gibson. The evolutionary accounts fail because they substitute proper or normal function for mechanism when explaining representations. This is akin to explaining how a car drives by reference to the factory procedures that produced it. One can indeed individuate mental states according to whether their underlying structures were nat-

urally selected, but this individuation does not map to a distinction between representational and nonrepresentational states. Finally, I will argue that Tye's externalism must shed both its openness to including evolutionary history in the determinates of representation and notion that computation underlies the internal processes of representations. This will set the stage for incorporating Tye's theory in a more complete theory of mind, one that addresses both representation and mechanism.

5.1 Twin Earth and the Case against Narrow Content

5.1.1 Putnam's Twin Earth Thought-Experiment

Imagine a world that is a perfect duplicate of the Earth in all regards except for one. On Earth, we have a substance consisting of H_2O, which we call 'water'. On the duplicate Earth, or Twin Earth, they have a substance consisting of XYZ, which they call 'water'. Water and twin water appear exactly alike in all respects, the only difference being their molecular composition. It is prior to 1750 on both Earth and Twin Earth, so they have yet to discover Daltonian chemistry, and so do not know anything more about water than how it appears and behaves. Imagine again that your duplicate on Twin Earth thinks the thought 'water is wet'. Is he thinking the same thought as you when you think 'water is wet'?

The answer to this question depends on what one means by 'thinking the same thought'. If, by this, we mean having the same internal state, whether neurochemical or functional (psychofunctional or machine functional), then the answer is yes, because it is assumed that they are microphysical/functional duplicates. But the thought 'water is wet' in your mind does not mean the same thing as the thought 'water is wet' in your twin's mind. The reason is that they refer to different things. To see this, imagine that you are somehow transported

to Twin Earth. You see XYZ and think ‘this is water’. Your twin sees XYZ and also thinks ‘this is water’. The meaning of your thought and your twin’s thought cannot be the same, because the truth value of the proposition contained in your thought is different from that of your twin’s. You are thinking that this is same substance as on Earth. So the reference is different and the difference between what you mean by ‘water’ and what you are viewing on Twin Earth can, in fact, be shown to you.

The upshot of Putnam’s thought-experiment (1975) is that external factors contribute to determining meaning. Therefore, the nature of mental representations, narrowly construed in the fashion of functionalists or identity theorists, does not determine meaning because it does not fix reference. Further, sameness of mental representation, narrowly construed, is not to be understood as sameness of meaning. Methodological solipsists must argue that the reference of a symbol is fixed by the symbol or symbol system itself, and that, as Fodor does argue (1975), that sameness of mental representation is sameness of syntactic structure. The latter is plausible. The former, for reasons discussed in Chapter 2, is not.

5.1.2 Burge on Why Psychology Does Not Need Narrow Content

Tyler Burge (1982) has argued that not only do *de re* attitudes (relational attitudes, such as attitudes about objects) vary in reference, but so do *de dicto* (nonrelational) attitudes. This would suggest that a cognitive psychology would have to abandon the notion of narrow content in general. Just the opposite is true, however, and many cognitive psychologists reject the idea that we must ascribe different mental states to individuals demonstrating the same behavior.

Burge's response to this is that psychology is not a science of behavior so crudely imagined as mere bodily motion. It is also a science of the relations between cognitive agents. Burge argues that because propositional attitudes are among the determinants of behavior, and that Twin Earth thought-experiments demonstrate that propositional attitudes do not supervene on brain states because of their essential reference to the environment, then individualistic states like brain states are not the only determinants of behavior.

To illustrate this proposition, Burge points to Marr's theory of vision. According to Burge, such mechanisms as edge detectors make essential reference to distal stimuli, i.e., essential reference to the environment. Marr's theory as a whole is intended to explain how mind/brains extract reliable information about their environments, and the notion of 'reliable information' cannot be understood without reference to the environment.

The externalist intuitions are therefore twofold. The first is that content is wide—that meaning is determined not only by internal states, but by reference to the environment. The second is that wide content states not only *can* be incorporated into explanations of behavior, they *must* be, and that the best examples of cognitive psychology demonstrate this.

5.2 Making a Fetish of Natural Selection: Evolutionary Approaches

If one accepts externalism, then one accepts the proposition that there is more to representations than their syntactic structure. What are representations, such that they are individuated in part by reference to the environment in which they are tokened? Most externalists accept a causal covariance model of representation. A representation is a state in the mind that reliably tracks an environmental feature. Their occurrences causally covary, which is

to say that their correlation is causally related. An environmental feature helps bring about a token in the mind, and this tokening in the presence of the environmental feature is what tracking consists of. But since this covariance can be thwarted, say, by malfunction of the token-producer, many externalists also append to this specification that causal covariance exists under *optimal conditions* or when *functioning normally*.

These notions imply that there is a purpose for which a device is suited and a set of conditions under which this purpose is fulfilled. The purpose of biological devices is not intentional in nature—no one chooses hearts for their capacity to pump blood—rather it is its function as selected for in the process of natural selection. And optimal conditions are understood as those conditions for which the token-producer was selected to handle. So for externalists with an evolutionary bent, the evolutionary history of representations is essential for their individuation. How these details are spelled out varies from philosopher to philosopher. I will examine two prominent variations on this theme, one which accepts the causal covariance model of representation, Dretske's, and one which rejects this model, Millikan's.

5.2.1 Millikan's Misapplication of Natural Selection

Ruth Millikan (1984, 1993) has worked out a highly sophisticated account of how 'biological function' explains the nature of representation. Her theory, however, rests on several misunderstandings and misapplications of the mechanisms of evolutionary theory, particularly those of natural selection. It is also chauvinistic in its scope, failing to ascribe content to devices that either clearly or possibly possess it, as well as liberal, ascribing content to

widely. To illustrate these failings, it is necessary to examine Millikan's notion of 'biological function' before proceeding to her discussion of content and representation.

5.2.1.1 Millikan's Notion of 'Function'

Biological categories, in Millikan's theory, are not determined by membership in a class of items related by physical structure, causal powers, or dispositions. Rather, membership is determined by possession of the appropriate proper function, understood as the function which an item is supposed to perform or was designed to do. Like any good Darwinian, Millikan understands "design" in terms of natural selection. Thus, malfunctioning kidneys or hearts still have proper functions, even if they are not capable of carrying them out.

Millikan states that an item has a *direct proper function* only as a member of a *reproductively established family*. To define a reproductively established family, we must first understand what a reproduction is. B is a reproduction of A if and only if three conditions obtain. First, B has some determinate properties in common with A. Second, that A and B have these properties can be explained by some natural law or laws operative *in situ* (laws derivable from universal laws under special conditions). Third, for each of the properties in common, the laws that explain why B is like A in this respect correlate determinates (like the color *red*) under a determinable (like *color*) so that whatever determinate A has, B must also. Although A is the model for B, B does not have to be a perfect copy of A, but the range of variation of a property of B must differ if the corresponding property of A differs. Next, reproductively established families are broken down into two types. First-order reproductively established families are sets of items that have similar characteristics as a result of repetitive reproductions of the same characteristic of the same model(s). Examples Millikan cites are specific genes and tokens of a specific word in various media (1984). Higher-

order reproductively established families are defined recursively, with first-order as the base case.

> Higher-order reproductively established families are of three types:
>
> (1) Any set of similar items produced by members of the same reproductively established family, when it is a direct proper function of the family to produce such items and these are all produced in accordance with Normal explanations...
>
> (2) Any set of similar items produced by the same device, when it was one of the proper functions of this device to make later items *match* earlier items, and these items are alike in accordance with a Normal explanation for performance of this function...(Millikan 1984, 24)

Examples are: hearts and kidneys, which are not copied from earlier hearts and kidneys, but rather result from the proper functioning of genes that were copied; any instinctual behavior; and learned behavior when it results from training or trial-and-error due to mechanisms that have as a proper function the reproduction of successful or rewarded behavior. Millikan loosens conditions (1) and (2) with a third:

> (3) If anything x (a) has been produced by a device a direct proper function of which is to produce a member or members of a higher-order reproductively established family R, and (b) is in some respects like Normal members of R because (c) it has been produced in accordance with an explanation that approximates in some (undefined) degree to a Normal explanation for production of members of R, then x is a member of R. (Millikan 1984, 25)

Millikan acknowledges the vagueness of this condition, suggesting it maps to vagueness in assigning biological categories to malformed items.

To define a direct proper function, Millikan first establishes the notion of an *ancestor* of a reproductively established family. There are three ways for an item to be an ancestor of *x*: if it is a member of a first-order reproductively established family from which *x* was derived by reproduction or a series of reproductions; if it is a temporally earlier member of a higher-order reproductively established family and was produced by an ancestor of the

device that produced *x*; if it is a temporally earlier member of the same higher-order reproductively established family as *x* and *x* is similar to it according to a proper function of a producer that produced both.

Now we can define a direct proper function. F is a direct proper function of *x*, with *x* belonging to reproductively established family R in accordance with R's Normal character C, if and only if: ancestors of *x* performed F; the direct causal connection between having C and ancestors of *x* performing F explains why C correlated positively with F over a set of items S that included these ancestors and other things that did not have C; and one explanation that can be given of *x*'s existence is that C correlated positively with F over S, which either caused the reproduction of *x* or explains why R proliferated. To distill this verbiage, we can understand a direct proper function in terms of a simple example. *Enabling people to screw screws* is a direct proper function of a screwdriver, because prior examples of screwdrivers performed this function, which gave them the characteristic of *being useful to screw screws*, and because of this were reproduced.

The notion of a *Normal explanation* keeps cropping up, and it is key to Millikan's account. A Normal explanation is not an explanation in terms of statistical averages. It is an explanation of "how a particular reproductively established family has historically performed a particular proper function" (Millikan 1984, 33). *Normal conditions* for biological items are conditions to which the item is adapted, including both environmental and internal conditions.

Proper functions that produce items bearing relations to the environment or other items are relational in nature. Millikan's example is of the chameleon's pigment-altering mechanism. When a device with a relational proper function produces something specific accord-

ing to its relation with a specific item, the device has acquired what Millikan calls an *adapted proper function*. To continue with Millikan's example, the chameleon's pigment-altering device has the adapted proper function of producing brown pigment in the presence of a brown environment. What is produced in this situation is termed an *adapted device*. Adapted devices have proper functions that are derived from the proper functions of the device producing them, and these proper functions are not merely the adapted proper function but the non-contextual proper function. The proper function of the pigment change to brown is not to blend in with brown specifically, but to blend in with the environment. These are *derived proper functions.*

Finally, the Normal explanation of an adapted or derived proper function derives from the Normal explanation of the relational proper function plus the specification of the particular situation as a Normal condition. Given these definitions, I turn now to how Millikan applies them to understanding mental content.

5.2.1.2 Millikan's Representational Hierarchy

Ruth Millikan has argued that there exists a hierarchy of types of representations—leading from mere reflections of the environment built into creatures to the most sophisticated of thoughts in humans—instead of a simple, discrete distinction between the representational and the nonrepresentational (Millikan 1993). This hierarchy is meant to explain how 'processes of information-transformation' arise.

At the lowest level are what Millikan terms 'tacit suppositions'. Tacit suppositions are those aspects of organisms that are so uniquely suited to a particular environment that they could not function normally outside that environment. The close fit of these organismic fea-

tures to their environment enables observers to deduce the features of the environment. Millikan contends that there are two cases in which these features "represent" the environment.

The first of these cases is the apparent mapping of an organism's features to aspects of the environment. For example, Millikan considers the motor homunculus in the cerebral cortex of humans to be a representation of the human hand (as well as other portions of the body), and the biological clocks of animals to be representations of the length of days. According to Millikan, this type of tacit supposition supports the counterfactual: if the environment were different in a particular dimension, the tacit supposition would be different in its corresponding dimension.

The second type of tacit supposition appears whenever a mechanism is required to implicitly represent aspects of the environment in order for an inferencer to function properly. She defines an inferencer as "a mechanism that, working properly, derives new true representations from old true representations" (Millikan 1993, 98). Millikan gives as an example the edge detectors in the visual system that "presuppose" the features of edges in the external world.

Intentional icons are the next step up in the hierarchy. Whereas tacit suppositions are built into organisms, intentional icons are acquired by organisms. Icons map to the environment according to a 'rule of projection', a mapping of variances in environment to variances in icons. Furthermore, intentional icons only exist where there are mechanisms designed (whether in the evolutionary sense, or by humans) to bring them about (not necessarily by causing them). When these mechanisms are successful, they "bring it about" that the icons to map to the environment. This does not mean that the presence of the feature iconed must cause the icon to occur. Images on the retina are intentional icons in Millikan's

estimation, because the eye lens that produces them evolved for the purpose of establishing a mapping between the world and the image on the retina (even though the actual images are themselves acquired, not evolved). Finally, an intentional icon must be used by an organism for guidance in its environment; there must be a "consumer" of the icon. Some further intentional icons are the dances of honeybees and the magnetosomes[1] of certain bacteria.

Intentional icons differ from Dretske's notion of 'natural information' in that they need not perfectly match the environment, i.e., be perfectly reliable, in order to "icon" the environment. Millikan explicitly rejects the causal covariance model. An intentional icon can fail in the vast majority of its tokenings to guide an organism, yet still map to the environment. In fact, the icon need not map at all and yet still have content, at least for the current consumer:

> the "content" of an intentional icon is described by telling what sort of structure or feature would have to be in the organism's environment, for the icon to map onto by its mapping rule, in order for its consumer to use it successfully in the normal way, that is, the way that historically accounted for the interlocking design of producer, icon, and consumer. (Millikan 1993, 100)

Note the subjunctive mood in the definition. The intentional content of an icon does not consist of a current mapping between icon and world. It must have a mapping that accounts for how it was selected in the past. Nor need the organism actually be guided by the icon in the present: "we should refer not to how the organism is in fact guided, but to the general principles in accordance with which it is *designed* to be guided . . ." (Ibid).

1. organs whose sensitivity to magnetic fields causes the bacteria to move toward geomagnetic north

The dances of the current generation of bees could universally fail to map to locations of nectar, and nonetheless be intentional icons.

From intentional icons, we move to representations proper. Millikan defines representations as those intentional icons that have the function of serving in mediate inferences. Mediate inference is the process of combining an intentional icon with other intentional icons to form new pieces of information. An example provided by Millikan is when an organism combines mental maps of the environment and correlates the information in each map: combining a map of where predators are with one detailing the location of food sources to determine how close predators are to food sources. These maps are not literally pictures of the world; they may be realized, for example, as activation patterns in neural nets. Millikan even holds that the way information is stored in neural nets could be thought of as an inference process, as was noted in the previous chapter:

> Nor is there any reason why the results of superpositional storage of information in neural nets should not be considered to yield conclusions of inference. That superimposed information gets stored correctly in the same net with certain old information clearly depends upon there being an overlap in semantics on some level . . . (Millikan 1993, 104)

Farther up Millikan's ladder of representation, we meet with beliefs. Unlike intentional icons and representations, beliefs (construed as 'sentences in a language of thought') can be negated. Nevertheless, the semantic information that characterizes a belief has its origin in the lower level of 'representations'. Beliefs are, according to Millikan, a subset of 'representations', which are themselves a subset of intentional icons. Accordingly, whether Millikan has provided an adequate account of what information processing is depends on the account she has given for intentional icons and representations. As I will now argue, the inadequacy of this account is revealed by its implausible implications.

5.2.1.3 Millikan's Liberalism

Faced with the question of how signals acquire semanticity, Millikan begins her answer with the implicit assumption that only organisms and artifacts can have contentful states. Tacit suppositions require an underlying design, whether produced by natural selection or human action. Yet, this requirement is insufficient to restrict tacit suppositions to organisms and artifacts. For example, the development of crystals is a selective process, and the design of the crystals enables scientists to read off features of the environment in which the crystals developed. Similarly, a rock positioned with half of its surface exposed to a waterfall will be shaped in such a way that the exposed portions reveal the influence of the water. The light of distant stars provides tacit suppositions about the nature of the stars emitting the light. In fact, it is questionable whether there exists a natural formation that does not bear tacit suppositions revealing facts about the formation's environment.

A further restriction on tacit suppositions is that the absence of the "represented" environmental feature causes them to cease "functioning properly." A rock does not have a function that it does properly or improperly. If the waterfall that had smoothed a rock's surface were to dry up, the smoothed surface of the rock would not cease functioning properly, because there is no sense in which it is functioning at all. Unless, of course, the rock were being used as part of a human artifact. The human intent would then impart design, and hence proper function, to the rock. But this proper function only exists relative to the intentions and actions of an intelligent agent; a human must intend for the rock to have a function and place the rock in an apparatus in which it carries out that function. This could be very simple: just placing the rock beneath the waterfall in order to very roughly estimate the waterfall's characteristics. Suddenly, a rock placed by a human bears a tacit supposition,

whereas the rock that just happened to be in a similar place does not. If this smacks of being ad hoc, that is because it is. And the account only gets stranger when we consider how natural selection, rather than human intervention, creates tacit suppositions.

Tacit suppositions developed in species through natural selection do not have proper functions by design. That the color pattern of the Viceroy butterfly 'mimics' that of the Monarch butterfly is due to a long process of selection, as well as possibly a fortuitous mutation in the Viceroy's line resulting in the resemblance (Millikan 1993, 98). Adaptationists consider the Viceroy's color pattern as having been 'selected for' its resemblance to the Monarch's. This phrase, and the phrase 'designed for', obscure important differences between the intentional activity of humans and the blind process of natural selection. All objects in nature, as well as human artifacts, have a passive design, namely a structural and causal description of them as a system. But only human artifacts are designed in the sense of being created to fulfill a particular function desired by their creators. Thus, what determines the proper functions of organisms's attributes is not that these attributes have a passive design, or that they were created for those functions, but that they contribute to the organisms's ability to stay alive and reproduce. The unanswered question is why this makes them contentful, whereas the attributes contributing to a rock's continuing to be a rock are not.

Intentional icons resemble tacit suppositions in that they map to some feature of the environment, and rely on a notion of proper function to distinguish them from the properties of rocks. A further constraint is that they 'guide' organisms, although

> we need not require that the guided mechanism reside in the same individual as the icon-producing mechanism. The production and "consumption" of the icon may be

> accomplished by any mechanism designed, biologically or some other way, to cooperate in the iconing project. (Millikan 1993, 99)

Here again, Millikan equivocates on the meaning of 'design'. The dances of bees indicating the location of nectar are not designed by the bees in any sense resembling the design of maps by humans.

Millikan also equivocates on the notion of 'to use'. When intentional icons map to more than one environmental feature, the function of the icon is to be determined by what it is used for. The example she gives, however, reveals how little this notion of 'use' relates to what is meant by 'used' in "he used the screwdriver to tighten the screw." She cites Dretske's example of magnetic-field-sensitive bacteria, which are drawn away by magnetosomes from the surface of the water in which they reside, and, thus, away from the oxygen that is toxic to the bacteria. In Millikan's parlance, these bacteria are 'guided' by their magnetosomes; the magnetosomes are 'used' as intentional icons:

> Considered as an intentional icon, the pulling of the magnetosome has just one intentional content. It intentionally icons the more-oxygen/less-oxygen polarity, for being wrong about that is what would guarantee its failure to perform its normal function. (Note that it is not one of its functions, for example, to move the bacterium either down or toward geomagnetic north—any more than it is one of its functions to move it toward true north, toward molten rock or toward snow (the arctic). None of these figures in a causal chain that helps effect its survival. Each is merely a correlate of performing its true biological function.) (Millikan 1993, 101)

It is odd that Millikan would use this as an example of an intentional icon, for presumably the magnetosome is an evolved, not an acquired mechanism. This, however, is not the strangest aspect of Millikan's account. That the pull of the magnetosomes toward geomagnetic north does not figure in a causal chain effecting the survival of the bacteria is just simply false. Moreover, that a simple causal mechanism such as this would be considered

‘guiding the organism’ means that any internal cause of external behavior is a form of ‘guidance’.

Millikan's conflation of the meanings of ‘use’ and ‘cause’ has serious consequences for the interpretation of animal behavior. Suppose we accept the idea that bacteria are ‘using’ magnetosomes to move away from oxygen. The debate whether certain primates use contentful signals then becomes moot, for all acquired modes of dealing with one’s environment must be considered contentful. How would we distinguish between primate tool use and the ‘use’ of hormones produced in response to environmental conditions?

Millikan’s treatment of representations only compounds the difficulties arising from her treatment of intentional icons. Although Millikan intends her notion of representation to accord with the common understanding, her definition of ‘intentional icon’ opens wide the flood-gates for what can count as a representation: “What makes an intentional icon into a representation is that one of its various jobs is to combine with other icons to produce icons carrying new information” (Millikan 1993, 103). In a footnote, Millikan adds that the combination must produce either new information in the form of another intentional icon, or an action based on the new product. Therefore, in addition to overlapping mental maps, secreted hormones that combine their effects must also be considered representations. In fact, every set of molecules that can combine to produce an effect different from the separate effects of the elements of the set, or to produce a new molecule, counts as a representation—assuming, of course, that these molecules perform some function in the body. So Millikan’s concept of representation extends from mental maps in the brain to molecules in the foot. This leads to the curious conclusion that, whereas, according to Millikan, the col-

oration change of the chameleon is not even an intentional icon, the molecular combinations that cause the change are actually representations.

Millikan's absurdities are multiplied further if we also accept the position of molecular Darwinism detailed earlier. Molecular Darwinists hold that the molecular processes leading to the appearance of DNA and RNA are not a product of mere chance combination in the primordial soup; rather, they are a result of natural selection (Eigen 1992). Extending natural selection to the molecular level means extending the notion of normal function to that level as well, and similarly extending Millikan's meaning of intentional icon and representation. The conclusion then must be that pre-organismic molecular strands can have representations. What reason would Millikan have now for not granting that crystals have intentional icons and representations, given that crystals are themselves produced through a form of selection?

Rather than a hierarchical account of how representations arise in organisms, Millikan has produced a form of panpsychism. Molecules have content, and combinations of molecules are representations. The only theoretical barrier between hormones and thoughts is that the latter can be negated. I take this inadvertent panpsychism to indicate a radical failure in Millikan's theory. The question is why Millikan's account collapses so easily into a position she clearly does not wish to advocate.

One source of this failure is her effort to take Dennett's Intentional Stance seriously. One of the tenets of the Intentional Stance is that the distinction between 'original' and 'derived' intentionality is illusory. Searle (1992), who defends the distinction, maintains that humans (and other beings like us) have 'original' intentionality, that is, intentionality not derived from the interpretation of our behavior by some other being. Computers and

similar artifacts, on other hand, do not actually possess intentional states. Instead, their intentionality is derived from human interpretations of their behavior, which describe them as if they possessed intentional states. In contrast, Dennett has long advocated the position that if a system can be described as if it had intentional states, and any other description, such as of its physical make-up or of its design, leaves out something captured by the intentional interpretation, then the system possesses those intentional states in the same sense that humans do (Dennett 1981, 1995). To behave as if one has intentional states is to have intentional states:

> To a first approximation, the intentional strategy consists of treating the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental stages exhibiting what Brentano and others call *intentionality*.
>
> . . . any object—or as I shall say, any *system*—whose behavior is well predicted by this strategy is in the fullest sense of the word a believer. *What it is* to be a true believer is to be an *intentional system*, a system whose behavior is reliably and voluminously predictable via the intentional strategy. (Dennett 1981, 132)

Dennett is not merely saying that intentional systems belong to the class of objects that are intentionally interpretable (the second half of the second paragraph), but that if a system's behavior is intentionally interpretable, then it is a true believer and hence an intentional system. The intentional stance is not merely a useful strategy, it is a strategy that reveals what qualifies as an intentional system. One plausible interpretation of Dennett's formulation is that it is simply a tautology, stating that a system whose behavior is well predicted by the intentional strategy is a system whose behavior is reliably predicted by the strategy. Yet, Dennett (1987) denies being a fictionalist, one who interprets the notion of belief as a useful falsehood; the reason he is not a full-blown realist about propositional attitudes is that he believes that belief/desire psychology will not end up being correct about how the internal states that cause behavior are individuated. Therefore, in Dennett's estimation, all systems

that can be interpreted as intentional systems are true intentional systems. Whereas Dennett specifically refers to beliefs and desires, Millikan extends intentional interpretations to the level of 'icons' (although this is in the spirit of Dennett's explanation of the origin of intelligence: successively dumber homunculi at the lower levels of the hierarchy of intelligence).

The intentional stance, however, is not taken by Dennett to be a definition of what makes something a belief. Dennett is defending a strategy that reveals whether a system possesses beliefs, not an account of what mental states really are, which accounts for the hesitation in his work on the intentional stance to present an alternative to belief/desire psychology. Yet, the latter is what Millikan must provide when she applies the intentional stance to hormones and magnetosomes. An intentional stance toward hormones, however, does not buy the endocrinologist much in the way of explaining their function. In fact, the functions of hormones are not to represent conditions in the environment affecting the body, but rather to regulate the body's behavior. Whereas the exact function of a belief cannot be pinpointed through chemical analysis, the functions of hormones can.

Without intentionalizing hormones and magnetosomes, however, Millikan cannot give a seamless account of the appearance of representations and 'sentences in the language of thought'. Representations are supposed to be just those intentional icons that have as a normal function 'serving in mediate inferences.' But if the icons themselves do not bear content, then how could the fact that they combine to produce new icons possibly impute content to their product? At least Millikan would have to have an account of how the content appeared from this combination. Here, Millikan is as stuck as Dennett when he tries to explain the origin of intelligence via a chain of successively dumber homunculi. Each step

in the chain is a change in degree—except at the point where an entity with no intelligence gives rise to one with just a smattering of it. How this happens remains a perfect mystery in Dennett's writings. To avoid this, Millikan has adopted the apparent strategy of denying such a change in kind occurs. But it is this change in kind that must be explained, if Millikan is to avoid panpsychism.

5.2.1.4 Millikan's Chauvinism

It is important to keep in mind that natural selection does not create anything. It merely determines what will undergo the next phase of mutation and crossover through reproduction. Attributes of organisms might be 'selected for', but they are not 'created for' any particular function. Therefore, there is a clear limiting case for Millikan's notion of tacit suppositions: the initial appearance of an attribute. In fact, it is unclear where this limiting case ends. It is unlikely that a trait will be clearly 'selected for' in the first generation due to the superabundance in nature of factors that could skew the results of the trait's contributions. Natural selection generally takes a long time (or many generations, if you are a virus). Even if such factors did not happen to affect the outcome, there is an epistemological problem determining whether a trait was 'selected for' in only a few generations: One cannot control for all the possible factors affecting selection, and, therefore, cannot discount their possible involvement. So tacit suppositions don't appear until after a mutation + n generations of organisms with that mutation. This is in keeping with Millikan's notion of proper function, which requires that whatever possesses a proper function does so by virtue of being a member of a reproductively established family associated with a characteristic that is correlated with the proper function. It takes time for these correlations to rise above the noise.

Therefore, according to Millikan's theory, newly emergent functions (in the mechanistic sense) cannot have intentional content. This does not imply the possible existence of zombies, creatures with all the trappings of a representational system but without representations.[1] Brains don't arise in a single mutation. But this also does not imply that the newly emergent functions are insignificant. Imagine a mutation that gives rise to primitive edge detectors, or to the ability to produce the first mediate inference.

Not only might the first n generations of an organism be considered lacking a particular representation, but this could conceivably extend throughout the entire history of the organism. Imagine a mutation occurs resulting in a new mechanism in the organism that produces tokens having an abstract mapping to environmental features. After this event, selection pressures are alleviated. Even though the organism uses the token for guidance (it being linked in with its other functions), perhaps only occasionally, the mechanism that produces the token is not selected. It being produced by a dominant gene, it finds expression as it gets passed along to the organism's progeny. Given the lack of selection pressure, it is highly unlikely that the gene will proliferate, and highly likely that it will get bred out. But not necessarily. And throughout its history, it will have no proper function, and therefore no intentional content.

Imagining such a scenario takes less of a leap than imagining Swampman—but let's do so nonetheless. Swampman, as Davidson imagined him, is a microphysical duplicate of, say, Davidson, created by a lightning strike in a swamp (Davidson 1986). He has no evolutionary history, no proper functions. Yet, he behaves exactly like Davidson. In Swamp-

1. Zombies are usually considered to be microphysical duplicates with no qualia. Millikan opens up the possibility of zombies with no representations, let alone qualia.

man's brain is a motor homunculi just like Davidson's. Yet, according to Millikan's theory, it is not a tacit supposition, even though it not only matches Davidson's motor homunculi, which is a tacit supposition, but also maps *in practice* to Swampman's motor functions. Swampman's visual system creates a 2 1/2-D sketch of the world (assuming Marr's theory is correct), which not only is a perfect copy of Davidson's 2 1/2-D sketch, but also enables Swampman to maneuver in his environment. Yet, according to Millikan's theory, it is not a representation. Something is amiss here.

We cannot extend Millikan's theory to cover Swampman by arguing that because Swampman is a duplicate of Davidson, the Normal explanations of Davidson's functions extend to Swampman's. The reason is that Swampman is not a *reproduction* of Davidson in Millikan's sense of the term. There is no causal link from Davidson to Swampman such that whatever properties Davidson has, Swampman must also possess, which then violates the third condition of being a reproduction. Nor are Davidson and Swampman both reproductions from the same model. Therefore, they cannot be members of the same reproductively established family, and so Normal explanations cannot be given for their similarity or extended to Swampman based on similarity.

Nor, given Millikan's theory, can we dismiss Swampman as missing something that is essential *mechanistically or phenomenally* for the production of representations. Had Swampman's history been like Davidson's, the functioning of his mind/brain would produce representations. He's got what it takes, except for his lack of breeding.

Millikan's theory fails to tell us what we want to know: what it is to represent something. How is it that a tokening of some sort "maps" to the world? In one sense, Millikan has a straightforward answer: isomorphism between token and world. Requiring merely

isomorphism leads to panpsychism, so she reduces the set of valid isomorphisms by means of an additional requirement, that of having been selected for. Presumably, this reduces the set of entities capable of representation to the known set of organisms bearing representations. So would a host of other artificial restrictions. For example: only living things with a certain level of complexity have representations.

Millikan's theory does not really explain the notion of representation as it pertains to the understanding of the mind. Instead, Millikan has identified a class of items—things capable of being mapped to the environment—that have contributed to the evolutionary success of organisms. Millikan is right when she says that the meaning of "representation" is not written in stone, but one would hope for more than this.

5.2.1.5 Missing Information: What Millikan Hasn't Told Us about Information

Computationalist approaches to explaining cognition were criticized in Chapter 2 for having failed to develop a theory of semantic information. Millikan's theory would appear to solve this problem by eliminating the need for such a theory, the correlation of tokens and environment being established by the evolutionary history of the tokening mechanism. Millikan, however, reveals that her theory rests on an undefined notion of semantic information when she writes: "What makes an intentional icon into a representation is that one of its various jobs is to combine with *other* icons to produces icons carrying *new* information" (Millikan 1993, 103). Here again, we have the elusive notion of a token *carrying* information. This is not meant to say that Millikan could not provide a theory of semantic information, only that she, like the computationalists, must.

5.2.1.6 Conceptual Problems with Millikan's Account

Natural selection must stamp its imprimatur on a "mappable" token in order for such a token to be a representation. Why? What is it about natural selection that not only establishes that something is a representation, i.e., enables us to read the historical record of the token and understand what it *was* used for and therefore *is* used for, but also grants representational status? A clue can be garnered from Millikan's understanding of how natural selection works:

> economy dictates that traits serving no purpose are highly likely to disappear, not necessarily because they get in the way, but because that section of the genetic code *could* serve a useful function if coded otherwise and sooner or later Nature will stumble on this discovery. (Millikan 1984, 27)

Natural selection is not really a messy process that often fails to clean up after itself. Rather, it is guided by a principle of economy that guarantees that, sooner or later, what is not useful will disappear. It is this guarantee that lends natural selection its special powers. Otherwise, the link Millikan posits between token and environment is broken; it becomes merely isomorphism, which she recognizes is not sufficient. All functions that are reproductively established are useful and all useful functions are reproductively established.

Millikan is wrong, however, in her characterization of natural selection. That natural selection is parsimonious in its use of the genetic code is contradicted by junk DNA—DNA that serves no useful purpose, and, in fact, constitutes the majority of DNA. Dawkins (1976) has offered one explanation for the existence of junk DNA, namely that it is a free-loader that serves its own purpose, having been adapted to ride along with useful DNA. That natural selection cleans up after itself is quite apparently false. No one is expecting the demise of the appendix. Vestigial organs are not uncommon. The idea that natural selection eventually stumbles on inventions shows an even deeper misunderstanding.

As Millikan acknowledges, natural selection does not create diversity, but only operates on a given pool of genes. Evolution is dependent largely on mutations to feed the gene pool, although such factors as random drift and pleiotropy contribute. Mutations are unbiased in the sense that they do not favor producing useful traits. Quite the contrary, they usually produce useless or even harmful traits. Suppose, however, we were given infinite time to cycle through all the possible mutations. If evolution were like this, then, yes, in all probability we would hit upon a useful mutation. Evolution, however, is a constrained search through gene space. It does not have the luxury of reengineering something from scratch, but rather must apply, through mutation, incremental modifications to existing genetic traits. If it takes a particular path in gene space, it may have closed off another possibility; it is not an exhaustive search of gene space. There is no guarantee that it will hit upon a use for a particular section of genetic code.

Not only is the link between evolutionary history and representation irrelevant with regard to explaining what representations are, but the link cannot even be established as Millikan wishes. Perhaps causal covariance can explain what 'being selected for' cannot.

5.2.2 Dretske's Representationalism

In Chapter 2, I argued that Dretske's causal covariance model of semantic information was inadequate, because, in part, the requirement for causal covariance to obtain was too strict, and as such, his theory had difficulty explaining misrepresentation. In *Naturalizing the Mind* (1995), Dretske more explicitly weds his causal covariance model with teleological concepts from Darwinism. He retains the notion of natural information—tokens are informational because they track environmental features by virtue of causal relations with them.

These causal relations, however, are due to design. Tokening mechanisms are selected for their ability to token reliably in the presence of the represented object.

If causal covariance were all there was to the story, then we would be stuck with the problem of misrepresentation. Teleological explanations, however, allow Dretske to handle this problem. Tokenings are causally correlated with the presence of environmental features under what Millikan would call Normal conditions. Misfirings of tokens result under abnormal conditions, and so do not break the causal covariance. Dretske expresses this by saying that tokens and environmental features *naturally* covary.

Dretske's later theory is twofold:

1. All mental facts are representational facts.

2. All representational facts are facts about information functions. (Dretske 1995, 1)

Stated otherwise: representation is causal covariance under natural conditions, and all mental states are representational states. Dretske is arguing both for externalism, and attempting to show how externalism can yield a complete theory of mind, one that encompasses experiences and qualia. Although I will argue that Dretske's version of externalism falls prey to many of the arguments against Millikan, his notion of explaining experience in terms of representation has merit in itself.

5.2.2.1 What Is a Representation?

Dretske's formal definition of what it means for a system to represent something is as follows:

> A system, S, represents a property, F, if and only if S has the function of indicating (providing information about) the F of a certain domain of objects. The way S performs its function (when it performs it) is by occupying different states s_1, s_2, . . . s_n, corresponding to the different determinate values f_1, f_2, . . . f_n, of F. (Dretske 1995, 2)

This definition encompasses the two aspects of Dretske's notion of representation: causal covariance ("indicating" or "providing information") and having been selected for this ("function of").

There are two kinds of representations. The first kind, systemic, has its indicator function by virtue of the system of which it is a state. It indicates because the system is so constructed as to produce this state in the given situation. Dretske's example is the state of a thermometer indicating that it is 32° F. The constituents of the thermometer will produce a distinct behavior when its environment is 32° F. The second kind, type representation, corresponds to the label placed on the thermometer, such as '32' or 'Freezing'. The first is hardwired, the second acquired. They correspond to phenomenal and conceptual representations, respectively. This is the crux of Representationalism: all phenomenal experiences are actually representations of the first sort. They are all sensory in nature, and have their content determined by the biological function (understood teleologically) of the sensory organs producing them.

Phenomenal experiences have the job of representing properties of objects, not that there are objects possessing these properties. An experience of redness indicates the presence of redness, that *this* is red, but does not indicate that *this apple* is red. Misrepresenting occurs when there is no *this* in the world corresponding to the experience's indication of a property, or when one is wrong about *this*, i.e., *this* is blue, not red. Discriminating that something is a particular shade of red, on the other hand, requires the concept *red*. Organisms that do not have concepts would still be able to experience different shades of red, and, presumably, react differently to them, but could not form the representation that 'X is red'.

Introspection, according to Dretske, is not an inner sense directed toward experience. Instead, it is a conceptual representation of experiences (sensory representations) as representations: "If E is an experience (sensory representation) of blue, then introspective knowledge of the experience is a conceptual representation of it as an experience of blue (as a color)" (Dretske 1995, 44). The act of introspection is not an additional experience that one has of sensory experiences: "One can, by introspection, come to know about experience, but the knowledge is obtained without any experiences beyond the ones one comes to know about" (Ibid, 63). Presumably, the act of conceptualization is not an experience that can be introspected, because it is not a sensory act that produces experiences. There is nothing like what it is like to conceptualize, no associated qualia.

Dretske's Representationalism denies to qualia the privileged access that most philosophers have come to associate with the concept. According to Dretske, qualia are simply the properties that objects are sensorially represented as having. There are two aspects of how something, A, looks phenomenally to someone, S, when A is an X:

1. A looks to S the way Xs normally look to S.

2. A looks different to S from other As.

These aspects are knowable by others. Knowing the normal functions of other creatures' sensory systems and knowing the actual discriminatory abilities of these systems tells us what the creatures' experience are like. There are a number of important consequences of these stipulations. First, if one cannot discriminate between different kinds of As, nothing can look to one as an A. Second, changing the discriminatory capability of an organism does change the organism's qualia, but it does not change it into qualia that are like those of creatures with similar discriminatory abilities but whose normal function is not similar

to the normal functioning of first organism. Dretske's example is of a speedometer that can register floating point values, such as 77.75 or 78.0 m.p.h., compared to one that only registers discrete steps, say from 77 m.p.h. to 78 m.p.h. If damage causes the first speedometer to no longer be able register increments of less than 1.0 m.p.h., its experience of 78.0 m.p.h. is nonetheless not the same as the second speedometer's experience of 78 m.p.h. The damaged speedometer retains its function of providing more precise information than the first one in spite of the damage to it.

Dretske's representational theory of qualia conflicts with the strongly-held intuition that we cannot know 'what it is like' to be a creature with sharply different experiences from ours. Frank Jackson's (1986) thought-experiment about a scientist, Mary, who has no direct experience of color, is one expression of this intuition. Jackson imagines a scientist who has grown up in isolation from colors, but who knows everything there is to know about the objective aspects of color and experiencing color—the various wavelengths associated with it, the nature of reflectance and illumination, how the brain creates representations of color, etc. Jackson intuits that this scientist is still missing information about color, namely, what it is like to experience it. The point of this argument is to show that experiences cannot be reduced to objective properties like brain states. Knowing everything there is to know objectively cannot tell us what it is like to experience color, or so the argument goes.

Dretske's argument against this intuition is that one can extrapolate from one's own experience and the given objective knowledge to conceive of what it is like experience what others experience. Dretske imagines Mary trying to understand what it is like for a dogfish to experience through electric sense (its ability to sense electric fields). According to Rep-

resentationalism, what the dogfish represents in its experience is not an electric field, but rather a geometric configuration—the information that the electric field provides through the dogfish's electric sense. If Mary knows the shape that the field is represented as, and can test to find the discriminatory abilities of the dogfish, then she knows everything that there is to know about a dogfish's experience of electric fields. The key to the argument is that the dogfish does not represent electric fields *as* electric fields.

Dretske acknowledges that there is one thing that Mary cannot do that dogfish can: form the conceptual knowledge about their sensory representation that *this* has pattern P. Mary cannot introspect a dogfish's experience, but, nonetheless, she knows all there is to know about what the dogfish is experiencing. This is because introspection is not itself experience. So while there is something the dogfish can do that Mary cannot, there is no experience associated with this action. This is odd, to say the least.

Consider pains. Dretske contends that pains are representations of bodily states. Pain is consciousness of the bodily state that the pain represents. A pricking feeling is a representation of being pricked. Phantom pains are representations of states for which no *this* exists. It is possible to know what pains in other organisms are like through the usual methods. But what is it like to have a pain in a poison sac? Well, it is like having a pain that indicates X about your poison sac. But I don't have a poison sac. So, it is like having a pain in something that is like having such and such connections to your body and being filled with poison. So now I know everything I need to know about what it is like to have a pain in a poison sac. Except, for some reason, I can't tell whether my abdominal pain is different from a pain in a poison sac. I can't seem to make this discrimination. This is an intuition about what we can know, and it conflicts with Dretske's. According to Dretske, I can know what it is like

to have a pain in my poison sac if I can know what such a pain would indicate—burning, tearing, poking, pricking. But what if a pain in a poison sac is not like any of these. What if there are quite different connections in a poison sac bearer's body from those in mine, and these connections cause a different kind of pain. How would I know what this pain is like? Dretske must assume that such cases are impossible.

Can we even know what a burning sensation in the abdomen is like without having experienced burning sensations? A burning sensation does not indicate burning in the body. Nor does it indicate a particular pathology, as this is conceptual knowledge. How does one discriminate between a burning sensation and a pricking sensation if one has not experienced either? Imagine Mary grew up in a padded world and had never been pricked. Could she know what it is like for a dogfish to be pricked? If not, can externalism be true without the objective knowability of experience?

5.2.2.2 The Inaccessibility Defense of Phenomenal Externalism

In his presentation of Representationalism, Dretske takes as granted Twin Earth considerations about the nature of thought, and assumes that thoughts do not supervene on brain states. Dretske's Representationalism is a theory about the nature of sensations, of experiences, that they are representational in character. Even among Externalists, the notion that sensations are external in nature is generally rejected. Dretske contends that if one accepts externalism about thought, there is no reason to reject externalism about experience. His argument hinges, ironically enough, on the inaccessibility of certain aspects of experience.

Take the usual suspects from Twin Earth examples, Fred and Twin Fred in this case, but have them look at the same stuff. Fred thinks it is flim. Twin Fred thinks it is flam. The

concepts they apply to their experiences of the stuff are different, so the stuff will seem, in the doxastic sense of 'how they believe it to look', different to them. Fred will believe it looks like flim, whereas Twin Fred will believe it looks like flam. How it looks phenomenally to each of them, however, is a different matter.

While it is possible that the stuff looks phenomenally the same to Fred and Twin Fred, Dretske argues that they cannot access this fact: "The access one has to the qualities of one's experience . . . is only through the concepts one has for having thoughts about experience" (Dretske 1995, 134). Consider someone who has no notion of a change of key. When listening to a symphony, they will hear a change of key, but they will not be aware of this. If asked whether they heard a change of key, they will answer no. If Fred is asked whether the stuff looks like flam to him, he will also answer no. So far so good; change of key is a concept that one has to acquire. But Dretske goes further than this. He argues that the inability to apply the concept reflects a lack of discriminatory capacity, not just a lack of the right label. This is no doubt true for identifying a change of key. Dretske, however, argues that this applies to such rudimentary discriminatory abilities as taste. That one can distinguish between Coke and Pepsi, without having the labels or concepts 'Coke' and 'Pepsi' in one's repertoire, indicate that one has some sort of conceptual scheme (taste in this case) by which to discriminate them.

There is one more leap that we need to accept in order to finish Dretske's argument. We must accept that being unaware of quality Q, the quality of 'sameness', about experience X, as Fred and Twin Fred are, implies that experience X does not appear to have quality Q: "From a subjective standpoint, it will be as if their experience . . . was not Q" (Ibid, 140).

Dretske's conclusion is that because there is no way that they can be aware of the sameness of experience, we shouldn't suppose it is the same.

This is meant to establish externalism about experience by offering a palatable disjunction. If one accepts that qualia are essentially knowable, then the only way to know them is through beliefs, and beliefs are externally grounded. On the other hand, if one denies the knowability of qualia, then we needn't conclude that qualia are necessarily the same in microphysical duplicates when they experience the same thing.

5.2.2.3 The Inaccessibility Defense Cannot Be Made A Priori

There are two important claims embedded in Dretske's inaccessibility argument. The first is that all discrimination is conceptual. The second is that acquisition of conceptual knowledge alters how experiences seem in what is for Dretske the only relevant sense of seem, the doxastic. Each of these is problematic as a universal judgment.

Imagine again our concert-goer who does not know what a change of key is. He hears the same symphony three times, but before the third time, someone explains to him what a change of key is and what to listen for. After the second time, he remarks on how much a certain passage sounded like the same passage in the first production. His memory of the how the first symphony sounded is like the experience of the second. Then he learns what a change of key is and goes back to hear the symphony again. That same passage in fact includes a change of key, and this he is able to recognize. He is also able to recognize that what he heard the first two times was a change of key, because he recognizes the phenomenal similarity of the three acoustic events.

Dretske must argue that this underlying phenomenal similarity that he seems to recognize is due to application of some manner of concept (not change of key) to all three. Perhaps, but perhaps not. Even if we concede this is true for the concert-goer, as a universal principle it seems specious. Must all "judgments" of similarity be conceptual? This is to argue that categorization is the same as conceptualization, a dubious proposition given results in cognitive psychology on preconceptual categorization. Or it is to expand the notion of conceptualization to include any form of categorization. This is to deny that categorization can be a perceptual experience. These are empirical questions, to be sorted out by psychology. Whether perceptual categorization is a direct experience is a matter of research. At the very least, hardwired discriminatory behavior is provably possible—robots that chase after orange balls display it.

The second claim, that acquisition of conceptual knowledge alters our experience, follows from Dretske's claim that conceptual knowledge is required for an awareness of experience. If we change how we are aware of sensation, we change how that sensation seems to us. Again, this is an empirical proposition, and likely to be false as a universal judgment. Certainly it is true that acquiring musical concepts changes, in one sense, how we hear a piece of music. We have a greater appreciation of the music, understand how it fits together, perhaps hear what we considered disjointed pieces to be clever combinations. But we do not lose the ability to identify memories of music acquired prior to this acquisition with what we hear now. What we hear now may seem richer, but there is a phenomenal core that we can identify regardless of these concepts. And if the first claim is false, then there is possibly an identical core that we can identify sans concepts. This, however, is something for

cognitive psychology to uncover by measuring and comparing reported, instinctive, and acquired responses, *as well as by looking at neural responses.*

5.2.2.4 Defending Externalism against Epiphenomenalism

The charge of epiphenomenalism is the charge that externalism implies that the mental is irrelevant *causally* for the production of behavior. If causality of behavior is a local activity of intrinsic properties, properties residing in the mind/brain, then historical differences and environmental conditions must be irrelevant to explaining behavior because they are extrinsic factors. If mental states do not supervene on brain states, instead being in part determined by evolutionary history and environmental conditions, as externalism contends, then mental states are causally irrelevant for the production of behavior. Hence, they are epiphenomena.

To counter this charge, Dretske argues that historical causes can and are understood as proximate causes of behavior. When we explain the reason for a brain-damaged individual's behavior, we often cite the reason for his brain damage, such as a brick having fallen on his head. More importantly, the selected functions of aspects of evolved organisms are generally cited as reasons for their behavior. Dretske illustrates this with a variation on Swampman. There is a plant, the Scarlet Gilia, that changes color during the summer in order to attract pollinators. Imagine a duplicate of the Scarlet Gilia, Twin Plant, with the same behavior, but having evolved this behavior for different reasons (to repel pests). Imagine a third plant, Swamp Plant, that spontaneously generated. There are no evolutionary reasons for its behavior. The behavior of each plant is explained differently. The Scarlet Gilia and Twin Plant are given their respective evolutionary explanations, whereas Swamp Plant must be explained in purely mechanistic terms.

If historical considerations can play a part in causal explanation, then that mental states do not supervene on brain states does not imply that they are causally inefficacious. Granted. But this is not the danger that Swampman poses to the externalist. Dretske is right when he asserts that there is no metaphysical problem with having historical conditions figure into the causes of a behavior. There is a problem, however, when we *require* historical conditions to figure into the cause of behavior.

For example, although Swamp Plant has no evolutionary history, it does have an organizing principle. The photosynthesis that it carries out has the function (high-level mechanistic) of sustaining the organism. Without it, Swamp Plant would die. That it succeeds in photosynthesizing sugars and other important chemicals has consequences. Among these are that it continues to absorb moisture, build cell structure, and produce pollen. Similarly, Swampman can digest, walk, see, and hear. His ingestion of nutrients is caused by the same processes that cause ours, hunger. An externalist might contend that digestion is only digestion in the presence of a purpose, in the evolutionary sense of 'purpose'. This is, however, to deny a middle ground between defining function in terms of input-output and evolutionary purpose. This middle ground is the system, or organismic, level, an idea that I will explicate in the following two chapters. Natural selection is just one way to explain the relation between organism function and environment. There is, however, no metaphysical reason why it should be the only way. If it is possible that Swampman can have the same experiences as you or I, then Phenomenal Externalism is wrong about experiences.

The wild improbability of Swampman might tempt some to think that this case can be ignored. This temptation can be easily overcome. With the development of powerful computers and robotic instruments, it is not implausible to conceive of situation in which com-

puters manufacture themselves by the billions. With the increasing knowledge in the field of neuroscience, it is plausible that brains could be broken down into their constituents at the neuronal level. Using our robotic assemblers, construct billions of robotic lab assistants whose job is to randomly put together the constituents of a brain. Each robotic lab assistant gets a subset of the possible combinations to construct, so the task is parallelized across billions of systems. The number of neurons in the human brain is on the order of 10^{10}, give or take an order of magnitude. The number of possible configurations of these neurons is staggering, but given the rules of chemistry, the number is reduced somewhat (the number of connections in the human brain is 10^{14}). Computers that can compute in the Teraflop range (one trillion, or 10^{9}, floating point operations per second) already exist, and robotic arms can operate with frightening speed. Going through all of the permutations of neuronal configurations would take a substantial amount of time; it may even be prohibitively so. If it is, we will set our 1 billion robots to constructing a new generation of robots to do the work (say each robot makes 10,000 more over its life, now we have 10 trillion robots). The probability of assembling a working brain now becomes much greater than Swampman's spontaneous generation. If the robots hit upon a working brain, this brain will not have been selected for, because the assembly process was an exhaustive one. It will not have an evolutionary history. If placed in a human body with all the appropriate connections setup, on what basis would we say that it does not experience? If we prick its body, does it not wince?

Dretske argues that psychology should not be reduced to the study of behavior as mere physical motion, and allows that the physical motion of Swampman is equivalent to its duplicate. So why is mental behavior different? Swampman runs, eats, and defecates—so

why doesn't he experience? Rather than treating experiencing as behavior, Dretske treats it as an organ with a teleological function.

Dretske hasn't given one good reason to overcome the intuition that Swampman doesn't have any experiences. He has given several bad ones. This is in the nature of evolutionary externalism.

5.2.3 Why Evolutionary Externalism Doesn't Work

Recently, robotics researchers at Case Western University developed a system to construct simple robots with configurations selected for by application of genetic algorithms. The more successful of these robots were capable of locomotion. If another set of researchers had randomly pieced together similar robots, which, given the simplicity of these robots, is plausible, would anyone argue that these robots are not capable of locomotion as they move themselves across a floor?

Evolutionary externalism seeks to carve out a special exception for representations. Only information gathering devices that have been selected for this purpose are truly information gatherers. An evolutionary past is a necessary condition for representing, though not for running, eating, or even digesting. Evolutionary externalists would even disqualify those functions that had appeared through a process other than natural selection. Dretske even holds that items chosen by artificial selection don't pass muster:

> Natural selection is quite different. Unlike artificial selection, an item cannot be naturally selected to do X unless it actually does X. It has to do X because the way it gets selected is by having its performance of X contribute in some way to the survival and reproductive success of animals in which it occurs. It is this contribution to reproductive success that, when it is selected for, confers a function on a system, and the function it confers is doing what the system *did* (in this case, provided information) the increased fitness. (Dretske 1995, 165)

A mechanistic function must contribute to the fitness of an organism on average to be a teleological function. But *functioning on average* does not tell us what something does. Suppose an antelope sees a lion. Its visual system has the function of warning the antelope that lions are near. Unfortunately, the lion is too quick in this particular instance, and the antelope is killed in spite of the proper functioning of its visual system. On average, its visual system gives it an advantage, but not today. This is how natural selection works; its notion of average is statistical, not Normal. There are no Normal conditions under which an antelope's visual system works, no conditions in which it always enables the antelope to escape. Nature does not care if there is a subset of conditions under which a function always works. All that needs to happen for selection to occur is a slight advantage being conferred.

We can illustrate this point as follows. Suppose a mutation occurs and an animal that otherwise could not detect lions now suddenly can 30% of the time. This is not because 30% of the situations it finds itself in are Normal conditions, but that its sensor goes off 30% of the time in the presence of lions. It is not very reliable, but 30% of the time it will help the animal escape. This animal now has an advantage over his compatriots, and this advantage will likely be selected for. But this sensor has no Normal conditions under which it works reliably.

Selection works over populations of animals, and therefore over the host of conditions that they encounter, and over time, and so the host of conditions that the generations encounter. These conditions do not even constitute the set of conditions under which an item will work, or the set under which it will not work. They are the conditions over which an advantage was gleaned. Presumably, there is a causal connection between the function developed and the nature of the conditions. But it is not a perfect relation. Thus, we cannot

say that under such and such Normal conditions, X has the function of providing natural information. Rather, we must say that over the set of conditions that an organism faced, X provided enough indication about the state of affairs that it gave an advantage. The connection between reliability of function or information and natural selection is much weaker than Dretske purports, though existent.

This becomes clear when one considers the other aspects of functioning that are relevant to providing an advantage. It makes no sense to have an organ that reliably provides information if it functions too slowly or if it is cumbersome to bear. Natural selection includes pressures other than fidelity of representation. Millikan sidesteps this by identifying the informational mapping with an isomorphism between representation and world, rather than the causal correlation used by Dretske. As van Gelder's example of the Watt governor illustrates, however, there may not be an isomorphism between system state and world, but rather a statistical correlation. Variations in the world may not always map to variations in representation, and specifying 'under Normal conditions' does not change this.

Ultimately, the failing of evolutionary externalism is that it looks at the wrong level to understand function. Natural selection answers the question of why an organism has the characteristics it does, not what these characteristics do. To answer this question, we must turn to the relation between organism and environment. Evolutionary history is not irrelevant to this, but is also not essential. It explains how an organism's species has come to have its relation to the environment *for naturally selected organisms*, not what that relationship is. Representation is, ultimately, a relation between organism and environment. As such, we need not refer to evolutionary history to understand it. Unlike Millikan, we should be

interested in figuring out what representations *are* rather than supplying a scheme for classifying them.

5.3 The Organism in its Environment: Gibson's Ecological Approach

In *The Ecological Approach to Visual Perception* (1979), J. J. Gibson argued against the physics-based approach to understanding visual perception and in favor of what he termed the 'ecological approach'. He offered an alternative theory to the standard understanding of information as composed of atomic elements gathered into meaningful structures by the brain, instead contending that light striking the eye already contains direct information about complex objects in the environment, as well as information about their value to the perceiver.

When pursuing a physics-based approach, one attempts to analyze visual perception in terms of the effects that photons have when they strike the retina. It is a bottom-up approach, building visual perception from the smallest of building blocks: the responses of single cells to photons striking them. Edges, surfaces, and depth are all perceived, according to the physics-based approach, by composing the information provided by these atoms of visual perception. Value is attached to the perceived object by the perceiver; it is not perceived.

In following the ecological approach, one considers surfaces to be directly perceived, not constructed in the perceptual system out of bits of information. According to Gibson, light reaching the eye carries information about surfaces themselves. Not only are surfaces directly perceived, but what these surfaces afford are directly perceived as well. An affor-

dance is what an aspect of the environment "offers the animal, what it provides or furnishes, either for good or ill" (Gibson 1979, 127). The ambient light is neither merely a signal nor a symbol to be interpreted. It bears semantic information that is directly perceived without the mediation of inferences or deductions within a symbol system.

Critics of Gibson have argued that he denies a place for information processing, or even for cognition, in the processes of visual perception. His defenders (Reed for one 1988) have contended that Gibson's theory shows how the environment augments the internal processes of the mind/brain, so that information processing can no longer be understood in terms of just internal factors. The environment provides structured information to the perceiver, reducing the amount of processing necessary for perception of complex entities. While there are aspects of Gibson's theory that suggest the more radical reading, such as his theory of affordances, the more moderate reading has much to suggest for externalism. I will examine each aspect separately.

5.3.1 How the Environment Augments Visual Perception

As Gibson and his defenders understand it, the physics-based approach to understanding visual perception posits a retinal image produced by incident light being transduced by receptors as the only source of information for visual perception. From this retinal image, the visual cortex and other portions of the brain must perform massive calculations to affect perception of 3D objects. One such calculation is the solution of the correspondence problem for stereopsis. Each eye produces a retinal image; to compute depth from these retinal images, the portions of the images that correspond to each other must be determined (hence, correspondence problem). This is computationally expensive, as computer implementa-

tions of this process have shown. Once correspondence is determined, disparity between points in the images can be calculated, and disparity can be translated into depth.

The ecological approach suggests that much of what the physics-based approach requires of the visual cortex is not necessary. The ambient light carries structured information about how the environment is configured, and the mind/brain often only needs to receive this information. Ambient light is light that converges on a point, as opposed to radiant light, which is light dispersed from a point. Every point in the environment is a point of convergence for light. The ambient light arriving at a point is structured in that, because it is light reflected from objects, it has solid angles of similar intensity and similar mixtures of wavelength, separated by intensity and wavelength distribution step changes. A solid angle is the somewhat homogeneous light with a particular mixture of wavelengths incoming from an object. This set of solid angles corresponding to objects and the space between them forms the optic array. Greatly homogeneous regions indicate spaces between objects. It is possible that ambient light has no structure, although this limiting case would be quite rare. Gibson suggests a dense fog could create such a situation.

Differentiation within a solid angle indicates further information about the object. For example, patterns of wavelengths offer information about the texture of the object. The structure of the optic array itself indicates the configuration of objects in the environment. Therefore, it is not necessary for the brain to engage in complex computations to create this information from the retinal image. The ambient light already contains these relationships. Much of the brain's work is offloaded, as Clark would put it.

Perception, however, is not a passive process. Motion is essential to visual perception, according to Gibson. What motion allows for is change of perspective, and change of per-

spective enables comparisons between optic arrays. Invariants within a set of optic arrays can be extracted, and these give vital information to the organism about the objects in its environment. Although Gibson is primarily concerned with visual perception, other sensory modalities come into play when an organism moves about its environment. For example, the semicircular canals in the ear offer information about the orientation of the head, which can be integrated with information from vision to get a better sketch of what is invariant in one's experience. The organism does not take uncorrelated snapshots that it processes independently. Rather it samples its environment, integrating information that the environment offers.

The information that the ambient light carries is "information" by virtue of the lawlike correspondences between the structure of the ambient light and the structure of the environment. Because of this unique mapping of particular optic array structure and environmental configuration, the organism is aware directly of how the environment is constituted.

What does this imply about a theory of semantic information? As Rowlands (1995) points out, information can no longer be understood in terms of Shannon information theory, or in terms exclusively intrinsic to the organism. To understand the information that the organism has, we must understand the information available in the optic array. What the optic array is *about* is contained within the optic array's structure.

There must exist a unique correspondence between the structure of the optic array and the structure of the environment for the perception of this information to be direct. Otherwise, the organism would have to deduce from other cues and internal processing what the array corresponds to in this particular situation. Just what the right formula is for how much the environment offers and how much is deduced from the information is up to empirical

science. *That* the optic array carries any of this information, however, is sufficient to rule out methodologically solipsistic theories of mind.

This theory is similar to Dretske's notion of natural information in two respects: the correlation between environment and that which carries the information must be reliable, and this correlation is causal. The difference lies in the relatio of the correlation. For Dretske, tokens in the mind/brain are correlated causally with aspects of the environment. For Gibson, the optic array is correlated causally with the environment's structure. Dretske's externalism, however, does not offload information processing as Gibson's does.

5.3.1.1 How the Environment Augments Cognition

The theory of information that flows from Gibson's ecological theory of perception can be easily applied to aspects of cognition in general. The idea is that external objects assist cognitive processes as place markers, shortcuts, cues, and promptings. Rowlands (1995) presents two possible extensions of Gibson's theory: memory and mental calculation.

Sights, sounds, and smells all help trigger memories. A Gibsonian views this fact as the environment being an adjunct store for memory. In some cases, it is intentional, such as when someone marks a spot to return to. In other cases, the environment offers cues to memory without active intervention of the organism. A particular spacing and type of trees might indicate where one has stored a nut.

Similarly, we can use external items to help us calculate. When we engage in calculations that exhaust our mental resources, we use symbols marked on paper to help us keep place. The action of writing and manipulating these symbols is itself information process-

ing. Thus, Rowlands contends that our acts of memory and calculation are hybrids of internal and external information processing.

The principles expounded by Gibson in his theory of the perception of surfaces and objects and as extended by others to cognition contain a reasonable balance of the external and internal. Gibson's theory of affordances, however, robs the mind/brain of functioning that is not reasonably ascribed to the environment.

5.3.2 The Radical Gibson: The Theory of Affordances

A number of authors, including Clark (1997), van Gelder (1995), Kelso (1997), and Edelman (1992), have suggested links between their repudiations or qualifications of computationalism and Gibson's ecological approach. In each case, the connection has been tenuous at best, supported only by the fact that Gibson and these authors reject locating meanings exclusively in the head. Gibson, however, went much farther than this:

> The theory of affordances is a radical departure from existing theories of value and meaning. It begins with a new definition of what value and meaning are. The perceiving of an affordance is not a process of perceiving a value-free physical object to which meaning is somehow added in a way that no one has been able to agree upon; it is a process of perceiving a value-rich ecological object. Any substance, any surface, any layout has some affordance for benefit or injury to someone. Physics may be value-free, but ecology is not. (Gibson 1979, 140)

Both Edelman and Gibson rely on a notion of intrinsic value to establish their theories of meaning, but they locate value in opposite places. Gibson considered the environment to be value-laden, whereas Edelman (1987, 1989) identifies value in the structures of the brain (although this value is established through processes that interact with the environment: evolution and development). At best, Edelman's theory may be seen as an explanation of how what seems to be direct perception actually occurs. This point will be developed fur-

ther in the next chapter. Clark attempts to enhance computationalism by using short cuts and constraints offered by the environment in order to fill out the meanings found in the head, but still considers these meanings as value-additions to the incoming signals from the environment.

Just how radical Gibson's theory is can be seen in his extension of affordances beyond what surfaces provide to the animal. It is a radical step from claiming that surfaces can be directly perceived to claiming that the use of these surfaces can also be directly perceived. All the more radical is the further step that all affordances can be directly perceived, even those which do not seem to have immediate connection with visual perception. Thus, Gibson argued that how a thing tastes can be directly perceived. This is possible because "a unique combination of invariants, a compound invariant, is just another invariant" (Gibson 1979, 141), and the taste of a thing is presumably just such a compound invariant. So "if the visual system is capable of extracting invariants from a changing optic array, there is no reason why it should not extract invariants that seem to us highly complex" (Gibson 1979, 141).

Gibson's theory identifies the semanticity of a signal within the signal itself. The problem of how semantic information arises from the syntactic operations of the mind/brain is traded in for the twin problems of how semantic information exists in a signal and how this information is perceived.

5.4 Problems with Direct Realism

Gibson's theory of affordances locates the value and meaning of signals externally to the perceiving agent. A given signal, an array of light in the case of visual perception, carries

information about what its source affords to the perceiver. But this information is different for different perceivers, because what an object affords is not the same for every possible perceiver. A piece of cheese affords nourishment for a human, but not for an adult tiger. How does this differentiation in semantic information occur? Gibson states that an affordance points in two directions, to the environment and to the individual. Perception is not merely perception of the external world, but is at the same time perception of one's self as well. A human perceiving a piece of cheese is also perceiving his relation to the cheese: his own tastes and need for nourishment.

While this blunts the obvious criticism of Gibson's theory of affordances, namely that meaning cannot be external to the individual because the meanings of signals are ambiguous, it also blunts the supposed radicalism of his thesis. Meaning cannot be exclusively external to an individual, because an internal sense is required to establish what an object affords. Therefore, information in ambient light does not specify what an object affords. In what sense, then, is the perception of affordances direct?

Gibson's theory founders when called upon to explain how learning occurs. As Gibson acknowledged, animals are not born capable of perceiving all of the affordances provided by the environment. They begin by perceiving only the affordances for themselves, later learning to perceive the affordances for others. How does this process of learning affordances not directly related to their own internal and external perceptions come about? Gibson provides few clues. He has described how they might come to perceive the same surfaces, but not the same tastes afforded by different foods. Does a child project his own internal perception on to another child, and thereby understand that the taste afforded to the other child is the same as his? If so, such a process is in no sense direct. If Gibson's theory

is to hold generally, then we must suppose that the ambient light carries the information that one child has the same affordances as another child, and that children directly perceive this. But the ambient light does not carry information about a child's internal sense, and so cannot specify entirely the affordance an object provides to that child. Learning would thus seem to require a process other than direct perception, which would mean some affordances are not directly perceived.

The more radical version of Gibson's theory also cannot adequately explain how misperceptions occur. In explaining the nature of misperception, Gibson imagined a scenario in which a sheet of plate glass is extended over the edge of a cliff. Although the cliff no longer affords falling, humans misperceive it as dangerous nonetheless. Gibson contended that the ambient light carries information about the cliff itself, but not about the glass that would prevent someone from falling off the cliff. Here Gibson is mistaken. The ambient light does carry information about the glass, if it 'carries' information at all. Such information is not merely haptic as Gibson asserted unless the glass is perfectly translucent (and no glass is). The human perceptual system cannot detect this visual information, however, and this is what causes the misperception. But a more sensitive perceptual system would be able to detect the presence of the sheet of glass. Furthermore, humans can be taught to recognize the presence of what seems imperceptible, such as the glass, by honing in on other visual cues. If a person learns to perceive the presence of the glass, does the ambient light suddenly carry information about its affordance?

Gibson's examples address only the case where a human does not see something that is actually present. A second case of misperception, illusory images, presents greater difficulties for Gibson's theory, a fact Gibson recognized (Gibson 1979, 243-244). If a person has

the illusory perception of a surface, is the misinformation due to the ambient light, or to the person's perceptual capacities. If we accept the radical reading of Gibson's theory, that meanings are external to the perceiver, then either the ambient light can provide false meanings, or the error is in the perceiver. The former explanation is untenable: In what way could light in itself be thought to have a false meaning? The latter explanation contradicts this reading of Gibson's theory, for some mechanism other than mere direct perception must account for the introduction of error. Therefore, the meaning of the perception is no longer external to the perceiver, but rather dependent upon him.

5.5 How Ecological Externalism Succeeds Where Evolutionary Externalism Fails

5.5.1 Defeating Chauvinism

As we have seen in the critiques of Millikan and Dretske, natural selection does not grant special status to its products such that a naturally selected mapping qualifies as a representation and an artificially produced one does not. Evolutionary externalism, therefore, excludes not only such metaphysical possibilities as Swampman from the class of representational systems, but also historical actualities, such as organisms that have yet to have their nascent sensory systems selected for. This strongly indicates that representation is something other than what evolutionary externalists claim, that there is a core to it that cuts across systems regardless of their history.

Even if the evolutionary history of a function enables us to discover what it is doing, this fact serves only as a means of discovery of the core function. It is not a necessary component, but rather an indicator of what we should look for in the organisms behavior.

How does ecological externalism fair with Swampman? Let's provisionally assume that Gibson is correct about the nature of optic array, at least as it concerns its relations to surfaces and objects (and not what it affords the organism in terms of value). The internal mechanisms of Swampman remain the same as Davidson. Place Swampman in the same environment as Davidson, and he will be confronted by the same sorts of optic arrays. Nothing about Swampman changes the structure of the optic arrays nor the environment that produces them, and so the information contained in them remains the same. Swampman can move around and sample optic arrays and extract the same invariants as Davidson, because Swampman has the same physical equipment as his double. Therefore, the specification of what information Swampman has, its external and internal components, is exactly the same.

If we are Representationalists[1] in the mold of Dretske, then we must conclude that Swampman has the same thoughts as Davidson. But accepting Dretske's story about the relation between experience and representation, we must also conclude that Swampman has the same experiences as Davidson. In each case, we remain externalists because the information that determines representation does not supervene on brain states.

This argument holds for our mutated-but-not-yet-selected-for organisms. They have internal characteristics that crudely process information, and their environment provides structured information that these internal functions can extract invariants from. So it seems that ecological externalism avoids the charge of chauvinism as it has been formulated here. What about liberalism?

1. It should be noted that Gibson rejects the notion of a representation as being something re-presented (Gibson 1979). There is no assumption made here as to whether Gibson is right or not. For the purposes of this section, 'representation' is shorthand for complex informational structure used by an organism.

5.5.2 Dodging Liberalism

If we were to require that for something to count as a representation for an entity there merely needs to be some informational structure transduced by the entity, we would again be faced by the prospect of panpsychism. Rocks are continuously bombarded by ambient light, and some of this light gets transformed into heat energy, yet rocks do not have representations. It would seem quite easy to sidestep this prospect.

Presumably, all we would need to do is require that the external information get "used" in some manner by the internal processes of the entity. This vague notion allows for a wide range of interpretation. We could wed ecological and evolutionary externalism, and require that the external information be used for the selected purposes of the organism, but this would simply lead us back to the chauvinistic position repudiated above. We could specify that the external information must be transformed into differently structured information, i.e., manipulated. Whether this avoids liberalism depends on what one means by 'information' in this context. If it denotes Shannon information, then one might again argue that rocks are representational systems.

One possible manner of specifying what it means for information to be used is that it enables behaviors that help to sustain the organism. This does not require an evolutionary context: a robot could produce behaviors that help keep it going. This would seem to exclude mental events like daydreams from being informational. Only if we insist, however, on using exclusively external terms to specify representations. A class of internal states can be specified according to similarity of intrinsic factors. Those that "use" external information are a proper subset of this class. Those that don't are another proper subset. We needn't religiously adhere to the notion that all representations are external. Or we might

apply Millikan's notion of a derived proper function to daydreams and the like, but instead of the original function being evolutionary in nature, it is systemic.

Thus, whether one avoids liberalism depends on one's notion of what information is. Deciding on a particular definition of information has implications for what sort of mechanisms can be present. My criticisms of computationalism's inability to solve the symbol grounding and frame problems implies that computationalism really does not fit with externalism. The next chapter seeks to describe the mechanisms through which an ecologically externalist form of representation could come about. This is not to imply that Gibson's arguments establish that ecological externalism is true. There are still many holes to be filled.

5.5.3 What Hath Gibson Wrought? Information and Misrepresentation

Ultimately, Gibson's theory is an empirical one. It is generally used by externalists to demonstrate how an externalist theory *might* play out in actuality. While its failure does not imply that ecological externalism is false, it is worthwhile to ask how far Gibson has gone in establishing his theory.

Though not conclusive by any stretch of the imagination, it ought to be pointed out that the majority of vision researchers reject Gibson's ecological theory of perception. Ullman's critique is generally seen as decisive; research in computer vision proceeds on the standard model, as does most research in the psychology of perception. Few if any attempt to model stereo vision in the way described by Rowlands, as sequential sampling by individual eyes, rather than as a synthetic process. A Gibsonian might argue that this is why computer vision is so hard—everyone is doing it wrong. Well then, let the Gibsonians show how it is done.

For some reason, not many are taking up this challenge. Again, this is not meant to be conclusive about the prospects of an ecological theory. In fact, as will be laid out in the next chapter, certain aspects of Gibson's theory have been taken up by dynamic systems theorists, although it does not yet appear that they will vindicate Gibson's specific claims about visual perception and the optic array, if they try to at all.

What is lacking from Gibson's theory that could yield more than just a theory of perception is a more developed account of what information is. What Gibson offers is fraught with difficulties. The primary of these has already been discussed but bears repeating: the relation between information and misrepresentation.

Information for Gibson consists of a relation between structure of the optic array (or external informational structure in general) and environmental structure. It is a unique mapping, though not necessarily one-to-one for all of the elements of structure in array and environment. The question then arises as to what misrepresentation is.

There are three possibilities for what could cause misrepresentation. First, there could be two identical optic arrays, each corresponding to different environmental features. This is ruled out by the specification of uniqueness. Second, there could be a malfunction in the internal processing that is independent of the external information. Third, there could be a problem of relating internal characteristics to external information. The more direct one reads Gibson's conception of perception to be, the less plausible these last two possibilities become. The possibility of misrepresentation vanishes.

The uniqueness of mapping is what must give, because it is false. It is theoretically possible to reconstruct ambient light without the usual environmental features that produce it.

It is even practically possible for controlled, simple conditions. The structure of ambient light does not uniquely point to anything.

It would seem that we have to invoke Normal conditions to save ecological externalism. But that would only bring along the baggage of evolutionary externalism. Dynamic systems theory offers another way. But one last glimpse at philosophical externalism before we turn to its possible rescuer.

5.6 Tye's Externalism: Why Tye's Representationalism Has No Foundation

In *The 10 Problems of Consciousness* (1995), Michael Tye expounded a Representational account of experience similar to that of Dretske's. Like Dretske's, it was founded on an externalist theory of representation, and attempted to show how the representational character of experiences helps explain away the more difficult problems of naturalistic theories of consciousness. Tye went further than Dretske, however, in claiming that even such states as anxiety and depression are representational. Whether Tye's Representationalism can solve the problems of consciousness depends in great part on whether he can formulate a plausible theory of representation. It is this aspect of Tye's theory that is examined here; whether Representationalism, once granted, can solve the problems of consciousness is discussed in Chapter 7.

Tye explicitly endorses the causal covariance model of representation:

> There are many different theories about the nature of representation, but one approach that seems well suited to sensory representations (although not to beliefs) is the causal covariation view. On this view, if optimal or ideal perceptual conditions obtain, sensory states of the sort found in perception track the presence of certain external; they thereby represent those features. (Tye 1995, 105)

There are some important qualifications in this statement. First, the causal covariance model is only intended for sensory representations. Second, there is the notion of optimal conditions obtaining. Third, there is the notion of "tracking" that needs to be explained. Tracking is just causal covariance of a state S in an object x with a state of world/object P; to put it more formally: "S represents that P = df if optimal conditions obtain, S is tokened in x if and only if P and because P" (Ibid, 101). So tracking only occurs under optimal conditions, conditions in which there are no abnormalities, malfunctions, or intervening circumstances. A misrepresentation occurs when these conditions do not obtain.

At first glance, the definition of optimal conditions seems circular. Optimal conditions are conditions under which a representation could occur, but a what a representation is depends on what optimal conditions are. So a representation happens under conditions under which a representation could occur. But optimal conditions are to be understood as those conditions under which "S would be tokened in x if and only if P were the case; moreover, in these circumstances, S would be tokened in x because P is the case" (Ibid, 223).

Optimal conditions are not necessarily equivalent to Normal conditions. Normal conditions are those conditions for which an attribute was selected for, not, as Millikan tries to argue, those conditions under which an attribute always fulfilled its function and hence was selected for. Optimal conditions need not refer to evolutionary history. They could simply be a subset of the conditions under which a state is perfectly correlated with what it represents in the world, a correlation due to causation. It is natural information without the evolutionary component.

Optimal conditions, if not the conditions for which an attribute was selected for, must be defined on a case by case basis. There is a great deal of variation between organisms of

the same species. The conditions under which my visual perception fails differ from the conditions under which yours fails. This variation is not covered by the specification of no abnormalities or malfunctions. There is no canonical human visual system against which to determine an abnormality. Bodily function is generally within a range, and abnormality is also generally a gross deviation. The only alternative to specifying optimal conditions in terms of selected function is to specify it in terms of average ability; unfortunately, average conditions do not get you the if-and-only-if relation between a tokening and the state of the world.

Tye seems in places to have tentatively thrown his lot in with the evolutionary reading of optimal conditions:

> It is still *possible* that two different organisms that evolved in different ways, while nonetheless sharing the same internal microphysical states (at some given time t), differ in their phenomenal state at *t*. (Tye 1995, 153)

But Tye does not consider sameness of evolutionary history to be a *necessary* condition for sameness of representation. Thus, he rejects the notion that Swampman does not experience anything. He does, however, accept the notion that microphysical duplicates could have duplicate sensory states causally correlated with different environmental features by virtue of different evolutionary histories and natural habitats, and thus they would have different experiences. What does this all imply?

Optimal conditions need not be specified in terms of evolutionary history (e.g., Swampman), but can be (e.g., for you and I, or duplicates of us). Tye has us imagine a duplicate of himself produced by the Transporter device on the series *Star Trek*. This duplicate is then sent to an unexplored planet. The experiences he has represent their earthly correlates—what they would represent on earth—because of the duplicate's causal connection to Tye.

What would Swampman's experiences represent if placed on this planet? The relevant causal relationship no longer exists, because the similarity between Swampman and Tye is accidental and independently caused. Swampman represents the external states his mental states track. Presumably then, Swampman's mental states on the unexplored planet would represent aspects of that planet, because there are no evolutionarily selected for earthly correlates.

So, if you have an evolutionary history, it matters to what you are experiencing; if you don't, don't sweat it, because you don't need it. Can this really work? Suppose we take Tye and his Swampman, crack open their skulls, and swap some neurons. We will specifically swap neurons operating within neuronal groups. If we find four neurons working together to produce edge representations, we will take two of them and swap them with the corresponding pair in the double's brain. We will do this for all neuronal groups. Now, when Tye and Swampman are placed on the unexplored planet, what are the causal correlates of their representations? A mix of earth objects and unexplored planet objects? Half of the neurons in neuronal groups in each of their heads were selected for their contribution in producing a particular type of representation, the other half were not. The only reasonable conclusion is that their causal correlates are the aspects of the environment that they currently track. This is the common core between Swampman and his double.

Tye cannot have it both ways. Either he must fully embrace evolutionary externalism, or reject it. Trying to split the difference only leads to absurdities. Since evolutionary externalism has been shown here to be untenable, the correct course is to drop that aspect. Having done so, an externalist like Tye must elucidate a theory of causal covariance that allows for the stochastic relationships between human representation and the structure of

the environment, rather than insisting on perfect covariance conditions. To do so means adopting an appropriate model of underlying mechanisms and how they relate to the environment.

5.7 Whither Externalism?

Philosophical externalism was initially motivated by the intuition that the internal structure of a representation does not determine its reference, and therefore content must have an external component. Evolutionary externalists tried to establish a nomic relationship between the evolutionary history of a representing device—emphasizing why it has the structure and, hence, function that it has, rather than what its structure is—and the reference of the representation. They have not only failed to establish this relationship, misconstruing aspects of Darwinism, but the consequences of their theories are to either implausibly deny the ability of plausibly cognizant beings to represent their environment or to count nearly everything as cognitive.

A plausible externalism begins with the understanding of the environment's informational contribution to organisms' representational capacities, as well as the need for internal mechanisms that can somehow "sync" themselves with the environment. For these internal mechanisms to achieve this, there must be both organizational capacities and organizational principles. These are not determined by the evolutionary history of the organism, though this can be contributing factor as to why they exist. Instead, they are properties of the organism as a system, not as a historical artifact. One of these organizational capacities is the integration of information through active sampling, as Gibson pointed out in his ecological theory of visual perception. Another, as I will discuss, is the capacity for self-orga-

nization of internal patterns around sensory signals. What these considerations highlight is the need for an understanding of the mechanisms involved in producing representations in order to understand what representations are.

The dynamic systems conception of mind offers the externalist these very mechanisms. But it is not only the externalist who can benefit from dynamic systems concepts. They offer the Representationalist a different understanding of the causal covariance relationship upon which their theories rely. Although Tye has endorsed the view that what is happening internally to the organism is computation, the mechanisms of dynamic systems provide a better foundation for his theory of phenomenal content. These issues will be raised in the final chapter when I sketch a theory of mind based on dynamic systems principles.

Chapter 6 Towards a Solution: Dynamic Systems and Information

We have seen a number of efforts toward understanding the nature of semantic information. Dretske identifies semantic relations with causal connections between representations and what is represented. Sayre equates semanticity with veridicality as defined by the mutual information between signal channels. The molecular Darwinists understand semanticity to be the relation between physical structures. Millikan considers mechanisms contentful by virtue of their having been designed for that function. Gibson and other proponents of direct realism have argued that semanticity resides in the environment and the affordances it provides. I turn now to detailing an alternative account of semantic information, one inspired by dynamic systems theory.

The dynamic systems understanding of semanticity is that the extremes of direct realism and methodological solipsism both make the mistake of considering the individual and his environment as separate entities (Kelso 1997). They must choose one of these antipodes as the place wherein semanticity resides: either the person or the world. Approaches such as Dretske's and the 'Robot Reply' to the Chinese Room argument also implicitly accept this divide when they attempt to bridge it via causal connections between a person's representations and the world around him.

Instead of separating the individual from the environment, these two poles should be understood to be part of one dynamic system. Semanticity, the relation between represen-

tation and represented, is then embedded in the parameters of the equations describing this system, or as Kelso writes in *Dynamic Patterns* (and as quoted earlier):

> Like mind and matter, the concepts of information and dynamics have long been held distinct and separate. Usually, they are taken to refer to fundamentally different but alternative modes of describing complex systems . . . But look at what is done here. Instead of treating dynamics as ordinary physics using standard biophysical quantities such as mass, length, momentum, and energy, our coordination or pattern dynamics is informational from the very start. The order parameter, φ, captures the coherent relations among different kinds of things . . .
>
> Notice, coordination dynamics is not trapped (like ordinary physics) by its (purely formal) syntax. Order parameters are semantic, relational quantities that are intrinsically meaningful to system functioning. What could be more meaningful to an organism than information that specifies the coordinative relations among its parts or between itself and the environment? This view turns the mind-matter, information-dynamics interaction on its head. Instead of treating dynamics as ordinary physics and information as a symbolic code acting in the way that a program relates to a computer, dynamics is cast in terms that are semantically meaningful. The upshot of this step, which, I stress is empirically motivated, is that intentions do not lie outside self-organized coordination dynamics. (Kelso 1997, 143-144)

At first glance, it would seem that what Kelso is suggesting in this passage is primarily a change in perspective, in how we view the mechanisms that give rise to intentions. What I will argue is that this different perspective describes different mechanisms than those supposed in computational theories of mind. Kelso also suggests that dynamic systems in general are inherently semantic. This would be an error, similar to that of Chalmers when he supposes all entities describable in terms of Shannon information have experiences. I will describe the properties that dynamic systems must demonstrate to bear semanticity, circumscribing this feature to a small set of dynamic systems. Finally, I will propose mechanisms underlying these properties and present a theory of mind that accords with these mechanisms.

6.1 The Nature of Dynamic Systems

A brief overview of one example of a dynamic system, the Watt governor, was given in chapter 2. Van Gelder used this example to flush out the differences between computational and dynamic systems. The impression this leaves is that computational systems cannot be dynamic systems. Van Gelder encouraged this impression by emphasizing the importance of continuous time operations in dynamic systems over the discrete nature of computational systems. In so doing, van Gelder is carving out his own notion of what a dynamic system is, as current dynamic systems theory includes such discrete systems as iterated maps under its rubric (Strogatz 1994). Rather than champion dynamic systems in general over computational systems, I will point out the properties of specific types of dynamic systems that recommend them over computational models of cognition. An overview of the relevant properties of dynamic systems is presented here.

Van Gelder is right to emphasize continuous time dynamic systems, because brains operate in continuous time. This is not the only, nor most important, aspect of time when considering its relation to cognition, although this will emerge later in this chapter. Researchers use differential equations to describe the evolution of continuous time dynamic systems, and most of these equations are nonlinear in nature. Nonlinear systems are not equivalent to the sum of their parts. This means that most equations describing dynamic systems cannot be solved analytically. In fact, nonlinear differential equations cannot be broken down into parts as their linear cousins can, making analysis difficult. Among nonlinear dynamic systems found in nature are the class of biological oscillators, such as neurons. Nonlinearity also holds importance beyond the fact that it is descriptive of the behavior of neurons; it is not just low-level physiological phenomena that are nonlinear,

but higher brain functions as well. Computational systems are capable of computing nonlinear functions—these are the basis of artificial neural networks—but nonlinearity in function is largely missing from classical AI.

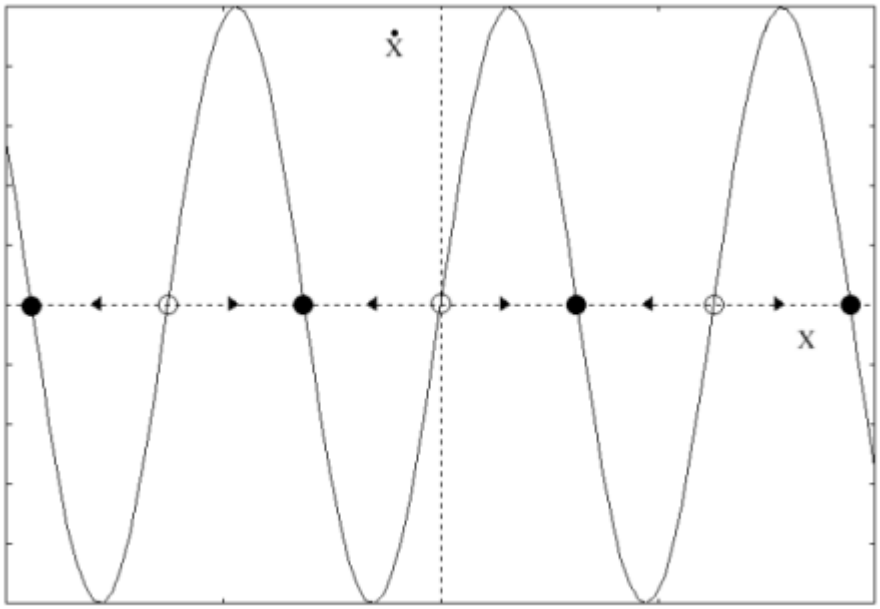
Differential equations trace out a trajectory or path for the system in what is called phase space. Phase space is the space of all possible states for the system over time. Its trajectory is the path a system takes through this space. The flow is the direction of the trajectory for a subset of time, or just a subset of the trajectory. Points at which there is no flow are called fixed points, and these may be either stable or unstable. Stable fixed points, or attractors, have the flow toward them, and unstable fixed points, or repellers, have the flow away from them. More precisely, an attractor is defined as follows:

> An attractor is a closed set A over states of the system with the following properties:
> 1.) A is an invariant set: any trajectory x(t) that starts in A stays in A for all time.
> 2.) A attracts an open set of initial conditions: there is an open set U containing A such that if $x(0) \in U$, then the distance from x(t) to A tends to zero as $t \to \infty$. This means that A attracts all trajectories that start sufficiently close to it. The largest such U is called the basin of attraction of A.
> 3.) A is minimal: there is no proper subset that satisfies conditions 1 and 2. (Strogatz 1994, 324)

A *strange attractor* is an attractor that "exhibits sensitive dependence on initial conditions," by which is meant that nearby trajectories will separate exponentially fast (Strogatz 1994, 325). Strange attractors are also known as *chaotic attractors*.

One way to visualize the dynamics of a system is to interpret the differential equation governing it as a vector field. Using Strogatz's (1994, 16-17) example, the nonlinear differential equation $\dot{x} = \sin x$, we can think of x as the position of a particle moving along the real line and $\dot{x}$ as its velocity. The vector field can be shown by plotting $\dot{x}$ versus x and placing arrows to indicate the velocity vector at x. Figure 6.1 shows the corresponding plot.

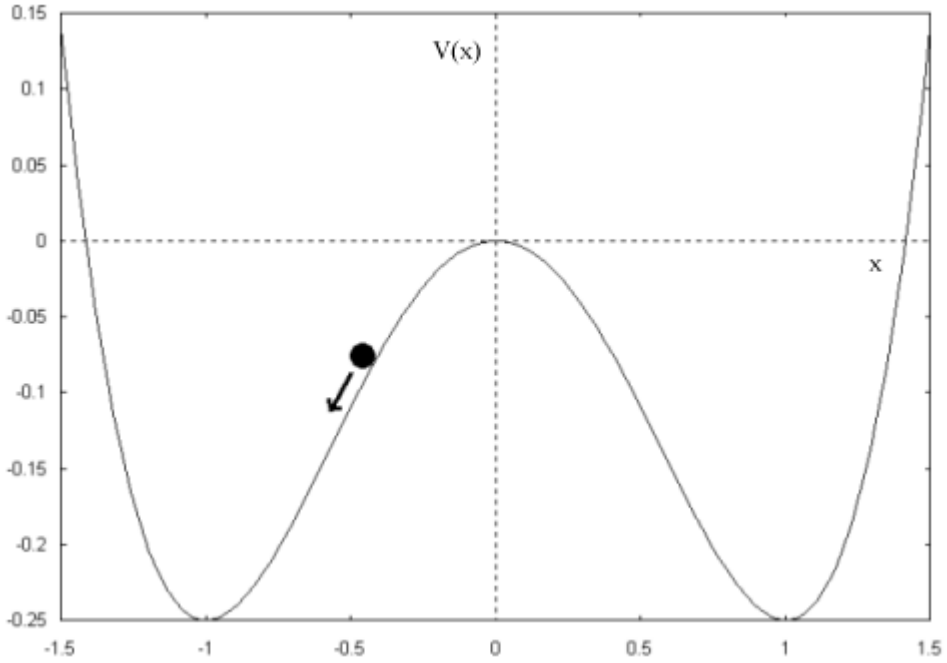
Figure 6-1. Plot of $\dot{x} = \sin x$



The arrows indicate the flow, which is to the right when $\dot{x} > 0$ and to the left when $\dot{x} < 0$. When $\dot{x} = 0$, there is no flow, and these are the fixed points. The solid circles represent the stable fixed points or attractors, which have the flow directed toward them, and the open circle represent the unstable fixed points or repellers, which have the flow directed away from them. Attractors need not be fixed points or simple intervals; the limit cycle attractor is an example of an attractor that is not a fixed point, and limit cycle attractors characterize the activity of pendulums. In fact, dynamic systems can produce arbitrarily complex attractors, with strange attractors being the most extreme example.

Another way to visualize the dynamics of a system (Strogatz 1994, 30-32) is to use the notion of potential energy, where the potential V(x) is defined by $f(x) = -\frac{dV}{dx}$. The idea is to imagine a particle sliding down the walls of a potential well. An example for the system $\dot{x} = x - x^3$ is given in Figure 6.2.

Figure 6-2. Plot of potential V as computed from $-\frac{dV}{dx} = x - x^3$



The solid circle is the imagined particle as it slides down the side of the potential well. Two attractors are shown at $x = \pm 1$ and one rippler at $x = 0$. This gives an example of a 'multistable' system, which is an important property of biological dynamics (Kelso 1997). This method of visualization is especially useful for illustrating the relative depth of attractors, which indicates how much energy is required to push them out of their stable state. Stability is a graded concept, with varying degrees of stability (instability) characterizing different attractors (repellers).

6.2 One Step Further: Self-Organizing Dynamic Systems

As one changes the parameters of dynamic systems, attractors and repellers can be eliminated or changed into one another. These changes are called *bifurcations,* and the points in parameter space at which the changes occur are called bifurcation points. Dynamic systems

that do not change behavior for most values in parameter space are structurally stable. Such systems show ordered behavior, which is to say that for large, connected regions in parameter space, these systems have fixed behaviors. The opposite extreme is chaotic behavior, characterized by the existence of chaotic attractors, i.e., the system can change behavior dramatically in only small steps in parameter space. In between the two extremes is complex behavior, in which fixed behaviors are breaking up and chaotic behaviors become fixed (Kauffman 1993).

Self-organizing dynamic systems are systems in which patterns spontaneously arise out of the interactions of their subcomponents. The interactions are nonlinear, which means the patterns are not merely the sum of their components and that the system is dissipative and far from equilibrium. A *dissipative* system is one in which the energy in the system is not uniformly distributed, instead being concentrated in flows that more easily dissipate the energy. This results in a system in which most of its degrees of freedom are suppressed; those that are not suppressed are known as *order parameters*. The order parameters are found at phase transitions where new patterns arise. Fluctuations in the system allow it to discover new patterns by breaking out of stable attractors, and are not just useless noise. *Control parameters* are those parameters that lead the system through changes in patterns, and are themselves not dependent on these patterns. Changing the direction of a control parameter, such as temperature for certain chemical reactions, can result in *hysteresis*, which is a condition in which an overlapping region of parameter space exists where the system can be in one or more states depending on the direction of change in the control parameter.

To illustrate these ideas, it is useful to look at a classic example of a self-organizing system: the laser (Haken 1996). A gas laser consists of a glass tube holding a gas of laser-active atoms. These atoms are excited by means of an electrical charge and emit light waves with random phases. The light waves are reflected by mirrors, trapping the light in the laser for a period of time. As the atoms are increasingly excited and the concentration of light waves grows, a process called *stimulated emission* occurs. Stimulated emission is when a light wave hits an excited atom, forcing the atom to release energy, which the light wave then picks up. The enhanced light wave may go on to hit another atom, and so on, causing a cascade of enhancement. When the light waves oscillate, they force the electrons of the atoms to oscillate in phase. But since there are many light waves in the laser with different phases, they are competing against one another. Eventually, one of the light waves 'wins' and all of the light waves become 'enslaved' to its amplitude, which is the order parameter of the system. Rather than a mix of randomly phased light waves, the laser now produces coherent light. This process will not happen with lower levels of excitation. The increasing electrical charge, which is the control parameter here, drives the system to instability. Once the system reaches an unstable state, it starts to self-organize. The relation between the order parameter and the component light waves is that of *circular causality*:

> The order parameters act as puppeteers that make the puppets dance. There is, however, an important difference between this naive picture of puppeteers and what is happening in reality. As it turns out, by their collective action the individual parts, or puppets, themselves act on the order parameters, i.e., on the puppeteers. While on the one hand the puppeteers (order parameters) determine the motion of the individual parts, the individual parts in turn determine the action of the order parameters. This phenomenon is called *circular causality*. (Haken 1996, 43)

The notion of circular causality is a bit troublesome, since order parameters are abstract entities. But the behavior characterized by order parameters emerges from the interactions

of the individual components and this emergent global behavior keeps these components in line. The *slaving principle*, which is the fact that order parameters govern the behavior of the system, enables us to describe such complex systems in simple terms. But to know how order parameters arise, we must understand the interactions of the individual components.

In the case of the laser, the control parameter, the electrical charge, was external to the gas interactions and, therefore, not dependent on them as the order parameter is. This need not be the case, and, in fact, is not in dynamic systems such as the mammalian brain. What leads the brain through changes is often originated within the brain itself and dependent on the dynamics of the brain.

6.3 Conscious vs. Nonconscious Self-organizing Dynamic Systems

Not all self-organizing dynamic systems are conscious, and so we need to be able to distinguish those aspects of self-organizing systems that are essential to conscious systems. Skarda and Freeman argue that 'global objectives' possessed by biological organisms with neural systems differentiate them from nonbiological systems and nonneural biological forms. The global objectives are motivations, characterized as

> complex process[es] whereby the organism predictively controls and maintains itself in the optimal condition given the circumstances in which it exists and acts. (Skarda and Freeman 1987, 173)

These global objectives regulate neural dynamics, according to Skarda and Freeman, in that they "limit the possible range of patterned neural behaviors and they mediate interaction among various neural subsystems" (1987, 173). So the distinction to be drawn between biological entities with brains and those without is that brains control bodies "for the self-pro-

moting purposes of search, attack, ingestion, escape, and reproduction" (1987, 173). Plants are an obvious example of biological forms that do not bear these characteristics.

Having described conscious self-organizing systems at the level of Millikan's intentional icons, Skarda and Freeman draw no further distinctions between conscious and nonconscious biological entities, leaving their account open to the same objections that plague Millikan's. The primary difference between Millikan's intentional icons and the Skarda-Freeman theory of consciousness, however, lies in the nature of self-organization. Millikan requires that there be a mechanism that has the function of producing an intentional icon in order for something to count as an intentional icon. Although intentional icons can be acquired (e.g., retinal images), they must be produced by organs selected for that very purpose (e.g., the retina). A mechanism that arose through self-organization would not have a normal function, according to Millikan's theory. This additional degree of freedom in Skarda and Freeman's theory is important—without it we would have to consider most neural mechanisms to be contentless—but not sufficient to explain why brains have content-laden, intentional states.

Nonetheless, the ability to self-organize is, I contend, a necessary condition for the production of such states. Self-organizing systems possess two essential properties of true representation-producing mechanisms: plasticity and nonlinearity. By plasticity, I mean the capacity to smoothly adjust representations, as well as the ability to spontaneously create new representations. By nonlinearity, I mean the ability to create representations that have arbitrary domains, including representations of ad hoc categories. An organism or machine that does not have these two capabilities will not be able to successfully navigate and manipulate a dynamic environment. It is because they lack these characteristics that current

robotic systems are unable to produce behaviors that humans find extremely simple to achieve. In short, it is because they lack these characteristics that robots suffer from the frame problem.

Skarda and Freeman have provided one piece of the puzzle, which is, to paraphrase their formulation, the ability to self-generate control parameters that guide a system toward maintaining itself. For organization to be maintained in a system, it must thwart the Second Law of Thermodynamics—all closed systems increase in entropy—by introducing energy into itself. Actively maintaining oneself requires being able to control one's actions within changing environments. This in turn implies an ability to dynamically adapt to environments. What enables a system to do this?

Watt governors and humans both have the ability to dynamically adjust to conditions in their environment, although the environment of a Watt governor consists merely of the valve to which it is attached. Nonetheless, a Watt governor does not support cognition. Although it is lacking the requirement outlined by Skarda and Freeman, this is not the principle reason to judge it as noncognitive. Missing from the Watt governor are any mechanisms that could be considered representational. The Watt governor does its job without internal representations because it does not need to plan for the future, or make any form of abstraction about its environment. The Watt governor in itself is a poor metaphor for the human brain. Instead, the Watt governor represents how perceptual subsystems might function. Although the Watt governor does not bear any internal representations, it is itself representational in that its states are statistically correlated with environmental conditions. A perceptual subsystem, such as edge detection through vision, can be conceived as like a Watt governor in that it does not have internal representations, but its behavior is represen-

tational in the same loose way as the Watt governor. There is a statistically correlated pattern of activity in this perceptual subsystem centered around the presence of edges. I say that the notion of representation here is a 'loose' one because this is not implying that the activity of the subsystem acts as a symbol for the larger perceptual system or the brain as a whole. Rather, the perceptual subsystem for the brain is like the valve and steam for the Watt governor: a component that influences its dynamics. This component has its own dynamics centered around features of the environment, but its dynamics are influenced by the higher level systems as it in turn influences them. This is not the 'circular causality' of Haken, but the feedback or reentry between neural systems as proposed by Damasio (1989) and others (Edelman 1987; Pribram 1991).

As we saw earlier, Damasio argued that representations arise through the time-locked multiregional retroactivation of perceptual sites. Pattern recognition and memory are activities of pattern *re*formation. To recast this in a dynamical systems context, the brain is a dynamic system composed of semi-modular dynamic subsystems corresponding to perceptual sites. The subsystems organize around input from the environment, and through connections to other subsystems, affect the dynamics of the brain as a whole. These in turn feedback to the originating subsystems. The subsystems are semi-modular because they have their own attractor states and do not necessarily affect the entire brain's dynamics. They are not truly modular in Fodor's (1983) sense of being informationally encapsulated, that is, sealed off from information from other systems. To answer Rapaport's (1996) question as to what is stored if not symbolic representations, memory consists of the adaptation of brain chemistry, through processes such as long term potentiation, so that the brain can reproduce, through feedback to various subsystems, patterns of neuronal behavior corre-

sponding to the patterns that emerged during experience. This is not *re*presentation literally. The reformed patterns may be adjusted according to a variety of factors, such as the work of processes that strip away aspects or bind with other patterns.

To summarize the dynamic systems interpretation of the phenomena described by Damasio: Patterns emerge from the dynamics of perceptual subsystems; these in turn affect the dynamics of the convergence zones, and continue to feed upward; the patterns are recreated in the perceptual subsystems through feedback as directed by global control parameters. The patterns here are not symbols, but stable patterns of activity, attractors, in the neuronal ensembles.

Dynamic systems theory has already been applied to explain the emergence of time-locked neuronal oscillations (Schuster 1991). Schuster analyzed the behavior of neuronal ensembles coupled via excitatory and inhibitory connections, and pointed to the emergence of a *Hopf bifurcation* leading to the time-locking. In a Hopf bifurcation, a limit cycle attractor (the kind of attractor that characterizes a pendulum) emerges from a stable fixed point attractor that loses stability. The Hopf bifurcation is an important phenomenon, because, as Kelso points out:

> The Hopf mechanism offers an intriguing way to spontaneously create and dissolve dynamic patterns of behavior. It is potentially important because this . . . process appears to be one way the nervous system achieves stability while flexibly adapting to environmental requirements. (Kelso 1997, 87)

Processes of neural self-organization enable organisms to adapt in real, not evolutionary, time scales to environmental conditions through reproduction of experience as guided by internal control parameters in the form of innate and self-generated goals. A system that lacks these features cannot be said to be aware of its environment or itself, and, therefore,

is not a conscious entity. This does not mean that intelligence cannot be created through artificial means. Rather, such intelligence would have to be instantiated through self-organizing pattern reformation functions (and not necessarily structure) found in real brains.

6.3.1 Self-organization, Order Parameters, and the Theory of Affordances

Both Skarda and Freeman and Millikan point to the need for including the behavioral function of representations into an account of what representations are. Skarda and Freeman explicitly link their effort along these lines to Gibson's theory of affordances:

> Nervous system dynamics is a self-organizing process constrained by the requirement that the system anticipate and incorporate the immediate consequences of its own output within the larger constraints of regulating its well-being and the long-term optimization of its chance for survival. This is subsumed in J. J. Gibson's theory of "affordances." (Skarda and Freeman 1987, 173)

Kelso also links his semantic understanding of order parameters to Gibson's theory of affordances. In each case, it is the less radical version of Gibson's theory that is embraced, the failings of which I detailed above. So how does the dynamic systems theory of semantic information escape these failings when it is so closely linked to their source?

Kelso sees order parameters as the essential component in a theory of semantic information. Order parameters capture the relations between the sensing and sensed systems, between organisms and their environment. Kelso argues further that order parameters are not merely descriptive tools, but rather have a metaphysical reality no different than the systems whose relations they capture. Furthermore, order parameters are not informational merely in the sense that they describe the relations of these systems for the dynamic systems analyst, but also, and more importantly, in the sense that they are what intentions influence in order to change the behavior of the system having the intentions. Because of

this, Kelso contends that order parameters are intentional in nature. Taken at face value, this is to confuse ‘intending to do something’ with the intentionality of beliefs and desire; it is the behaviorist maneuver of replacing thoughts about objects with dispositions to behave toward them in particular ways. But if we understand Kelso's description of intentions to include acts in which the brain directs its attention to objects, and thereby changes its order parameters, we have the beginnings of an account of what intentionality and semanticity are.

This additional component, the action of the mind/brain to alter itself, distinguishes the dynamic systems approach from Gibson’s direct realism. Whereas Gibson considers semantics to be provided from the environment, dynamic systems theorists view semantics as a function of both the environment and the mind/brain. The latter view may seem self-evident: semantics has to do with the relation between a concept and what it is about. However, dynamic systems theorists, unlike computationalists, do not view the environment and the mind/brain separately. All semantic information is context-dependent, because the state the mind/brain is in when attending to the world is dependent on the state of the environment. The mind/brain self-organizes around environmental signals; it does not create abstract symbols and then assign them to environmental features.

This is true even in the case where the mind/brain is focused on its own ideas. In this case, environmental signals include the mind/brain's own functioning, i.e., feedback from its own mechanisms. Remembering and reflecting are not merely dispositions to act, but reenactments of processes that accompanied perception and action with attention drawn to these reenactments. Here, attention ought not to be confused with conscious awareness.

The mind/brain attends to an event or object when it allocates resources to process signals corresponding to that event or object. Awareness of this process need not accompany it.

Although humans are capable of drawing their attention to arbitrarily defined categories, the process of evolution has led to mechanisms for resource allocation based upon the needs of the human organism. This introduces the concept of value into the dynamic systems account of the mind/brain, a concept that, as we will see, Gerald Edelman makes much of in his selectionist theory of brain function. Order parameters contributed by the mind/brain are not arbitrary. They have been constrained by evolution for the benefit and reproduction of organisms. Value is this inherent predilection of mind/brain mechanisms to guide the organism to beneficial outcomes. The world is seen by the organism through a prism of what is good or useful to it. The concept of value makes another connection to the less radical version of Gibson's theory: objects are perceived in terms of their affordance to the perceiver. The question is how 'direct' this perception is.

The perception of affordances is direct when it is guided by inherent value mechanisms in the sense that the order parameters necessary for the system attaining the 'correct' state need not be contributed by the efforts of the mind/brain, because they are already given parameters of the mind/brain. [Value is the predisposition of an organism to perceive an object as useful to it. Without value, there cannot be affordances in Gibson's sense, because semanticity is not a property of a signal, and, therefore, the optic array alone cannot provide meaningful information. Perception is only as direct as the dynamic neural processes that result in an organism's taking a particular attitude (e.g., perceiving something as 'climbable') toward what is perceived. That semanticity depends on an organism's attitude toward

the world indicates that semantic analysis can only be applied to states of an organism, not bits in a computer or action potentials in a neuron.]

The discussion so far has been an abstract analysis of how dynamic systems theory applies to questions about the origin and nature of semantic information. I turn now to actual applications of dynamic systems theory to understanding human semantic capabilities. In particular, I will describe how dynamic systems theory casts human categorization abilities in a new light.

6.4 How Dynamic Systems Theory Helps Explain the Nature of Categorization

In chapter 3, I analyzed Phillipe Schyns's application of SOFM networks to model prototype effects in categorization. An important assumption of Schyns's work is that prototype effects reveal the nature of the conceptual structures underlying human categorization. This is a hotly disputed assumption (Armstrong, Gleitman, and Gleitman 1983; Barsalou 1987; Lakoff 1987), especially given Rosch's (1978) insistence that the prototype theory of categorization does not imply anything about the psychological mechanisms that give rise to prototype effects. The debate centers around the differences between categories, category effects, and concepts.

6.4.1 Most Categories Possess Graded Structure

The classical notion of categories is that they are collections or sets of individuals grouped together according to well-defined membership functions. These membership functions provide a list of the features that an individual must possess in order to be a member of the category. Wittgenstein's (1953) famous example of the category *game* called into question

whether the necessary and sufficient conditions could be enumerated for all categories. Rather than a set of necessary and sufficient conditions for membership, some categories seem to be defined by family resemblances, none of which are in themselves necessary for membership.

The pioneering work of Rosch and Mervis (1975) on prototype effects gave additional credence to the family resemblance theory of categories. Categories appeared to possess *graded structure*, which is to say that there are degrees of membership for categories. Subjects were asked to rate individual objects on how well they represented the category to which they belonged, and what emerged was that people consistently apply similar graded measures of how well an individual represents a class. Rosch's explanation was that categories consist of clusters around a prototype. Later, Rosch and others explained the appearance of graded structure as just that, appearance. Graded structure was really a category effect, the result of applying a concept of a category in a particular situation. This reflects the competence-performance distinction employed by some linguists (Chomsky 1965) to distinguish between an individual's abstract knowledge of language and their ability to apply it. Concepts, as opposed to categories, are the actual psychological structures employed by humans represent real world and abstract categories. The question is whether graded structure effects reveal anything about the nature of concepts.

Graded-structure effects have been found for a wide variety of categories, including categories that are generally held to be classical. Armstrong, Gleitman, and Gleitman (1983) discovered graded structure effects in categories such as *odd numbers*, with the accompanying difference in recognition and learning times between prototypical and non-prototypical members. Armstrong, Gleitman, and Gleitman conclude from this that the

notion that graded-structure effects reflect human conceptual schemes is flawed since it is inconceivable to them how a graded structure category of odd numbers could be applied to math. Rather surprisingly, Pinker and Prince (1999) have recently discovered graded-structure effects in the linguistic category of *irregular verbs*. When they applied their methodology to regular verbs, however, they discovered no graded-structure effects. They concluded that the category *regular verbs* is classical in nature. Barsalou (1987) has also found graded-structure effects in ad hoc categories, such as *things to eat on a diet*.

The early work on graded-structure effects suggested that the results were quite robust, with people agreeing with one another as well as with themselves over time about the degree of membership for various individual items. Barsalou (1987), however, demonstrated that this is not the case. In fact, degree-of-membership judgements vary over time for individual persons as well as between people. Barsalou found three different types of typicality used for making judgements—closeness to central tendency, ideals associated with a category, and how frequently an individual is perceived as instantiating its category—and that different measures of typicality were used in different contexts. He also noted that linguistic context can alter the graded structure, as can change of perspective. People perform poorly at conceiving the typicality judgements of others. Even for categories that are supposedly classical in nature, where people employ definitions to determine membership, it has been found (Bellazza 1984) that the propositions involved in the definitions that people use change over time, with changes occurring in time scales of a week or less.

One approach to dealing with graded structures and their instabilities is to dismiss them as mere noise or epiphenomena (Armstrong, Gleitman, and Gleitman 1983). As Barsalou

points out, someone taking this approach must assume that "there are invariant cognitive structures associated with categories that we should be trying to discover" (Barsalou 1987, 114). This implies a pristine symbolic structure hidden beneath the messy wetware of the brain. There are a number of reasons that mitigate this explanation of graded structure.

First, classical categories are late-comers to, as well as exceptions in, the conceptual universe. Real world categories such as *bird* and *tiger* preceded those of *odd number* and *regular verb*. It is likely that the appropriate wetware for handling family resemblance categories evolved first, and that the ability to handle classical categories was developed out of this.

Second, this reasoning denies the usefulness of graded-structure categories. For example, the degree of belonging to the category of *dangerous things* could be immediately correlated with degree of response in an organism to the presence of a dangerous thing.

Third, there is a lack of evidence confirming the existence of such pristine structures. Graded-structure effects are real, and the human brain demonstrates immense variability both between individuals and within individuals. This is not to draw the fallacious conclusion that lack of evidence demonstrates the theory's falsehood, but rather to put the onus on those making the claim in the face of graded-structure effects. An alternative to taking graded structure as indicative of category structure, the core+identification theory, which holds that there are classical cores to categories and incidental properties that give the appearance of prototype structures, fails to explain how hedges like *strictly speaking* and *technically* incorporate incidental properties into the core of a category (Lakoff 1987).

Finally, what would count as evidence in favor of such a cognitive structure? Given that graded structure effects are to be ignored, aren't we then in danger of arbitrarily selecting

the results that back the theory? Nongraded-structure effects would be arbitrarily designated to be truly indicative of cognitive structure.

What graded-structure effects and their instability show is that humans qualify and update their conceptual structures over time. This indicates that humans not only face the general frame problem, but adequately overcome it. Arguments such as that from McDermott (1987) that AI systems generally don't face the frame problem and therefore we shouldn't be too concerned about it only further indicate the distance between AI and true cognitive systems.

6.4.2 Why Dynamic Systems Theory Is better Suited to Explaining Graded Structure

The argument that graded-structure effects do not reveal the nature of the conceptual structure of categories stands in contrast to another inference that computationalists normally make. The symbolic nature of language is supposed to reveal symbol manipulation mechanisms working in the brain. In both cases, however, we may infer mechanisms that produce the effects in question. This does not necessarily indicate precisely which mechanisms are needed to produce the effects; a range of mechanisms may be consistent with what is observed. Whatever mechanisms are postulated, they should not be inconsistent with observed behavior.

The dynamic systems approach to explaining categorization effects and the mechanisms underlying categorization is to take graded structure to be indicative of how the brain carries out these operations (see Thelen and Smith 1994). There is no classical core concepts hidden behind the observed phenomena. The instability of concepts due to context observed by Barsalou points to a mechanism sensitive to variations in context. What many

researchers discard as noise is really a feature of the brain's categorization system. A category, according to the dynamic systems approach of Thelen and Smith,

> is created in context in a trajectory of internal activity in time. The trajectory *always* is a complex product of the immediate context, the just prior internal activity, and the history of reentrant mappings between the heterogeneous processes that make up the system. (Thelen and Smith 1994, 182)

Reentrant mapping refers to a mechanism by which various neural systems coordinate their responses. Simplistically, reentrant connections are connections from systems receiving input back to their sources. According to Gerald Edelman (1987), reentrant connections are not a form of feedback of information, but a means for selection of function. More will be said about this later. Thelen and Smith's contention is that categories are formed through the self-organization of neural systems as they move into attractor states, a process dependent not only on the internal working of the brain but also the environmental context. This is to identify categorization not with a symbolic structure, but with an activity of a system of brain, body and environment. Even the prototype theory of conceptual structure, that concepts are clusters as suggested by prototype effects, is rejected:

> the behaviors that suggest criterial properties, essential properties, and graded structures are all temporally specific manifestations of interacting processes. They are the behavioral and context-specific products of the activity of knowing, not the structural components of knowledge. (Thelen and Smith 1994, 182)

Graded structure effects occur when the systems responsible for categorization are characterized by shallow attractors. The emergence of classic categories is due to the deepening or strengthening of these attractor states. What this means is that the various inputs that could previously move the system out of one attractor state into another no longer effect the state of the system.

The correlated or time-locked activity of neural subsystems do not produce 'codes' that other neural systems 'read' and interpret. No part of the brain reads off attractor states of neural subsystems to determine what they are representing. The dynamic systems interpretation is that activity in neural systems that becomes correlated for specific inputs *is* categorization. Connections from these correlated subsystems to other systems then also can correlate their activity, and a process of *re*categorization proceeds. Recategorization can reduce sensitivity to features of the input, thereby abstracting the initial categories.

Beyond categorization effects and the processes in the brain that produce these effects, there is nothing more, no hidden conceptual structure. There are performance and process, but no competence. This explanation of mind/brain may be understood as a form of Behaviorism, the characteristics of which will be explicated later. First, I take a closer look at the relation between symbols and attractors.

6.5 Can Attractors Take the Place of Symbols?

Pinker and Prince (1999) suggest that two separate and distinct mechanisms are needed to explain the existence of classical and graded categorization. Pattern associators, such as those proposed by PDP modelers, are postulated for graded categorization. Pattern associative mechanisms, however, are inadequate, according to Pinker and Prince, for explaining how classical categories arise:

> Classical categories are the product of formal rules.
> Formal rules apply to objects regardless of their content—that is what "formal rule" means.
> Pattern associators soak up patterns of correlation among object's contents—that is what they are designed to do.
> Therefore, pattern associators are not suited to handling classical categories.

> We conclude that the brain contains some kind of non-associative architecture, used in language, and presumably elsewhere. (Pinker and Prince 1999, 27)

As a critique of PDP methods, this is entirely unobjectionable. PDP systems do indeed "soak up" patterns of correlation among object's contents: "soaking up" is Pinker and Prince's phrase for the superposition of information. Similar inputs generally produce similar outputs in most ANNs. As pattern correlators, this is a desirable trait, but, as Pinker and Prince point out, this makes them unsuited for activities that consist of applying formal rules.

As has been noted, nonlinear dynamic systems have bifurcation points, which are points in parameter space where the dynamics of the system change dramatically. Also, chaotic systems have strange attractors, which are extremely sensitive to initial conditions. Similar inputs do not necessarily mean similar outputs in such systems. In fact, quite similar starting points for such a system can lead to radically different behavior. So, even if PDP models "soak up" correlations, making it difficult if not impossible to produce substantially different behavior on similar inputs, this does not imply that a biological neural network must be implementing a symbol manipulation system in order to produce non-associative behavior. The question still remains whether nonlinear dynamic systems can produce language, whether their behavior can conform to rule-following.

6.5.1 Can Attractors Figure into Formal Systems?

The answer to the question that heads this section is clearly no. Attractors are not stand-ins for symbols, which a symbol manipulation system reads off and applies formal rules to. Attractors describe relatively stable behaviors of a system, and as such are not tokens or

structural components; for example, a limit cycle attractor describes the oscillatory behavior of coupled neuron-firing.

The challenge facing the dynamic systems theorist is to explain the emergence of symbolic behavior without reference to internal symbol manipulation. Thelen and Smith (1994) argue that a bootstrapping theory such as Pinker's (1989), which holds that syntax arises out of semantics, explains how categorization mechanisms in dynamic systems give rise to syntactic mechanisms. But this would only explain the origin of syntax, not its functioning. If the mechanisms that bootstrapping procedures produce are symbol manipulators, then bootstrapping gains the dynamic systems theorist nothing. If they are not symbol manipulators, what are they?

Dynamic systems theory is not poised to answer this question at the moment. In fact, a general answer from dynamic systems theory is probably not in the offing. The reason is that the complex systems that give rise to language are likely to be too specific to explain in terms of, say, Hopf bifurcations. They undoubtedly have quite unique dynamics, which can only be discovered through investigation of brain structures responsible for language.

The explanation that dynamic systems theorists can offer is that the manipulation of external symbols in behavior such as language use or map-making is an aide to the internal dynamics of thought and not a reflection of it. But this is little more than a slogan until an explanation of the internal dynamics underlying language use is given. All that is known is that some nonlinear dynamic systems can produce what can be characterized as non-associative behavior. This leaves open the possibility of a dynamic systems account of language use.

6.5.2 Why Internal Symbols Are Not Necessary

Earlier chapters investigated the question of whether internal symbols and symbol manipulators are sufficient for cognition. The question now turns to whether they are necessary for language and other blatantly, if only externally, symbolic behavior. As Brooks (1990) points out, symbolic behavior is the exception, not the rule, among biological organisms. If dynamic systems theory can account for perception, categorization, and simple associative reasoning, it will have covered not only the majority of cognition functions, but also the bases for semantic interpretation of language.

This returns us to Clark's notions of 'representation-hungry' processes and the partial programs that implement them (see sections 2.6-2.7). Recall that partial programs are programs that carry out the representation-hungry portion of a task and interface with dynamic systems for input and output from the rest of the task. Although enticing as an explanation of the dichotomy in types of cognition (pattern recognition vs. rule following), we have already seen why such an approach is implausible. However, if we were to suppose that the dynamic systems theorists are correct in rejecting all internal symbols, we must nonetheless admit that the behavior of dynamic systems can be mapped to external symbol systems. This happens every time someone writes or speaks a sentence. How might this be done?

A minimal requirement is a system to organize the representational dynamic systems into sequential processes. Lieberman (1984) contends that the sequential organization of representations carried out by the language subsystems of the brain are an adaptation of neural mechanisms for controlling motor sequences. But the ability to organize sequences does not imply the need for symbols and rules. Kelso (1997) has developed dynamic systems explanations for the emergence of horse gaits and finger-tapping sequences without

recourse to symbols and syntactic mechanisms for sequencing them. But mere sequencing is not sufficient to explain language ability, because language demonstrates recursive structure.

Horgan and Tienson (1996) claim that it is possible to have syntactically structured representations, which would allow for recursion, within a dynamic systems framework, and, therefore, without symbols and rules for manipulating them. They propose that the mind/brain is a dynamic system that gives rise to a language of thought, and this language of thought has syntax that demonstrates 'systematicity' and 'productivity' in encoding semantic relationships. Systematicity means that the fact that two or more representations predicate the same feature to an object is encoded in the structure of the representations. Productivity means that the acquisition of representations of new properties automatically determines representations that predicate the properties to individuals. These two properties are meant to capture the constituent nature of syntax without relying on a part/whole relation. To illustrate that dynamic systems can demonstrate these properties, Horgan and Tienson cite two examples of connectionist systems that possess them: Pollack's (1990) Recurrent Auto-Associative Memory (RAAM) and Smolensky's (1990) tensor-product representations. RAAM networks and their variants have been employed (Berg 1992; Callan and Palmer-Brown 1997) as parsers, and have demonstrated some ability to handle recursive structure. Tensor-product representations consist of vectors standing in for representations. Representations can be of linguistic roles, such as *subject*, and fillers, such as the noun *John*, which are bound by tensor multiplication (a dot product). Sentences are constructed through vector addition of bound representations. A neural network trained by a

variant of recursive backpropagation is then used to learn sentences and unbind fillers from roles in response to queries.

One last property completes Horgan and Tienson's characterization of the language of thought: representations must play an appropriate causal role within the cognitive system. In other words, the syntactic structure of a representation helps determine the causal role of the representation.

Research in connectionist linguistics has produced alternatives to the part/whole understanding of constituency. Language production might be realized not through the application of rules of composition to symbols, but through a process similar to these connectionist alternatives. So rather than being saddled with the necessity of rules and symbols, we are left with Horgan and Tienson's contention that a language of thought as they conceive it is a necessary condition for a cognitive system:

> Cognizers must have a language of thought because getting around and surviving in the world requires a representational capacity so vast that it is possible only for a cognizer that has a systematic way of constructing representations—i.e., syntax. Thus, natural cognitive systems must have syntactically structured representations, and hence a language of thought. (Horgan and Tienson 1996, 71)

But if a language of thought is necessary, why wouldn't a classical symbol system suffice? In fact, wouldn't a classical symbol system be more desirable, given that it could handle constituent structure in terms of part/whole relationships?

6.5.3 Horgan and Tienson's Dynamical Systems Hypothesis

Horgan and Tienson reject classical (rule-based) solutions to the supposed need for syntactically structured representations because they believe that classical solutions require exceptionless, programmable rules that operate over tokens acting as representations. Their

reasoning for rejecting such rule-based systems is as follows. Exceptionless rules imply the need for a vast number of rules enumerating every possible exception, i.e., each exception must be turned into a rule. There must also be rules for transitions (Cognitive Transition Functions or CTF) between 'total cognitive states', which are the intentional states of the cognitive system that can be composed of one or more simultaneously instantiated cognitive states, which are in turn characterized by the various rules concerning how representations are formed. The transition rules must also be exceptionless, creating an even greater combinatorial explosion. The transitions between total cognitive states are, therefore, not 'tractably computable', which is to say that they are not computable "by a *physical* computational device of the relevant kind" (Horgan and Tienson 1996, 26). Horgan and Tienson identify this as the frame problem, although it corresponds more closely to the update problem.

Horgan and Tienson's argument in favor of dynamical systems theory over classical approaches is rather crude:

1. CTFs are not tractably computable by classical approaches.
2. The transitions between states in most dynamical systems are not tractably computable.
3. Therefore, CTFs must be subserved by noncomputable dynamical systems.

This negative argument is quite different from the positive argument that certain dynamical systems possess properties which classical systems do not, which are in turn necessary for realizing cognitive functions. Horgan and Tienson only briefly mention nonlinearity and entirely disregard self-organization and the slaving principle. One reason for the latter

omission is that their conception of dynamical systems is hierarchical. They envision three levels of what they call 'Noncomputable Dynamic Cognition', which is simply cognition realized by a dynamical system, the transitions of which are not computable. The levels are: cognitive-state transitions, mathematical state transitions and physical implementation. The first is what they consider to be the level of the mental *qua* mental; it is the level of transitions between total cognitive states. The second is the level of the dynamic systems, which is a mathematical description of the cognitive system. The third is the neural network that subserves the dynamic system. There is a unidirectional realization relation that links these three levels:

> Cognitive states are realized by mathematical states, and mathematical states are realized by physical states; thus by the transitivity of realization, cognitive states are realized by physical states. (Horgan and Tienson 1996, 146)

Points on the activation landscape realize cognitive states, and the semantic relations between cognitive states is structurally embodied through relative-position relations. Positions that are close, as measured in the n-dimensional activation space or in one of the innumerable subspaces, on the activation landscape are close semantically.

Ironically, Horgan and Tienson's position closely resembles the eliminativist form of connectionism proposed by Paul Churchland. Churchland (1989) similarly conceives of semantic relations being expressed through relations within the activation space of a connectionist network. Churchland, however, rejects the idea that points in activation space correspond to representations and that distance relations between points determine similarity. Instead, he views representations as partitions of the activation space, with the partitions being determined by the functions computed by the network. A representation space could therefore have a highly nonlinear contour and need not consist of continuous regions.

Points in activation space do not necessarily correspond to similar representations if they are in similar positions, as they may fall within different partitions. Although Horgan and Tienson stipulate that the topography of the activation landscape can also determine similarity—two points on the same attractor basin could be similar due to the shared topographical position—they nonetheless hold that points in activation space are what subserve representations.

Like Churchland, Horgan and Tienson argue that a connectionist architecture subserves cognition, and that representations produced by the system correspond to portions of the activation space of the connectionist network. Like Churchland, they reject the view that cognition consists of symbolic computation, with rules being applied to manipulate and produce symbol tokens. Yet, Churchland also rejects the notion that there is a language of thought, whereas Horgan and Tienson argue that a language of thought is a necessary condition not only for linguistic behavior, but for cognition in general. So who is right?

6.5.4 Why a Language of Thought Is Not Necessary

Horgan and Tienson's argument that cognitive systems need a language of thought can be characterized as an *argument from bigness*. Only a systematic method of constructing and structuring representations can enable an organism to handle the vast complexity of the world in which it lives. This argument from bigness is really an argument from complexity. It is both the number of representations and their many interactions that impress upon Horgan and Tienson the need for syntactic structure. But explaining the emergence of structure from vastly complex and unsystematic interactions is the very thing that the dynamic systems approach offers. The production of coherent light in lasers is not the result of syn-

tactically structured interactions among its constituents. If coherent light can be produced without syntactically structured interactions among its elements, why can't syntactically structured speech be produced without syntactically structured manipulation of its elements? Should we postulate a language of lasers to explain the production of coherent light from such vast complexity?

Horgan and Tienson's argument extends not merely to behavior that is obviously syntactically structured, such as language use. They hold that all representations must be constructed via a language of thought. What they mean by a language is quite different than what is traditionally understood. Or so they think. Consider their example of a language with nonclassical syntax:

> Imagine, for example, a language in which there is one class of pure wave forms that can be used as proper names and there are other classes that can be used as general terms, relational terms, adverbs, connectives, and so on. When a general term is predicated of an individual, the general-term wave form and the individual name wave form are produced simultaneously. Sentences are analogous to *chords*, not to tunes. Slight conventional modifications of the basic pure wave forms indicate that a word is a direct object, an indirect object, etc. Sound waves, like all waves, superimpose; so in the chord none of the individual waves that went to make it up is tokened. (Horgan and Tienson 1996, 73)

This example nicely demonstrates how a language could have constituency relationships without part/whole relationships or decomposability. It is, however, an example of a rule-based symbol manipulation system, albeit an external one. Wave forms are symbols in this language, and their are rules for composing them to produce symbols with syntactic roles. As an external system, it does not conflict with the dynamic systems approach, because it does not imply that there is a rule-based symbol manipulation system in the mind/brain. But Horgan and Tienson argue that what goes on in the mind/brain could be analogous to this language production system:

> If there are creatures with such a system of communication, it would hardly be reasonable to deny that they have a language, or to say that their language lacks syntax, on the grounds that their system of communication lacks classical constituency. Likewise, if a system of mental representations encodes predication in a similar manner, it would be inappropriate to deny it is a *language* of thought. Our suspicion is that the language of thought is more like this than it is like the first-order predicate calculus or LISP. (Horgan and Tienson 1996, 73)

If a language of thought is indeed like this example, then Horgan and Tienson's position that mind/brains are not rule-based systems manipulating symbols is false. The only aspect of this language that could be considered non-symbolic is the superposition function to produce the chords. This could be achieved through a numerical solution. At the very least, Horgan and Tienson would have to give up the *noncomputable* aspect of their theory.

An important reason for adopting the dynamic systems approach is that self-organizing complex systems are more powerful, in the sense of the range of behavior they can produce, than rule-based symbol manipulation systems. Horgan and Tienson seem to acknowledge this in their argument that cognition must be subserved by a system the transitions of which are not tractably computable (assuming their argument is not just a crude analogy). Their case for a language of thought, however, contradicts this, suggesting that only a rule-based symbol system can construct representations that enable an organism to survive. It would make more sense to wed dynamic systems theory with classical syntax, as Clark has tried to do, than to posit a weaker form of syntax underlying thought, because it would at least be able to explain more easily the production of language that does contain part/whole relationships.

Rule-following behavior does not imply rule-following mechanisms underlying it. Explaining how such behavior arises remains the primary challenge for dynamic systems theory. It is an empirical challenge, because, as we have seen with lasers, there is nothing

about structured behavior that in principle prohibits nonlinear dynamical complex systems from producing it. The processes through which perceptual representations arise, however, have received greater attention, with theories such as Edelman's Theory of Neuronal Group Selection (TNGS) offering possible explanations.

6.6 Filling in the Details: How The Theory of Neuronal Group Selection Explains Why Certain Dynamic Systems Are Capable of Cognition

Edelman's Theory of Neuronal Group Selection has been discussed throughout this dissertation in a variety of contexts. In this section, I endeavor to overview its salient features and show how it fits into a dynamic systems theory of mind/brain. A similar effort has been undertaken by Thelen and Smith (1994), although they primarily highlight its importance for a theory of development and give less attention to explaining higher cognitive functioning. I will focus on how the TNGS explains the emergence of perceptual categorization, and how the mechanisms underlying perceptual categorization lead to memory and more abstract categorization.

6.6.1 What Is Neural Selectionism?

As a *selectionist* theory, the TNGS stands in contrast to *instructionist* theories of the brain that posit that the function of neurons and neuronal groups is to pass information and instructions between neural structures indicating what the neurons and neuronal groups have detected or what the receiving neural structure should do next. Instructionists often liken the brain to a computer, with instructions being passed between neuronal groups as if they were executing a program. What is important to instructionist theories is the informa-

tional function of neuronal behavior. Individual variation in neurons is unimportant and generally considered noise.

The TNGS, as a selectionist theory, takes the existence of individual variability among neurons, neuronal connectivity, and the make-up of neuronal groups to be essential to the emergence of global neural behavior. The TNGS applies the population thinking of evolutionary theory to neuronal groups, and, therefore, the emergence of massive variation in the above-mentioned neural components is essential for useful neural functioning to arise. Competition between neuronal groups that perform similar functions, referred to as *degenerate* structures, leads to a process of selection. Successful neuronal groups are not, however, differentially reproduced, as is the case with species or individuals according to standard evolutionary theory, but rather they are differentially *amplified*, meaning that their connections are strengthened and they are more likely to be triggered in similar situations in the future.

In addition to providing an explanation for the existence of extreme variability in neural structures within species and individuals in terms of evolutionary utility, selectionist accounts also make sense of the fact that the majority of connections within brains are not functionally expressed (Edelman 1989). These unexpressed connections are from degenerate groups that have lost out in the process of selection.

6.6.2 Basic Mechanisms of Neuronal Group Selection

There are three basic mechanisms postulated by the TNGS: developmental selection, experiential selection, and reentrant mapping. In development, neuronal connections branch and diversify, forming what Edelman refers to as the *primary repertoires*. These are neuronal

groups within given anatomical regions that vary in connectivity and function. There are massive numbers of these variant groups within each region, and they form through cell division, migration, death, process extension and neural activity (Edelman 1989). Unlike Darwinian natural selection, the selective processes during development are not merely eliminative, but also productive.

Experiential selection begins after most of the connections of the primary repertoires are fixed. Experiential selection primarily produces synaptic alteration—the strengthening and weakening of connections—rather than producing new connections:

> during behavior particular functioning neuronal groups are dynamically selected by the action of various signals and mechanisms of synaptic change. This selection occurs among *populations* of synapses, strengthening some synapses and weakening others, a process that leads to the formation of *secondary repertoires*. The consequence is that certain circuits and neuronal groups in such repertoires are more likely to be favored over others in future encounters with signals of similar types. (Edelman 1989, 46)

It is important to note that selection is not governed by a global, regulatory mechanism that picks out winners and losers any more than Darwinian natural selection is. Instead, selection occurs through localist interactions between signals and neuronal groups. Also, selection specifically works on neuronal *groups*, and not individual neurons, and so should not be confused with mechanisms such as long-term potentiation.

Reentrant mapping is "temporally ongoing parallel signaling by separate maps along ordered anatomical connections" (Edelman 1989, 49). Reciprocal connections between neural maps are simple examples of reentrant mapping. Whereas these connections are generally interpreted as feedback loops in instructionist theories, reentrant mapping, according to the TNGS, enables more complex selections to occur:

> Signaling in either a phasic or continuous fashion across reentrantly connected maps permits temporal correlations of the various selections that occur among neuronal groups within these maps. (Edelman 1989, 49)

Local processes of selection become global through the contribution of reentrant connections. The emergence of global maps is what underlies perceptual categorization.

6.6.3 Creating Minds: Reentrant Mapping, Categorization, and Memory

The correlation of selected groups within maps with specific types of signals leads to reactivation of these groups when similar signals arrive. This correlation, plus the connections of the correlated maps to behavioral and memory areas, produces the global mapping necessary for perceptual categorization:

> A global mapping is a dynamic structure containing *multiple* reentrant local maps (both motor and sensory) that interact with nonmapped regions such as those of the brain stem, basal ganglia, hippocampus, and parts of the cerebellum. The activity of global mapping connects neuronal groups selected in one set of local maps (as a result of the activity of feature detectors) to neuronal groups selected in other sets of maps (as a result of the correlation of features that is established, for example, by the continuity of motion.) (Edelman 1989, 54)

Perceptual categorization occurs as neural maps self-organize around incoming signals through processes of selection and reentrant mapping.

Categorization involves the grouping of disparate signals and responses to signals to produce a unified response. Edelman identifies three ways in which reentrant mapping is able to integrate signals:

1. By resolving conflicts between responses of different areas or different groups within areas to the same signal

2. Through *cross-modal construction* of responses, which is when one area uses the output of another area for its own particular function (e.g., when features such as color and illusory contour feed into areas for motion detection, and thereby perceptually undergo apparent motion)
3. Through *recursive synthesis*, which is when higher-level maps influence the inputs they receive from lower areas such as those responsible for perceptual categorization

Reentrant mapping also allows for the recategorization of signals. Neuronal groups reentrantly connected to the groups involved in perceptual categorization can correlate their activity with selected aspects of the activity of the perceptual groups. Memory arises through such reentrant connections and further associations with what Edelman refers to as *value*. Value is an inherent bias in the brain for achieving particular goals, which consist of evolutionarily determined needs such as homeostatic regulation and reproduction. Memory is the 'storage' of previous connections between perceptual categorizations and value systems. Recall is the recategorization of the activity of these associated areas. What Edelman refers to as *primary consciousness* arises through the activity of these memory repertoires and reentrant signaling between them and sensorimotor groups:

> Primary consciousness thus emerges from a conceptually based recategorical memory (relating *previous* value-category sequences) as it interacts with current input categories arising from the neural systems dedicated to *present* value-free perceptual categorization. It constitutes a *discrimination* of the acquired self-nonself memory from current ongoing perceptual categorizations. (Edelman 1989, 97)

Systems related to the self are the homeostatic brain functions, i.e., the value systems, including the autonomic, hedonic and neuroendocrinal systems. Conceptually based memory recategorizes associations between past perceptual categories and value, produc-

ing both modal (directly related to a particular sensory system) and amodal concepts. Primary consciousness is the process of relating, comparing, and updating priors concepts and current perceptual categorization. A precursor to higher-order consciousness is the system, presyntax, which, according to Edelman, consists of reentrant links between temporal regions and the primary mechanism for conceptualization, the frontal cortex (with its connections to value systems), and which orders concepts successively. This system is the basis for thought, and is present in primates such as chimpanzees.

Higher-order consciousness frees the individual from the demands of the moment and enables him to think about objects and events without connection to the real-time flow of the world. Edelman defines higher-order consciousness as "the capacity that sustains direct awareness related to plans" (Edelman 1989, 173). Essential to the development of higher-order consciousness is the emergence of linguistic capabilities. Edelman adopts a semantic-bootstrapping account of the development of syntax in humans similar to that of Pinker (1989). Semantics arises from the conceptual systems described above, and is coupled to phonological elements produced by a phonological system that appears on the scene as an evolutionary novelty. The coupling of semantic and phonological elements, again through reentrant connections, provides the first requirement for syntax: the lexicon. Then,

> when a lexicon is sufficiently developed, the conceptual apparatus may recursively treat and classify the various productions of language themselves—morphemes, words, sentences—as entities to be categorized and recombined without any *necessary* further reference to their initial origin or to their bases in perception, learning, and social transmission. (Edelman 1989, 174)

In contrast to Horgan and Tienson, Edelman considers linguistic capability an enhancement to thought, one that allows for concept formation not fettered by the immediate demands on the organism. This account coheres much better with the natural world. It is odd that we

should consider a lamprey or beetle in need of a language of thought in order to just survive, and yet not observe them using its potential. What might pass as 'planning' in a rough sense in the natural world, such as storing food for the winter, is really scripted behavior in that it is preset in the organism to carry out this behavior—as we have seen, this is the distinction highlighted by cognitive roboticists to distinguish their work from robots executing scripts. So, although a language of thought is not necessary for representing the world or surviving in it, linguistic capability is the essential component in the production of symbolic behavior. Linguistic production is achieved, according to Edelman, without the need for explicitly represented, pre-programmed rules; instead, rule-like behavior emerges through reentrant connections and learning.

6.6.4 Why Neuronal Group Selectionism Does Not Imply the Perception of Affordances

Although Edelman has linked his theory to Gibson's ecological approach, he has given few indications of where the two theories intersect. I propose that a selectionist theory of brain function is essential to explaining how direct perception of complex entities such as surfaces might arise. This is due to the fact that selectionist theories alleviate the need for representations to be formed by composition of primitive informational elements. According to Edelman's theory, however, mechanisms for perceptual categorization and value systems are distinct neuronal structures, and their connection requires another layer of reentry. If true, this rules out direct perception of what the environment affords.

Selectionism enables direct perception of complex entities in that the neural ensembles responsible for perceptual categorization are selected from a host of degenerate (in Edelman's sense) ensembles for the way they react to a set of incoming signals—in other words,

how they treat the set of signals as a whole event. Each degenerate ensemble receives connections from the input sites of perception and correlates its activity with these sites. The winning ensemble determines how the system reacts to the signals as a whole. This allows the system to bypass hierarchical steps of combining simple features into more and more complex features. The key is the existence of degenerate structures from which an appropriately behaving structure can be selected. With sufficient variation in the structures, a useful one is likely to exist. The selected structure would be equivalent to combining the functions of lower-level feature detectors without requiring this lower-level as input.

An analogy to this can be found in Darwinian evolution itself. It is rare to find a one-to-one mapping between trait and gene; more often it is a one-to-many or many-to-many mapping between traits and genes. This is to say that gene complexes and not single genes are usually responsible for traits, and that gene complexes can produce more than one trait. Traits, such as complex behaviors, are often not decomposable into primitive elements, and primitive elements of traits generally cannot be mapped to primitive elements of the gene complex responsible for them. What is selected for is a complex trait in itself and a gene complex in itself, not individual genes to build the gene complex (in which case it would not really be a gene ‘complex’). Similarly, a complex attitude toward a complex signal might be selected for by selecting a particular neuronal ensemble. Thus, a neuronal ensemble might categorize a collection of primitive incoming signals as a surface without piecing together edges, lines, and planes.

It is important to note that neural selectionism only provides a *plausible* mechanism for the emergence of direct perception. It is also plausible in a selectionist account that perception is not direct. The perception of surfaces might be the result of several layers of struc-

ture selected for their responses to various levels of structure in the incoming signal. It is also important to reemphasize that Edelman conceives of perceptual categorization and value systems as separate entities. According to the TNGS, the perception of affordances is *not* direct, but rather requires reentrant activity between distinct systems. Only if the selection of perceptual categorization structures were dictated by the value systems would it be possible for direct perception of affordances in the manner described above for surfaces.

Neural selectionism allows for varying degrees of directness in perception. It enables a system to 'skip' steps in composing signals, to produce complex behaviors in response to complex signals without having to decompose the signal into informational primitives and analyze their relations. In Darwinian evolution, species evolve through selection of complex traits from an abundance of alternatives. It seems likely that complex mental behaviors would evolve through a similar process, rather than nature having to produce an architecturally pristine structure such as a computer.

6.7 Selectionist, Self-organizing Dynamics Underlies the Functioning of Minds

Timothy van Gelder illustrated the differences between dynamic and computational systems by comparing the Watt centrifugal governor with a hypothesized computational governor, an illustration that captures the basic differences between the two types of systems but does not address how cognition arises in dynamic systems. To address this, dynamic systems theorists such as van Gelder point to connectionist systems as examples of how dynamic systems implement cognition. But as we have seen, connectionist systems do not

bear important differences from traditional AI systems, and computer implementations of neural networks do bear many of the drawbacks for which van Gelder criticizes traditional AI. Even if we establish that certain dynamic systems are not implementable via computational systems, and that the mind/brain is one such system, we have not established what it is that makes the mind/brain different from a Watt centrifugal governor.

The aspects of minds that set them apart from dynamic systems in general are their ability to produce and reproduce representations. Representations arise as attractor states of the dynamic system through a process of selectionist self-organization. Selectionist self-organization occurs through the competitive interaction of neuronal groups, which leads to coordinated activity between neuronal groups through the strengthening of connections between successful groups. This process occurs without explicit or implicit rules being executed in the system and produces coherent behavior correlated with signals from the environment. Reentrant connections between groups enables reproduction of this behavior absent the presence of environmental signals, allowing for reproduction of representations. This is the basis of memory.

The nature of attractors in the mind/brain dynamic system helps to explain cognitive phenomena such as graded structure in concept formation and production. Shallow attractors are what produce graded-structure-concept effects. Deeper attractors enable cognitive systems to employ classical categories. Although connectionist systems are capable only of associative operations on these concepts, dynamic systems can perform a wider range of operations, as small changes in parameters can give way to starkly different behaviors in the system. Demarcations need not be gradual, but rather can be sharp and dramatic.

Thus, a language of thought is not required to explain how dynamic systems carry out what is considered syntactic behavior. In fact, employing a language of thought to explain how dynamic systems can do syntax weakens their appeal, for then they reduce to neurally implausible connectionist systems. A more neurally plausible explanation of how neural dynamic systems achieve categorization, memory, and perception is given by Edelman's Theory of Neuronal Group Selection. How this sketch of the mechanisms behind the brain bears on what the mind is I turn to in the next chapter.

Chapter 7 Putting it all Together: An Eliminativist Theory of the Mind

Computationalism introduces a layer of mental architecture between the behavioral outputs of a system and the mechanisms that produce them. This layer consists of chunks of information, symbols, which figure into rules, whether implicit or explicit, which in turn produce more chunks of information. This additional level of explanation was developed to avoid the objections that doomed behaviorism. In so doing, computationalists have created their own set of problems. Computationalism places constraints on the realizing physical hardware: it must have mechanisms that subserve symbols, such as stored electric charges in computers. A theory that proposes mechanisms that cannot be decomposed according to their informational content must be rejected out of hand.[1] This in turn places constraints on what the hardware can do, and from this emerges the frame problem.

But the frame problem is more a symptom of a deeper disorder rather than the disorder itself. This deeper affliction is computationalism's inability to explain what makes a piece of information *information-about-the-world*, what makes it *semantic information*. One aspect of this affliction is the symbol-grounding problem. But accepting the symbol-grounding problem as formulated by Harnad (1990) suggests that there is a solution, that we can overcome computationalism's fundamental difficulty if we just figure out the

1. This is the gist of one of Clark's objections to van Gelder's Dynamic Systems account of mind (see Clark 1998).

proper way to link symbols to the world. And this in turn rests on the assumption that symbols can be grounded in the first place.

Not all computationalists believe that symbols even need to be grounded. If we follow this line further, we are led to the conclusion that grounding symbols does not change their fundamental nature. In this case, we must reject the notion that symbols can be grounded in any meaningful sense of the term. In psychological terms, we are left with a pure competence theory of mental events. Rules are churning within the machine, producing particular mental events regardless of the environment. The variations of performance in perceptual categorization, for example, are simply noise. Whether the machine is situated within the world or not is irrelevant to what is going on inside the machine's head.

Computationalism has diverged into two extremes: one that recognizes computationalism's practical flaws and attempts to patch it, and another that rejects the need for these patches and isolates mind from the world. Those who reject the need for fixing computationalism with kludges such as partial programs linked to dynamic systems have correctly assessed the consequences of computationalism. If computationalism is correct about the nature of mind, then linking symbols to the world is irrelevant. A token is a symbol for something by virtue of its formal properties, not due to a causal link to the world. Thus, those seeking to improve the performance of computational agents by adding noncomputational enhancements are not moving any closer to constructing a mind than those who use purely computational methods.

The alternative to pure computationalism is, therefore, not to link it to the world, but to reject it completely. This means rejecting the intermediate layer between behavior and realizing mechanisms, rejecting the layer of concepts and information-processing. Cognitive

scientists reintroduced this layer to explain what behaviorism could not. They are left unable to explain their own explanatory device. They have introduced entities with properties so distinct from their realizing hardware that they require accepting the principle of property dualism. An approach that rejects the existence of these mental entities or properties over and above the properties of the systems that supposedly bear those mental characteristics would, however, reintroduce a form of behaviorism. The question is whether it would also harbor the same failings that lead to the almost universal rejection of behaviorism and the cognitive turn.

7.1 Dynamic Systems Behaviorism: An Alternative to Computationalism

It is the belief that the mind/brain creates and stores discrete concepts that underpins the notion that the mind/brain is a symbol processor. It is this belief that must be rejected. I propose that the mind is instead a dynamic system incorporating brain, body, and environment. The states of this system are not symbols standing in for concepts, but attitudes of the brain/body to the environment. These attitudes are formed by interaction with the environment, and are only abstracted from their environmental context by directing neural resources toward the processes that form these attitudes: the process of recategorization as described by Edelman. As one moves up levels of abstraction, contributions from the body and environment become less and less, and the attitudes become dependent more on neural processes. They are attitudes toward attitudes, neural processes directed toward brain/body processes that emerged within a particular environment. Computationalism tries to skip this hierarchy and move immediately to abstract symbols, but it is the lower level that enables the production of more abstract attitudes. By abstract attitude, I mean a state of the system

that does not require immediate environmental stimulus for it to form, or that can be formed in response to individuals across a type. This is to say that its formation is less sensitive to environmental context.

So what is an 'attitude'? At the most basic level, that of perception and perceptual categorization, it is an attractor state of the brain/body/environment system. It is a neural pattern that emerges from the interaction between signals from the environment that are sent through the body as a result of bodily interactions with the environment and the neuron ensembles. Attitudes are not like the dispositions of behaviorism. They do not necessarily dispose the system toward overt behavior, although that is one function of them. They can also dispose the system toward particular internal behavior, namely the formation of other coherent patterns within the brain. This is done by alteration of brain chemistry in connections between and within groups of neurons, and not by storing a particular pattern at a particular spot. The formation of attitudes through the self-organization of neuron ensembles with reentrant connections makes possible the recreation of these attitudes by the system itself without environmental stimulus—the abstract attitudes described above.

Thus, in addition to the formation of a particular type of neural pattern within a particular context, whether the context is environmental or internal to the brain, attitudes are also distinguished by their effects, whether internal or external. For example, the process of categorization consists of the formation of an attitude that produces category effects, such as enabling the organism to react to a type of predator rather than having to decide if each individual of the type is a threat. Following Thelen and Smith (1994), the nature of the pattern formed is what produces the instability effects related by Barsalou (1987); these effects are the result of unstable patterns. There is nothing beyond the formation of attitudes and their

effects, no informational structure that stands in for an object and is stored somewhere in the brain. What are stored are changes in neuronal chemistry that dispose a particular neuron or neuron ensemble to participate in the formation of a pattern. These dispositions are distributed across many neurons, and are superposed with other dispositions.

The attractor states of a system or its subsystems are not analogs of symbols. They are not abstract tokens produced through the application of rules. They do not figure into rules. They are continuous-time activities, firing patterns that a system moves into. What this implies is that thought is not a discrete entity produced by the brain, nor the process of producing such entities. It is an ongoing action, the formation of a pattern of activity over time. This activity also does not cease once the pattern is formed, firing once in synchrony and then ending, although it does dissipate as other thoughts are generated.

Functionalists individuate mental states according to their function and their relations to other mental states. Dynamic Systems Behaviorism (DSB) individuates mental states by their function and their interaction with other states. What is required is not a mere semantic relation nor a possible causal link between states, but actual interactions. This is not to say that unexpressed causal links—links between states that could be expressed if, say, the system drew its attention to these links—are unimportant. They also figure into the specification, but of primary importance is actual behavioral interaction between states and systems. That the interactions between particular systems are important distinguishes DSB from computationalism, because these interactions are not between abstract entities. What a system does and how it is causally linked to the world affects the meaning of the links from other systems to it. Thus, Damasio emphasizes the retroactivation of areas related to perception as important for determining the nature of a representation, and Edelman simi-

larly points to the reentrant connections to perceptual areas for producing perceptual categorization. A symbol system would have no need to have links to particular areas, since the symbols they produce could be stored elsewhere and links could be maintained to these symbols.

The notion of a 'mental state' in DSB therefore bears a different meaning than in computationalist theories. It is simply a specification of the parameters of a system, its structure and behavior. Mental states are eliminated in DSB not because they are reduced to more fundamental physical properties, such as reduction to a neuron ensemble firing, but because they are exhausted by a neural dynamic systems description. There is nothing to a mental state beyond its specification according to a dynamic systems explanation of the neural activity involved and the effects it produces. The slaving principle captures the interaction between 'mental state' (the global order) and the neural elements 'realizing' it.

Thus, DSB avoids anti-reductionist arguments such as that of Crane and Mellor (1990) by accepting them. Crane and Mellor argue that reductive physicalism founders on the fact that mental states bear properties that cannot be found in the realizing elements. One example they give to illustrate this is the emergence of behavior in a gas conforming to the ideal gas law that cannot be derived from the individual behaviors of the gas molecules. Such emergent, global behavior is exactly what dynamic systems theory is directed toward explaining. And the nature of this behavior—its causal origin, interactions, and future behavior—is exhaustively explained by dynamic systems theory. The mistake of antiphysicalists is to conclude that there must be a form of property dualism separating the 'mental' from the physical. The only dualism is between the global order and the local interactions from which it arises. They are connected in dynamic systems theory by the slaving princi-

ple and the notion of ‘circular causality’. The illustration given by Crane and Mellor can be turned against them as an argument for physicalism, albeit a nonreductive physicalism. The emergence of global order can be explained by considering the local interactions of its underlying elements. The difficulty lies in the fact that the relationship between the global and local orders is nonlinear. In such a system, we would not expect to find global properties in local elements, nor would we expect to be able to linearly compose local properties to produce global properties. But this does not mean that we must postulate a ‘mental’ order to explain the appearance of global properties not found in the neurons. The relatively young fields of synergetics and complex systems are directed toward explaining such emergence; Crane and Mellor have rejected physicalism too soon.

7.1.1 Behaviorism’s Flaws and How to Avoid Them

DSB does not identify being in a particular mental state either with behaving in a particular publicly observable manner or a disposition to behave in such a way. Instead, DSB identifies being in a particular mental state with a relation between environmental input, internal behavioral output, and the causal relations between the neural subsystems involved. Even this characterization is somewhat incorrect, because a traditional input-output relation presupposes a single direction of causality, which is violated by the circular causality of self-organizing systems.

This is a move similar to that of functionalism (just how similar will be discussed in the next section). The question is whether this move protects DSB from the criticisms leveled against behaviorism.

Behaviorism is open to the charge of chauvinism in its attribution of mental states, which is to say that it denies that a class of persons to whom we would normally ascribe mental states actually possess them. An example, offered by Putnam (1963), is that of super-spartans, who show no publicly observable behavior when experiencing pain, and, to defeat the objection that they would still have dispositions toward pain behavior, super-super-spartans, who do not even have a propensity for pain behavior when experiencing pain. There are two possible responses to such a charge against DSB. First, states like pain, hunger, thirst, etc., could be denied to be attitudinal states. There need not be an attractor state formed for an organism to feel pain. Pain is plausibly a direct stimulation of certain value regions. These value regions would then instigate behavior, unless some other state or subsystem intervened. The possibility of thwarted behavior also answers the objection that behaviorism cannot explain the defeasibility of a behavior-mental state link with invoking other mental states. Second, pain could be allowed to be an attitude of the organism without any necessary behavioral outputs, but with 'normal' outputs (in the sense of normal function). Again, other internal states could intervene under abnormal conditions, preventing normal pain behavior.

Another objection to behaviorism is that complex mental states, such as believing that Martin Bormann died while trying to escape Berlin in 1945, cannot plausibly be explained in terms of publicly observable behavior or dispositions to publicly observable behavior. Again, DSB makes the similar move as functionalism and relies on an intricate web of attitudes toward attitudes, all behaviorally expressed in the brain, but not necessarily by the body.

The fact that DSB uses the same maneuvers as functionalism to escape behaviorism's flaws suggests that DSB is just a flavor of functionalism. There are, however, important differences, to which I turn now to highlight.

7.1.2 How Dynamic Systems Behaviorism differs from Functionalism

As Block (1978) points out, behaviorism and functionalism differ only by virtue of functionalism's stipulation that mental states are defined not only by their input-output relations, but also by their dispositions to cause other mental states. Turing-machine functionalism interprets the input-output relations to be specified by a machine table that maps symbolic inputs to symbolic outputs. Although numerous variations on functionalism have been proposed, most are based on the information-processing model of mind. An exception to this pattern is the teleological functionalism of Sober (1985) and Millikan (1984, 1993), who identify function of an object not with its informational input-output relations, but rather with its purpose or role within a system. As we have seen with Millikan, the purpose of an object is interpreted in terms of its evolutionary selected role.

DSB follows teleological functionalism by rejecting the computer model of mind and its underlying Turing-machine notion of function. But it also rejects Millikan's position that function should be understood in terms of evolutionary selected role. Instead, mental states are individuated by the environmental context (both the sensory input and its source), the behavior elicited within directly related neural subsystems, and the internal and external behavioral outputs or dispositions (if any).

Stipulating that it is the environmental context that helps determine mental states and not just sensory inputs grants ecological considerations a limited role. The reason environ-

mental context is important is that mental states do not form merely in response to signals coming in through sensory organs, but also through the active investigation of these signals by an organism traversing its environment. For example, Thelen and Smith (1994) document the fact that human infants learn the dangers of navigating slopes through actively navigating them, and not by generalizing from their fear of cliffs. An infant who learns that a slope presents danger when it crawls needs to relearn this fact for when it walks on slopes. That a slope is a dangerous is thing is not learned by the infant through visual inspection and generalization to past experiences or innate fears. It requires active investigation. This is, it will no doubt be argued, just another form of sensory input. It is, however, highly context dependent and requires the activity of the body within the environment.

Some variation of Descartes's evil genius is often trotted out to demonstrate that the environment is not relevant for determining what mental states are, since, it is supposed, it is possible in principle to simulate all the signals coming from the environment and pump them into an isolated mind/brain. Presumably, this mind/brain would form the same mental states as if it were in a real environment. That it is probably impossible in fact to achieve is often brought up (Dennett 1991) as a reason to doubt its relevance, but there is also reason to suspect that it begs the question of what mental states are against ecological and functionalist approaches. The evil genius would not only have to generate signals corresponding to what emanates from the environment, but also provide feedback for the isolated mind/brain's voluntary actions. This means the evil genius would have to know what the individual is thinking. This would either require a priori knowledge of which set of input-output relations and neural states map to which mental states for the particular individual, because neural architecture is highly variable and the effects of experience reinforce this variability,

or require that we assume that behaviorism is true and only input-output relations matter. If we take the former position, where would this a priori model come from, if not from an individual located in a real environment? This implies that the evil genius himself can only determine mental states in terms of their environmental context. Supposing that the evil genius is omniscient and just knows begs the question of whether omniscience is possible in principle. If we take the latter route, we have a thought-experiment that applies only to behaviorism.

The behavior elicited within the directly related neural subsystems refers to the activity of pattern formation in neuronal ensemble firing. The self-organization of neural firing around incoming signals through neural selection that leads to the formation of firing patterns is, in itself and without need for causal connections to other mental states, a defining feature of mental states. This is one way in which DSB augments functionalism.

Internal behavioral outputs consist of the internal causal relations between neural subsystems. This is the same characteristic of mental states that functionalists introduced to fix behaviorism, although the mechanisms are different than those proposed by Turing-machine functionalists, as the previous chapter makes clear. This difference in mechanism is of no little importance, as the next section highlights, for it frees DSB of the charge of liberalism in ascribing mental states.

These characteristics are not meant to be necessary and sufficient conditions for a physical process to qualify as a mental state. The category of mental state is not classical in nature, but rather consists of states bearing family resemblances. The only necessary condition is the formation of patterns of neuronal activity, although it is not a sufficient condition.

DSB could be considered a form of teleological functionalism but for three differences. The function of a mental state, its causal relations and environmental context, is understood not in terms of its evolutionary selected role, but its neurally selected role. The context of neural selection is both the brain and the environment over the life of the individual. Evolutionary selection occurs within populations of individuals and species and the environment over large time-scales. Second, DSB contains an additional component, the pattern formation of neural systems, that constitutes the phenomenal content of mental states. Finally, DSB explicitly eschews the notion that the function of neural structures is to pass pieces of information. Teleological functionalism is perfectly consistently with the information processing model of mind, because the evolutionary purpose of neurons could be to pass pieces of information just as the evolutionary purpose of the heart is to pump blood. Accepting teleological functionalism does not necessary imply accepting an alternative understanding of neural structures to that of the information processing model, whereas accepting the DSB model requires rejecting the information processing model.

7.1.3 Avoiding Liberalism

Because Turing-machine functionalism is a plausible expression of teleological functionalism, the latter does not necessarily avoid the defects of the former. Thus, contrary to Sober's (1985) claim, teleological functionalism is open to the charge of liberalism, which is the mistake of attributing mental states to systems that we would not normally recognize as possessing them. Ned Block's (1978) 'homunculi head' thought-experiment illustrates this failing of functionalism.

Block imagines a situation in which a vast number of people, the population of China, are given two-way radios that connect them with one another and with an artificial body in a way that mimics the connections in the brain and body of a particular person. Each person mimics the activity of a neuron in transmitting the appropriate signals to the person-neurons to whom he is connected. In this way, the population of China could mimic the functional organization and activity of a real brain, and therefore, according to Turing-machine functionalism, would have be considered as having mental states.

Teleological functionalism is supposed to avoid liberalism by defining function as the role of object or structure within the system it resides. Thus, the teleological function of the heart is to pump blood. So what is the function of a neuron? Teleological functionalism does not proscribe the neuron's role from being that of passing information to other neurons. If it is, then teleological functionalism falls prey to the homunculi head though experiment. And this is an empirical matter, not a theoretical one.

DSB can be seen as a subset of teleological functionalism that is disjoint from Turing-machine functionalism. It makes more specific claims about the functions of neurons in the brain, and these functions do not conform to a machine table description. It is not plausible that a group of humans could mimic the self-organization of the brain by following a set of input-output rules. To do so would require specifying a set of rules for each human to obey, rules that would have to conform not only to the nonlinear behavior of neurons, but also the interactions of neurons. It would also have to take account of the slaving principle. This means that somehow a global order would have to control the local interactions of the person-neurons without it having to backpropagate to each person-neuron. The time for backpropagation to occur is not the issue here, and therefore Block's argument that the

length of time for the processes to occur is irrelevant is itself irrelevant. Backpropagation of a global order simply does not occur, because the global order emerges contemporaneously with the local interactions. Furthermore, because nonlinear behavior cannot be decomposed by its very nature, it is implausible that people could be given a set of rules to mimic it.

What teleological functionalists were attempting to argue is that the behavior of neurons cannot be captured by input-output mappings of information. DSB fills in this argument. The burden of demonstrating that a homunculi head could reproduce the behavior of the brain lies squarely with those who believe it can, and there is little reason to believe they will succeed. Success would require abandoning a rule-based approach, and therefore, abandoning computationalism.

7.2 Dynamic Systems Behaviorism as Ecological Externalism

All forms of externalism reject the notion that mental states supervene on brain states. The varieties of externalism are distinguished by the type of non-neural component they deem necessary for mental states to be individuated. Evolutionary externalists point to the organism's evolutionary history as what distinguishes similar internal states. Ecological externalists point to aspects of the environment as that additional component. In Chapter 5, I argued why evolutionary externalism should be rejected. I also suggested that the general principles of ecological externalism were correct, although the specific details of the variant of ecological externalism examined there, Gibson's ecological approach, may be incor-

rect. Here, I will flush out how dynamic systems theory serve as a foundation for ecological externalism.

It was noted that the mapping between external informational structure and the environmental features that cause it may not be unique, and that some further specification of the relationship, perhaps akin to Tye's optimal conditions, needs to be applied to this correlation. The notion of optimal conditions, however, was also found to be too stringent. How then should we relate the internal and external in such a way that both captures the flexibility of representation found in human categorization as well as helps explain the interaction of the two?

I suggest that the notion of an attractor in state space is just this needed conceptual framework. Attractors are not tokens that require a mapping function. They are sets of states of the system, positions in state space that the system moves to for a certain set of parameters. Thus, there are ranges for the parameter sets that, when taken on by the system, result in its settling into an attractor. We have seen in van Gelder's example that the state of the Watt governor's flywheel does not map uniquely to a state of the steam engine, and *vice versa*. But what is the "system" when we consider cognitive beings? The system here consists of both the internal and external factors, the organism and its environment. We capture the relations between them through dynamic systems analysis. The environment adds parameters to our description of representational states.

How the environment and how the organism may vary, as well as *how they may covary*, such that a coherent pattern is produced is again captured by the dynamic equations of this system. A range of external parameters—configurations of the environment and the positioning and motion of the organism—and a range of internal parameters contribute to the

same attractor state. How they are interrelated, and thus, how they interact, is therefore described by the dynamic equations governing them.

The notion of optimal conditions and the unique mapping of informational structure to environment is replaced by dynamic equations governing the interrelation of external and internal parameters. This does not imply that Gibson could not be right that visual perception uses an optic array as he describes, only that the information contained therein is not determined by a uniqueness mapping.

The "fuzziness" of human categorization is a product both of the variability of environment and variability of human neural and physiological makeup, including the variability within an individual over time. A specification of representational mapping based on if-and-only-if conditions does not capture this variability. Specifying Normal conditions is impossible in the face of stochastic relationships between environment and neural behavior. Thus, we must adopt a framework that can account for these stochastic relationships, not merely as relations of probability distributions (a distribution corresponding to environmental tokenings and a distribution corresponding to internal tokenings), but also as *relationships* over time. The fuzziness of human categorization has, as we saw in Chapter 6, a temporal dimensions—representations are patterns of the system over time and change over time. Representational mappings are not necessarily stable over time. How do we account for this? Dynamic systems theory has the tools for expressing such temporal relations.

Because of its temporal aspect, dynamic systems theory is particularly suited for understanding how an organism's actions over time contribute to its representation. Gibson contended that an essential aspect of visual perception was the organism's motion in its

environment, allowing it to sample the scene. The position of the organism over time, the ambient light it receives over these positions, and the temporal behavior of its neural processes all become parameters of the representational system. This ability to integrate over time, as well as its ability to capture continuous, stochastic relationships, are what commend dynamic systems theory over symbolic approaches.

7.2.1 Swampman Returns

If representation consists of an environment-organism system governed by dynamic equations, it would seem to imply that mental states supervene on system states. There is one problem with this implication. If we interpret this to mean that if the system's parameters change, there must be a corresponding change in the mental state, then the description of supervenience must be rejected. This is because parameters may change and yet the system remain in a stable state: change in parameters does not imply change in mental state. We can't even say that mental states supervene on parameter sets, because this would view them as fixed in their relation to mental states. A counterexample of such a fixed relationship can be seen in the phenomenon of hysteresis. As a system evolves from an initial parameter configuration and moves from one stable state to another, this evolution is not necessarily reversible. The boundary between stable states changes as the system moves in the reverse direction (passing through the same parameters over time, except in reverse), and so we can't map a particular set of parameters to a particular state of the system. Initial condition and evolution over time is essential. Thus, mental states do not supervene on parameter configurations. Rather, they supervene on attractor and repellor states.

What does this mean for Swampman? Swampman is once again a microphysical duplicate of Davidson—the hardware implementing his internal dynamic processes are the same—placed in the same environment as Davidson. However, we must add one other specification for Swampman, namely that his internal dynamics have the same initial conditions as Davidson's. In this case, Swampman upholds our strong intuition that he thinks the same thoughts as Davidson. Given the same implementing mechanisms and same starting conditions, the two representational systems that are Swampman-in-his-environment and Davidson-in-his-environment will reach the same attractor states. If we violate the initial condition requirement, all bets are off, although, depending on how strong the particular attractor corresponding to a mental state is, they may still have the same mental state. It just isn't guaranteed.

Does this latter fact violate what we know or intuit? Given different initial conditions, Swampman and Davidson are actually different temporal slices of the same individual. It is important to note: *initial conditions cannot be specified merely by means of atemporal values of parameters.* Swampman and Davidson are inherently temporal beings. For one, hysteresis reminds us that the directionality of the evolution of their representational systems is important. *Directionality* is not a magical concept, but rather is simply the notion of where you are coming from in parameter space. So same initial conditions includes the directionality of change at that point in parameter space. The essential temporality of Swampman and Davidson is not surprising given what we know about human categorization. If Swampman is the Davidson of tomorrow, he might apply a category differently than the Davidson of today would. As was noted in Chapter 6, many categories are unstable over time.

If Swampman and Davidson are the same temporal slices of Davidson (this does not mean that Swampman was copied from Davidson, just that his initial conditions are the same), they will form the same representations. Notice that their evolutionary history is irrelevant. Dynamic Systems Behaviorism, however, relies on the notion of neuronal selection. Doesn't Swampman imply that neuronal selection is not necessary?

The answer is Yes and No. What Dynamic Systems Behaviorism implies is that a mechanism for self-organization is necessary, and that, in practice, this mechanism is neuronal selection. Because computationalism cannot provide a similar mechanism, it is not a plausible theory of mind. However, if by some miracle a lightning strike was both able to make a microphysical of Davidson (the structural component), and "wind him up" correctly (the temporal component), so to speak, then Swampman would have the same representations as Davidson (assuming he is in the same environment). The past of Swampman is, in this case, irrelevant, although his current neural ability to organize and learn is dependent on a mechanism like neuronal selection (or continued lightning strikes). So, no, neuronal selection is not necessary, assuming lightning can have these powers. But, yes, a mechanism like neuronal selection is necessary in the absence of magical lightning, and it just so happens that neuronal selection is the mechanism that naturally occurs. It does not threaten Dynamic Systems Behaviorism to imagine that lightning can have the power to form a system as if it were formed by neuronal selection. However, lightning strikes cannot give Swampman an evolutionary past, even if they can make him in the image of something that does have one. If an evolutionary history is necessary, then Swampman is doomed. Rather, evolutionary externalism is doomed, because it requires that Swampman have an evolutionary history.

Correctly created, Swampman has the same representations as his double. But does he experience what his double experiences? What does Dynamic Systems Behaviorism imply about the phenomenology of experience?

7.3 A Scientific Foundation for Phenomenology

As Edelman points out, neuroscientists have tended to avoid formulating theories of consciousness, focusing instead on lower-level brain functioning and very specific aspects of the higher-level. This has been a quite reasonable, divide-and-conquer approach, one that has yielded astounding results. It is an approach constrained in part by the analytic tools available. Dynamic systems theory offers the neuroscientist a tool more suited to explaining global aspects of functioning. Efforts to apply dynamic systems theory are still in their infancy, but are likely to yield similarly important results as their predecessors.

Even granting this possibility, there remains a nagging question about the applicability of objective techniques to studying consciousness. Can the objective tell us what it is like to have an experience? Can it explain the phenomenal qualities associated with experiences? What leads to this question is a host of strange aspects of phenomenal consciousness. Unlike physical objects, mental objects are inherently owned; there are no unowned pains floating about, they are all someone's pains. Furthermore, all experience is experience from a perspective. Without having had this perspective, i.e., without having experienced something for one's self, it seems that one cannot have knowledge of what it is like to have that experience (the point of Jackson's example of Mary). Following Tye (1995), let's call these problems the Problem of Ownership and the Problem of Perspectival Subjectivity, respectively.

It would seem that the mental has some odd properties not attributable to physical objects. If this is true, how can we explain how the physical causes the mental, since we are all (well, almost all) physicalists of some sort? Conversely, how is it that the mental can cause physical behavior? Again following Tye, we will call these problems the Problem of Mechanism and the Problem of Phenomenal Causation, respectively. The first two problems set the stage: there is something about the mental that makes it not like the physical. The second two draw uncomfortable consequences from this disparity: we can't explain the causal interaction of mental and physical. Although Tye identifies ten problems relating to phenomenal consciousness, I will focus on these four, as they are the most fundamental.

So it would seem that the study of phenomenal qualities, Phenomenology, is divorced from the study of objective cognitive properties, the Cognitive Sciences. This needn't be so, as van Gelder (1999) has attempted to demonstrate. Aspects of Tye's Representationalism (and Dretske's, insofar as it is similar), can help bridge the gap between the phenomenal and physical. I will now examine those features of Representationalism that enable us to understand the four problems, pointing out how a dynamic systems understanding of representation takes the place of Tye's notion of representation. Van Gelder's illustration of how dynamic systems theory and phenomenology is one possible step in this general direction.

7.3.1 Dynamic Systems Behaviorism as Representationalism

7.3.1.1 What Is the Phenomenal Character of Experience?

Phenomenal Representationalists like Tye and Dretske hold that experience itself is representational. What does this mean? It is sometimes assumed that there are phenomenal

objects, such as afterimages, that the mind views through introspection. Supposedly, one can direct one's inner sense toward these objects and discover their qualitative properties. The properties of these objects are the qualitative properties of experience. Phenomenal Representationalists reject this view. In its place, Tye proposes that phenomenal objects be identified with phenomenal events, so, for example, my pain is just the event of me undergoing pain experiences. There is nothing over and above this event, no extra qualitative properties of an internally presented object to discover.

In the Representationalist view, sensory modules produce representations of external and internal objects. The phenomenal character of experience is just the intentional content of these representations. According to Tye, "[p]henomenal content . . . is not a feature of any of the representations occurring *within* the sensory modules" (Tye 1995, 136). So what is phenomenal content? Tye argues that it is Poised Abstract Nonconceptual Intentional Content. For contents to be *poised* is

> to be understood as requiring that these contents attach to the (fundamentally) maplike output representations of the relevant sensory modules and stand ready and in position to make a direct impact on the belief/desire system. (Tye 1995, 138)

Contents are *abstract* if the only objects that enter into these contents are the subjects of experience. They are nonconceptual in the sense that the features of the contents are not necessarily the features that figure into the subject's corresponding concepts. Something is intentional if it can be about something else, without that something else existing.

For the contents to be poised, the outputs of the sensory modules must have a maplike relationship to what they are triggered by under optimal conditions. But I have elsewhere denied that Tye's notion of 'causal covariance under optimal conditions' is adequate for establishing a mapping. How do we understand the notion of 'poised' in terms of Dynamic

Systems Behaviorism? To be poised is to have an attitude toward something. In the case of sensory perception, it is to form a coherent pattern of activity in the sensory modules with regard to the environmental conditions. The system parameters that produce this attractor state are internal and external, and the dynamic equations governing the system determines their interrelations (a more complex mapping than that envisioned by causal covariance under optimal conditions).

This would suggest that contents are not abstract. External factors must be present in order for the sensory modules to move into the attractor state. It would seem that we are consequently faced with the problem of explaining phantom limb pains. Not so. Remember, the interrelation of parameters that produces a particular state is extremely complex. A dynamic system does not necessarily cease to function if a subset of its possible parameters don't play a part. It is still possible to move into a particular state with the contribution of only a subset of its components. So, phantom limb pains are possible. The pain sensory modules in the brain form coherent patterns of behavior in the absence of nervous activity.

This would seem to imply that patterns of behavior in the sensory modules are "about" what they were selected for, i.e., that they behave as they would under Normal conditions. The dynamic systems approach does eschew evolutionary considerations, but does not thereby throw out temporal explanations in general. Sensory modules organize around input from the external world, and this organization cannot be explained without this. The history of the sensory module is important because, for example, it determines the directionality of the system's behavior.

Does the brain-in-a-vat thought-experiment undermine this? The thought-experiment is that we place a brain-in-a-vat and feed it sensory inputs just like the ones that are presented

to a real brain sensing the world. We must, of course, assume that we can read its mind to provide the sensory input of its desired actions. Its sensory organs organize around these inputs. What are they representing? Tye's answer to this thought-experiment is to cite the evolutionary history of the brain. But evolutionary history, however, does not determine fully how the brain is wired up, or even whether it will be wired for a particular function. Hubel and Wiesel's (1979) experiments with cats make that clear.[1] Real-time adaptation also determines how the brain-in-the-vat is wired up. In any case, given that Tye acknowledges that Swampman has experiences, he cannot even refer to evolutionary history to establish the content of the brain-in-the-vat's experiences. We could just pry Swampman's brain out his skull and put it in the vat, and his argument would be moot. Swampman's brain in the vat would presumably have the same experiences as his brain out of the vat and in his body.

How does a Dynamic Systems Behaviorist solve this problem? By realizing that is only a problem for Representationalists who insist on a causal covariance model of representation. Just as sensory modules can form patterns of behavior in rare instances without external inputs, they can form these attitudes with a variety of inputs. This does not make representation arbitrary; there are distinct subsets of external inputs that can produce a particular attitude, even if there is not a one-to-one mapping between external input and attitude. Note that Swampman's brain in his body and his brain in a vat are two distinct dynamic systems. We are not asking if a duplicate of Swampman can have the same experiences, but rather whether something neurally like Swampman can have the same experi-

1. Hubel and Wiesel deprived kittens of sensory input to their eyes and demonstrated that environmental input is necessary for full development and maintenance of these sensory organs.

ences. This does not mean that experiences supervene on neural states, because in each case, the attractor associated with the attitude is a result of the entire system.

7.3.1.2 How Representationalism Helps Solve the Problems of Consciousness

Attitudes, as envisioned in Dynamic Systems Behaviorism, are patterns of behavior, not physical objects themselves. They are states of a system. As such, they belong to a system, they are particular to the system that gives rise to them. In this sense, they are owned. Having a particular experience therefore means being that particular system. Thus, the Problem of Ownership (and the privacy of experience) does not affect Dynamic Systems Behaviorism.

This directly mirrors the arguments made by Tye as to why the Representationalist notion of experience is not affected by the Problem of Ownership. Representations are events, and events belong to systems having them. The difference between Tye's representationalism and Dynamic Systems Behaviorism in this context lies in how the events are conceived.

Tye's solution to the Problem of Perspectival Subjectivity, on the other hand, does require accepting Tye's argument that awareness of phenomenal quality is achieved through application of phenomenal concepts. There are two kinds of phenomenal concepts: predicative and indexical. Predicative phenomenal concepts correspond to the features we conceive our experiences to be of (particular colors, shapes, etc.). Indexical concepts enable us to single out particular features represented in our experiences contemporaneously with the experience. Together, they enable us to conceive of an experience as an experience of *this* shade of blue. Understanding phenomenal content requires applying these concepts to

experience, which means taking a particular perspective toward experience. Perspectival subjectivity is nothing more than this. In the language of Dynamic Systems Behaviorism, the perspectival subjectivity of experience is the fact that understanding a phenomenal state is a system's taking the appropriate attitude toward it.

Once the spookiness of consciousness is gone, the mechanisms for causal interaction become evident. For the dynamic systems behaviorist, if phenomenal character is representational, then phenomenal experience is simply the activity of the sensory attitudinal *systems*. These systems are themselves causally efficacious through their contribution to inner attitudinal systems directed towards them (the correlate to Tye's notion of applying phenomenal concepts).

This, of course, is just a rough sketch of Tye's sketch of a solution to the problems of consciousness. However, if Tye is right that Representationalism can explain the problems of consciousness, then it would seem that Dynamic Systems Behaviorism can as well, and without the use of vague notions of information.

7.4 Symbols as Social Constructs

Humans display a wide range of behavior that can be classified as symbol manipulation. From language use, to mathematics, to painting and sculpture, the use of symbols for the expression of thoughts distinguishes humans from other species. Computationalists explain human symbol manipulation in terms of symbol manipulation in the mind/brain, casting the external use of symbols back into the psychological structures that lead to this external behavior. This move is tempting for its simplicity. The existence of external rule-following

behavior is explained in terms of rules in the mind brain; symbols in the world are place-markers for symbols in the mind.

DSB rejects this projection of external symbolic behavior back into the mind. Instead, symbols are conceived as objects only for a community. They exist only within a community capable of interpreting them. A speech act acts as a symbol for a thought only insofar as others are capable of understanding it as such. Speech acts and other symbols do not capture the nature of a thought—they are not external expressions of the symbols that appear in the mind. Symbols act as facilitators for enabling other members of a community to approximate the thought processes of the symbol producer. Symbols are necessary simplifiers for the act of conveying what one is thinking, and because they are simplifiers, they do not fully capture the nature of thought.

7.5 Conclusion: Representation and Information

The information-processing model of mind attempts to decompose thoughts into primitive elements according to their syntactic structure and assign a semantic interpretation to these syntactic primitives. This approach fails to explain the representational capabilities of real organisms and to provide a foundation for designing adaptive artificial cognitive agents. An alternative understanding of representational capacities in real organisms has emerged with dynamic systems theoretical accounts of organisms' brain/body interactions with their environments.

Kelso (1997) has outlined one possible interpretation of the semantic features of dynamic systems. He views the order parameters that emerge through the interaction between brain/body and environment as inherently meaningful because they capture the

relations between the two domains. Order parameters describe the manner in which patterns form in brain/body behavior and how these patterns correlate with the environment.

Kelso argues that these order parameters are *meaningful to* the organism, that they are meaningful bits of information about its coordinative relations to its environment. What this suggests is that the organism has access to the order parameters, that it analyzes its own dynamic systems and distills these order parameters from its understanding of how it is itself behaving.

Unlike a symbol, which has no inherent connection to what it represents and therefore requires 'grounding', an order parameter is intrinsically meaningful to the observer analyzing it because it captures the relations between domains. But like a symbol, it can be meaningful only to an observer who identifies its role within a system. By suggesting that the order parameter itself is meaningful to the system, Kelso duplicates computationalism's error of identifying a representation of a system with the system's representations. The meaningfulness of order parameters indicates not that these parameters are meaningful to the system, but that the processes which they describe are themselves meaningful to the system. Order parameters capture the relations between domains—environment, body, and brain—because the systems they describe are composed of these domains.

DSB is therefore neither a radically environmentalist nor a methodologically solipsistic theory of mind. Mind is instead a unity of environment, body, and brain in that the self-organization of the neural elements responsible for cognition only occur within a total system including the body and its environment. Mental states are behavioral states of a brain/body within a particular environment, and it is necessary to explain the contribution of the environment in order to explain the nature of these states. For what are considered

‘abstract’ mental states, the environment consists largely of the behavior of neural sub-systems.

This approach captures what Andy Clark (1997) is attempting with his augmentation of symbol systems: it brings the environment into the system. But Clark also wants to be able to informationally decompose the system, and so requires that all representations beyond direct perceptual representations be symbolic in nature. It avoids, however, the necessary (for Clark), and unexplained, link between the symbolic and dynamic systems.

Representations are not bits of information gathered from the environment, stored in the brain, and reproduced in thought. They are behavioral attitudes taken through interaction of the environment, neural maps, and internal value systems, and reproduced through environmental triggers and internal controls. Attitudes are ‘stored’ by modification of neural chemistry to make the emergence of similar behavioral patterns likely in the appropriate contexts. Behavioral patterns are not ‘codes’ for interpretative neural systems to decode and act upon. In some contexts, they are control parameters similar to the electrical charge applied to active elements in the gas laser: they drive the process of self-organization in other neural ensembles external to their particular module. In other cases, they become part of a larger system and become enslaved to its higher-order behavior. The symbol and its interpreter is removed from the machine, and with them, the homunculus.

Any system can be analyzed in terms of Shannon information flow. This fact should not lead to the conclusion that semantic information is coextensive with Shannon information. It also does not imply that every system is conscious or has experiences. Although neural dynamic systems can be analyzed in terms of information flow, their role is not of information processors.

The computationalist information-processing model of mind seeks to explain the operations of the mind by casting the external symbolic behavior of language use back into the mind. Because of this move, it offers a more compelling explanation of the bases of language behavior than connectionist or dynamic systems theories. And because of this move, it cannot explain how non-linguistic representations arise. Dynamic systems theorists and neural selectionists build from perceptual processes upward, and so are able to offer an explanation of the origin of non-linguistic representations, but have yet to fill out the details on how overtly symbolic behavior is produced. Linguistic behavior stands as the great challenge to any dynamic systems account of cognition.

References

Anderson, J. 1991. The adaptive nature of human categorization. *Psychological Review* 98: 409-429.

Armstrong, S., L. Gleitman, and H. Gleitman. 1983. What some concepts might not be. *Cognition* 13:263-308.

Barinaga, M. 1990. Neuroscience models the brain. *Science* 247: 524-526.

Barsalou, L. 1987. The instability of graded structure: implications for the nature of concepts. In *Concepts and conceptual development: ecological and intellectual factors in categorization*, ed. U. Neisser. Cambridge: Cambridge University Press.

Berg, G. 1992. A connectionist parser with recursive sentence structure and lexical disambiguation. In *AAAI-92: Proceedings of the Tenth National Conference on Artificial Intelligence*. Cambridge, MA: The MIT Press.

Block, N. 1978. Troubles with functionalism. In *Perception and cognition: issues in the foundations of psychology*, ed. C. Savage. Minneapolis: University of Minnesota Press.

Brooks, R. 1990. Elephants don't play chess. *Robotics and Autonomous Systems* 6:3-15.

Brown, T. 1989. *Genetics: a molecular approach.* New York: Chapman and Hall.

Burge, T. 1979. Individualism and the mental. In *Midwest studies in philosophy*, 4, eds. P. French, T. Uehling, and H. Wettstein. Minneapolis, MN: University of Minnesota Press.

———. 1982. Other bodies. In *Thought and object*, ed. A. Woodfield. Oxford: Oxford University Press.

———. 1986. Individualism and psychology. *Philosophical Review* 95:3-45.

Callan, R., and D. Palmer-Brown. 1997. An analytical technique for fast and reliable derivation of connectionist symbol structure representations. *Connection Science* 9(2):139-159.

Carlson, N. 1994. *Physiology of behavior*. Needham Heights: Allyn and Bacon.

Chalmers, D. 1992. Subsymbolic computation and the chinese room. In *The symbolic and connectionist paradigms: closing the gap*, ed. J. Dinsmore. New York: Lawrence Erlbaum.

———. 1996. *The conscious mind: in search of a fundamental theory*. Oxford: Oxford University Press.

Cherniak, C. 1991. The bounded brain: toward quantitative neuroanatomy. *Journal of Cognitive Neuroscience* 2(1):58-68.

Churchland, P. M. 1989. *A neurocomputational perspective: the nature of mind and the structure of science*. Cambridge, MA: The MIT Press.

———. 1995. *The engine of reason, the seat of the soul: a philosophical journey into the brain*. Cambridge, MA: The MIT Press.

Churchland, P. S., and V. Ramanchandran. 1993. Filling-in: why Dennett is wrong. In *Dennett and his critics*, ed. Bo Dahlbom. Oxford: Blackwell Publishers Ltd.

Churchland, P. S., and T. Sejnowski. 1989. Neural representation and neural computation. In *Neural connections, mental computations*, ed. L. Nadel. Cambridge, MA: MIT Press.

———. 1992. *The computational brain*. Cambridge, MA: The MIT Press.

Clark, A. 1997. *Being there: putting brain, body, and world together again*. Cambridge, MA: MIT Press.

———. 1998. Time and mind. *Journal of Philosophy* 95(8):354-376.

Cormen, T., C. Leiserson, and R. Rivest. 1992. *Introduction to algorithms*. Cambridge, MA: The MIT Press.

Cover, T. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers* EC-14: 326-334.

Crane, T. and D. Mellor. 1990. There is no question of physicalism. *Mind* 99:185-206.

Crockett, L. 1994. *The turing test and the frame problem: AI's mistaken understanding of intelligence*. Norwood, NJ: Ablex.

Cummins, R. 1983. *The nature of psychological explanation*. Cambridge, MA: The MIT Press.

Damasio, A. 1989. Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. In *Neurobiology of cognition*, eds. P. Eimas and A. Galaburda. Cambridge, MA: The MIT Press.

Daugman, J. 1986. Communication theory and intentionality. *Behavioral and Brain Sciences* 9:140-141.

Davidson, D. 1986. Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association* 60:441-458.

Dawkins, R. 1976. *The selfish gene*. Oxford: Oxford University Press.

Dennett, D. 1981. True believers: the intentional strategy and why it works. In *Scientific explanation: papers based on Herbert Spencer Lectures given in the University of Oxford,* ed. A.F. Heath. London: Oxford University Press.

———. 1987. *The intentional stance.* Cambridge, MA: The MIT Press.

———. 1991. *Consciousness explained*. Boston: Little, Brown.

———. 1995. *Darwin's dangerous idea: evolution and the meanings of life*. New York: Touchstone.

———. 1997. An exchange with Daniel Dennett. In Searle, J. *The mystery of consciousness*. New York: The New York Review of Books. First published in The New York Review of Books.

———. 1998. *Brainchildren: essays on designing minds*. Cambridge, MA: The MIT Press.

Dretske, F. 1981. *Knowledge and the flow of information*. Cambridge, MA: The MIT Press.

———. 1990. Putting information to work. In *Information, language, and cognition*, ed. P. Hanson. Vancouver: University of British Columbia Press.

———. 1995. *Naturalizing the mind*. Cambridge, MA: The MIT Press.

Dreyfus, H. 1972. *What computers can't do: the limits of Artificial Intelligence*. New York: Harper and Row.

———. 1992. *What computers still can't do*. Cambridge, MA: The MIT Press.

———. 1996. Response to my critics. *Artificial Intelligence* 80:171-191.

Dreyfus, H., and S. Dreyfus. 1988. Making a mind versus modelling the brain: Artificial Intelligence back at a branch-point. In *The philosophy of Artificial Intelligence*, ed. M. Boden. New York: Oxford University Press.

Dyer, M. 1995. Connectionist natural language processing: a status report. In *Computational architectures integrating neural and symbolic processes: a perspective on the state of the art*, ed. R. Sun and L. Bookman. New York: Kluwer Academic Publishers.

Edelman, G. 1987. *Neural darwinism: the theory of neuronal group selection*. New York: Basic Books.

———. 1989. *The remembered present: a biological theory of consciousness.* New York: Basic Books.

———. 1992. *Bright air, brilliant fire: on the matter of mind*. New York: Basic Books.

Edelman, G., and L. Finkel. 1984. Neuronal group selection in the cerebral cortex. In *Dynamic aspects of neocortical function*, eds. G. Edelman, W. Gall, and W. Cowan. New York: Wiley.

Eigen, M. 1992. *Steps towards life: a perspective on evolution*. Oxford: Oxford University Press.

Elman, J. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:195-225.

Fodor, J. 1975. *The language of thought*. Hassocks, Sussex: Harvester.

———. 1980. Searle on what only brains can do. *The Behavioral and Brain Sciences* 3: 431-432.

———. 1983. *The modularity of mind: an essay on faculty psychology*. Cambridge, MA: The MIT Press.

———. 1986. Banish DisContent. In *Language, mind and logic*, ed. J. Butterfield. Cambridge: Cambridge University Press.

———. 1993. Fodor's guide to mental representation: the intelligent auntie's vade mecum. In *Readings in the philosophy and cognitive science*, ed. A. Goldman. Cambridge, MA: The MIT Press. Originally published in MIND (1985) 94:55-97.

———. 1994. *The elm and the expert: mentalese and its semantics*. Cambridge, MA: The MIT Press.

Fodor, J., and Z. Pylyshyn. 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition* 28:3-71.

Gammaitoni, L., P. Hanggi, P. Jung, and F. Marchesoni. 1998. Stochastic Resonance. *Review of Modern Physics* 70:223-288.

Gazzaniga, M., ed. 1995. *The cognitive neurosciences*. Cambridge, MA: The MIT Press.

———., ed. 2000. *The new cognitive neurosciences*. Cambridge, MA: The MIT Press.

Gibson, J. 1979. *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gleason, J. 1993. *The development of language*. New York: MacMillan Publishing Company.

Globus, G. 1992. Toward a noncomputational cognitive neuroscience. *Journal of Cognitive Neuroscience* 4:299-310.

Goonatilake, S. 1991. *The evolution of information: lineages in gene, culture and artefact.* London: Pinter Publishers.

Hanks, S., and D. McDermott. 1986. Default reasoning, nonmonotonic logics, and the frame problem. *Proceedings AAAI* 86:379-412.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42:335-346.

Hayes, P. 1977. In defence of logic. *Proceedings IJCAI* 77:559-565.

———. 1987. What the frame problem is and isn't. In *The robot's dilemma*, ed. Z. Pylyshyn. Norwood, NJ: ABLEX Publishing Corporation.

Haykin, S. 1994. *Neural networks: a comprehensive foundation*. New York: Macmillan Publishing Company.

Hecht-Nielsen, R. 1989. Theory of the backpropagation neural network. *Proceedings of the International Joint Conference on Neural Networks* 1:593-611.

Hobson, J., and R. Stickgold. 1995. The conscious state paradigm: a neurocognitive approach to waking, sleeping, and dreaming. In *The cognitive neurosciences*, ed. M. Gazzaniga. Cambridge, MA: The MIT Press.

Holzmüller, W. 1984. *Information in biological systems: the role of macromolecules*. Cambridge: Cambridge University Press.

Horgan, T., and J. Tienson. 1996. *Connectionism and the philosophy of psychology*. Cambridge, MA: The MIT Press.

Horgan, T., and J. Woodward. 1985. Folk psychology is here to stay. *The Philosophical Review* 94:197-225 .

Horst, S. 1996. *Symbols, computation, and intentionality*. Berkeley: University of California Press.

Hubel, T., and D. Wiesel. 1979. Brain mechanisms of vision. *Scientific American* 248(1):54-64.

Ifeachor, E., and B. Jervis. 1993. *Digital signal processing: a practical approach*. New York: Addison-Wesley Publishing Company.

Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy* 83(5):291-295.

Janlert, L-E. 1987. Modeling change—the frame problem. In *The robot's dilemma*, ed. Z. Pylyshyn. Norwood, NJ: ABLEX Publishing Corporation.

Jochem, T., D. Pomerleau, and C. Thorpe. 1995. Vision guided lane transition. *IEEE Symposium on Intelligent Vehicles*, Detroit (1995).

Kaas, J. 1995. The reorganization of sensory and motor maps in adult mammals. In *The cognitive neurosciences*, ed. M. Gazzaniga. Cambridge, MA: The MIT Press.

———. 2000. The reorganization of sensory and motor maps after injury in adult mammals. In *The New Cognitive Neurosciences*, ed. M. Gazzaniga. Cambridge, MA: The MIT Press.

Kelso, J. 1997. *Dynamic patterns: the self-organization of brain and behavior*. Cambridge, MA: The MIT Press.

Kohonen, T. 1995. *Self-organizing maps*. Berlin: Springer Verlag.

Kosslyn, S. 1983. *Ghosts in the mind's machine*. New York: Norton.

Küppers, B-O. 1990. *Information and the origin of life*. Cambridge, MA: MIT Press.

Lakoff, G. 1987. Cognitive models and prototype theory. In *Concepts and conceptual development: ecological and intellectual factors in categorization*, ed. U. Neisser. Cambridge: Cambridge University Press.

Cormen, T., C. Leiserson, and R. Rivest. 1990. *Introduction to Algorithms*. Cambridge, MA: The MIT Press.

Lenat, D., and R., Guha. 1990. *Building large knowledge based systems: representation and inference in the cyc project*. Reading, MA: Addison-Wesley.

Lespérance, Y., H. Levesque, R. Lin, D. Marcu, R. Reiter, and R. Scherl. 1994. A logical approach to high-level programming—a progress report. In *Control of the physical world by intelligent systems, papers from the 1994 AAAI Fall Symposium*, ed. B. Kuipers, New Orleans (1994).

Levesque, H., and R. Reiter. 1998. High-level robotic control: beyond planning. *AAAI 1998 Spring Symposium: Integrating Robotics Research: Taking The Next Big Leap* Stanford University (1998).

Levesque, H., R. Reiter, Y. Lespérance, F. Lin, and R. Scherl. 1994. GOLOG: a logic programming language for dynamic domains. *The Journal of Logic Programming* 31:59-84.

Lieberman, P. 1984. *The biology and evolution of language*. Cambridge, MA: Harvard University Press.

Lin, F., and R. Reiter. 1997. How to progress a database. *Artificial Intelligence* 92:131-167

Lorenz, E. 1993. Does the flap of a butterfly's wings in Brazil set off a tornado in Texas? In *The essence of chaos*. Seattle: University of Washington Press. Address firtst presented at American Association for the Advancement of Science, 1979.

Madsen, P., S. Holm, S. Vorstrup, L. Friberg, N. Lassen, and G. Wildschiodtz. 1991. Human regional cerebral blood flow during rapid-eye-movement sleep. *Journal of Cerebral Blood Flow and Metabolism* 11, 502-507.

Margolis, H. 1987. *Patterns, thinking, and cognition: a theory of judgement.* Chicago: University of Chicago Press.

Marr, D. 1982. *Vision: a computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman and Company.

McCarthy, J. 1979. Ascribing mental qualities to machines. In *Philosophical Perspectives in Artificial Intelligence*, ed. M. Ringle. New York: Humanities Press.

———. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence* 13:27-39.

———. 1986. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence* 26:89-116.

———. 1996. Book review of Hubert Dreyfus, What computers still can't do. *Artificial Intelligence* 80:143-150.

McCarthy, J., and P. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine intelligence 4*, eds. B. Meltzer and D. Michie. Edinburgh, Scotland: Edinburgh University Press.

Miikkulainen, R. 1993. *Subsymbolic natural language processing: an integrated model of scripts, lexicon, and memory*. Cambridge, MA: The MIT Press.

———. 1997. Natural language processing with subsymbolic neural networks. In *Neural network perspectives on cognition and adaptive robotics*, ed. A. Browne. Philadelphia: Institute of Physics Press.

———. 2000. Text and discourse understanding: the DISCERN system. To appear In *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, eds. R. Dale, H. Moisl and H. Somers. New York: Marcel Dekker.

Millikan, R. 1984. *Language, thought and other biological categories*. Cambridge, MA: The MIT Press.

———. 1993. On mentalese orthography. In *Dennett and his critics*, ed. B. Dahlbom. Cambridge, MA: Blackwell Publishers Inc..

Minsky, M., and S. Papert. 1969. *Perceptrons*. Cambridge, MA: The MIT Press.

Monod, J. 1972. *Chance and necessity*. New York: Vintage Books.

Newell, A. 1982. The knowledge level. *Artificial Intelligence* 18: 81-132.

———. 1990. *Unified theories of cognition.* Cambridge, MA.: Harvard University Press.

Nisbett, R., and L. Ross. 1980. *Human inference: strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Nosofsky, R. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115: 39-57.

O'Brien, G., and J. Opie. 1999. A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* 22(1):127-147.

Ollis, M. 1997. Perception algorithms for a harvesting robot. Ph.D. Dissertation, Carnegie Mellon University. CMU-RI-TR-97-43.

Penrose, R. 1989. *The emperor's new mind: concerning computers, minds, and the laws of physics*. Oxford: Oxford University Press.

Pierce, J. 1980. *An introduction to information theory: signals, symbols and noise*. New York: Dover.

Pilarski, T., M. Happold, M. Ollis, H. Pangels, K. Fitzpatrick, and A. Stentz. 1998. The Demeter System for Automated Harvesting. *Proceedings of the 8th International Topical Meeting for Robotic and Remote Systems*, American Nuclear Society, April 25-30, 1999.

Pinker, S. 1989. *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: The MIT Press.

Pinker, S., and A. Prince. 1988. On language and connectionism. *Cognition* 28: 73-193.

———. 1999. The nature of human concepts. In *Evolution, structure and representation,* ed. P. van Loocke. London: Routledge.

Pollack, J. 1990. Recursive distributed representations. *Artificial Intelligence* 46:77-105.

Pomerleau, D. 1992. *Neural network perception for mobile robot guidance*. Boston: Kluwer.

Pomerleau, D. 1995. RALPH: Rapidly Adapting Lateral Position Handler. *IEEE Symposium on Intelligent Vehicles*, Detroit (1995).

Pribram, K. 1991. *Brain and perception: holonomy and structure in figural processing* Hillsdale, NJ: Lawrence Erlbaum and Associates, Inc.

Putnam, H. 1963. Brains and behavior. Reprinted in *Mind, language, and reality: philosophical papers*, vol. 2. London: Cambridge University Press, 1975.

———. 1975. The meaning of 'meaning'. In *Language, mind and knowledge*, Minnesota Studies in the Philosophy of Science, 7, ed. K. Gunderson. Minneapolis, MN: University of Minnesota Press.

———. 1988. *Representation and reality*. Cambridge, MA: The MIT Press.

Rapaport, W. 1988. Syntactic semantics: foundations of computational natural-language understanding. In *Aspects of Artificial Intelligence*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

———. 1995. Understanding understanding: syntactic semantics and computational cognition. In *AI, connectionism, and philosophical psychology, philosophical perspectives* Vol. 9, ed. J. Tomberlin. Atascadero, CA: Ridgeview.

———. 1996. Understanding understanding: semantics, computation, and cognition. preprinted as *Technical Report 96-26*. Buffalo, NY: SUNY Buffalo Department of Computer Science.

———. 1998. How minds can be computational systems. *Journal of Experimental and Theoretical Artificial Intelligence* 10:403-419.

———. forthcoming, 2001. How to pass a Turing Test: syntactic semantics, natural-language understanding, and first-person cognition. *Special Issue on Alan Turing and Artificial Intelligence, Journal of Logic, Language, and Information, Journal of Logic, Language, and Information.*

Ratcliff, R. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review* 97: 285-308.

Reiter, R. 1991. The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression. In *Artificial Intelligence and mathematical theory of computation: papers in honor of John McCarthy*, ed. Vladimir Lifschitz. San Diego, CA: Academic Press.

Redish, A., A. Elga, and D. Touretzky. 1996. A coupled attractor model of the rodent head direction system. *NETWORK* 7(4): 671-685.

Regier, T. 1992. The acquisition of lexical semantics for spatial terms: a connectionist model of perceptual categorization. Ph.D. Dissertation. University of California at Berkeley.

Richard, M., and R. Lippmann. 1991. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation* 3:461-483.

Ritter, H., T. Martinetz, and K. Schulten. 1992. *Neural computation and self-organizing maps*. New York: Addison-Wesley.

Rosch, E. 1977. Human categorization. In *Advances in cross-cultural psychology* v. 1, ed. N. Warren. London: Academic Press.

———. 1978. Principles of categorization. In *Cognition and categorization,* eds. E. Rosch and B. Lloyd. Hillsdale, NJ: Erlbaum.

Rosch, E., and C. Mervis. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* 7:573-605.

Rosch, E., C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8:382-439.

Rowlands, M. 1995. Against methodological solipsism: the ecological approach. *Philosophical Psychology* 8(1):5-24.

Sayre, K. 1986. Intentionality and information processing: an alternative model for cognitive science. *Behavioral and Brain Sciences* 9:121-166.

Schuster, H. 1991. Nonlinear dynamics and neuronal oscillations. In *Nonlinear dynamics and neuronal networks*, ed. H. Schuster. New York: VCH.

Schyns, P. 1991. A modular neural network model of concept acquisition. *Cognitive Science* 13: 461-508.

Searle, J. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417-424.

———. 1992. *The rediscovery of the mind*. Cambridge, MA: The MIT Press.

———. 1997. *The mystery of consciousness*. New York: The New York Review of Books.

Shanahan, M. 1997. *Solving the frame problem: a mathematical investigation of the common sense law of inertia*. Cambridge, MA: The MIT Press.

Shannon, C. 1949. The mathematical theory of communication. In *The mathematical theory of communication*. Urbana: University of Illinois Press.

Skarda, C., and W. Freeman. 1987. How brains make chaos to make sense of the world. *Behavioral and Brain Sciences* 10:161-195.

Smith, B. 1990. Putting Dretske to work. In *Information, language, and cognition*, ed. P. Hanson. Vancouver: University of British Columbia Press.

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11: 1-23.

———. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159-216.

Sober, E. 1985. Panglossian functionalism and the philosophy of mind. *Synthese* 64(2):165-193.

Sterelny, K. 1990. *The representational theory of mind: an introduction*. Cambridge, MA: Blackwell.

Stich, S. 1983. *From folk psychology to cognitive science: the case against belief*. Cambridge, MA: The MIT Press.

Strogatz, S. 1994. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Reading, MA: Addison-Wesley Publishing Company.

Stufflebeam, R. 1998. Representation and computation. In *A companion to cognitive science*, eds. W. Bechtel and G. Graham. Oxford: Blackwell Publishers Ltd.

Sun, R. 1995. An introduction: on symbolic processing in neural networks. In *Computational architectures integrating neural and symbolic processes: a perspective on the state of the art*, eds. R. Sun and L. Bookman. Norwell, MA: Kluwer.

Tam, K. 1998. Experiments in high-level robot control using ConGOLOG—reactivity, failure handling, and knowledge-based search. M.Sc. Thesis, Department of Computer Science, York University, Toronto.

Thelen, E., and L. Smith. 1994. *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: The MIT Press.

Touretzky, D., and G. Hinton. 1988. A distributed connectionist production system. *Cognitive Science* 12(3): 423-466.

Touretzky, D., A. Redish, and H. Wan. 1993. Neural representation of space using sinusoidal arrays. *Neural Computation* 5: 869-884.

Ullman, S. 1980. Against direct perception. *Behavioral and Brain Sciences* 3:333-81.

Van Gelder, T. 1991. What is the 'D' in 'PDP'? In *Philosophy and connectionist theory*, eds. W. Ramsey, S. Stich, and D. Rumelhart. Hillsdale, N.J.: Lawrence Erlbaum Associates.

———. 1995. What might cognition be, if not computation? *Journal of Philosophy* 91(7):345-81.

———. 1999. Wooden iron? Husserlian phenomenology meets cognitive science. In *Naturalizing phenomenology: issues in contemporary phenomenology and cognitive sci-*

ence, eds. J., Petitot, F. Varela, B. Pachoud, and J.-M. Roy. Stanford: Stanford University Press.

Vera, A., and H. Simon. 1993. Situated action: a symbolic interpretation. *Cognitive Science* 17:7-48.

Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-37: 328-339.

Weaver, W. 1949. The mathematics of communication. *Scientific American* 181: 11- 12.

Weizsäcker, C. von. 1971. *The unity of nature*. Translated by F. Zucker. New York: Farrar, Straus, Giroux.

Wittgenstein, L. 1953. *Philosophical Investigations*. Translated by G. Anscombe. Oxford: Blackwell Publishers, Inc.