BRAIN AND COGNITION **34,** 48–71 (1997)
ARTICLE NO. BR970906

Symbol Grounding: A Bridge from Artificial Life to Artificial Intelligence

Evan Thompson

Department of Philosophy, York University, North York, Ontario, Canada

This paper develops a bridge from AL issues about the symbol–matter relation to AI issues about symbol-grounding by focusing on the concepts of formality and syntactic interpretability. Using the DNA triplet-amino acid specification relation as a paradigm, it is argued that syntactic properties can be grounded as high-level features of the non-syntactic interactions in a physical dynamical system. This argument provides the basis for a rebuttal of John Searle's recent assertion that syntax is observer-relative (1990, 1992). But the argument as developed also challenges the classic symbol-processing theory of mind against which Searle is arguing, as well as the strong AL thesis that life is realizable in a purely computational medium. Finally, it provides a new line of support for the autonomous systems approach in AL and AI (Varela & Bourgine 1992a, 1992b).

INTRODUCTION

What is the relation between form and matter? An echo of this rather ancient philosophical question can be heard in the debates within two recent scientific endeavors, the computational approach to the mind-brain in computational neuroscience and artificial intelligence (AI) and the synthetic ap-

An earlier and considerably abridged version of this article was published in R. S. Cohen and M. Marion (Eds.), *Quebec Studies in the Philosophy of Science, Volume II* (Thompson, 1995). Versions of the paper have been presented to various conferences and institutions: to a symposium on Connectionism at the 1992 meeting of the Canadian Philosophical Association at the University of Prince Edward Island, to the Department of Philosophy at the University of Toronto, and at the 1993 workshop ''Artificial Life: A Bridge towards a New Artificial Intelligence,'' organized by the Department of Logic and Philosophy of Science and the Faculty of Computer Science and Institute of Logic, Cognition, Language and Information, at the University of the Basque Country, San Sebastian, Spain. I am grateful to the audiences at all of these occasions for their comments. I have also benefited greatly from discussions with Kenneth Cheung, Ronald de Sousa, Stephen Grossberg, Stevan Harnad, Alvaro Moreno, Sonia Sedivy, Tim Smithers, Paul Thompson, and Francisco Varela. Address reprint requests to Evan Thompson, Department of Philosophy, York University, 4700 Keele Street, North York, Ontario M3J 1P3, Canada.

0278-2626/97 $25.00

proach to living systems in artificial life (AL). In AI and AL the question arises mainly in connection with the nature and status of *symbols*, that is, items that are physically realized, formally identifiable, and semantically interpretable.

In classical or ''symbol processing'' AI the issue about symbols has two sides, one having to do with semantics and the other having to do with syntax. The semantic side is the problem of how symbols as syntactically (formally) individuated tokens of a given type get their meaning. The syntactic side, which has been much less discussed, is the problem of how something physical can also be syntactic. As a conceptual and practical design problem in AI the semantic side of the issue has been dubbed by Stevan Harnad the ''symbol grounding problem'' (Harnad, 1990). This term is better suited, however, to cover both the semantic and syntactic sides of the issue. The symbol grounding problem can thus be seen to comprise both a ''semantic grounding problem''—what fixes the semantic interpretation of the symbol system?—and a ''syntactic grounding problem''—what fixes the syntactic (formal) interpretation of the physical system?[1]

In AL the issues about symbols are broader. The main issue has been dubbed by Howard Pattee the ''symbol–matter problem'' (Pattee, 1989). The symbol–matter problem takes many forms. To cite a few examples: In a biological system, how can a material structure operate as a symbolic form? Does evolution require certain types of interaction between material structures and symbolic forms? Must something have a symbolic genotype and a material phenotype to be a living system? In other words, is the symbolic genotype/material phenotype coupling a necessary feature of life? Or is it only contingent, accidental to life as we know it on Earth?

According to Christopher Langton's influential AL manifesto (Langton 1989a), AL seeks to answer the above sorts of question by placing life as we know it within the wider theoretical context of life as it could be. Thus the AL research strategy involves trying to discover the organizational principles that define life by simulating the behavior of actual and possible living systems on a computer. But Langton also puts forward a more controversial statement of the AL program, one which raises another type of issue about the symbol–matter relation. This more controversial version of the AL program has come to be known as ''Strong AL,'' a name drawn from the supposedly analogous idea of ''Strong AI.'' (The latter term was coined by John Searle who opposes the position [Searle, 1980]). Whereas ''Strong AI'' holds that the mind is a computer program (that instantiating the right program suffices for having mental states), ''Strong AL'' holds that the properties necessary and sufficient for life are purely formal, and so it is possible not only to simulate but also to realize living systems within a computational medium (Langton 1989a; Morris 1991; Sober 1992). How cogent is this no-

[1] This way of looking at the issue was originally suggested to me by Kenneth Cheung.

tion of a system that qualifies as living yet has only a formal identity? And to what extent can a formal system clarify the symbol–matter relations that are characteristic of life on Earth (Emmeche, 1992; Pattee, 1989)?

The general aim of this paper is to suggest a possible bridge from these issues about the symbol–matter relation in AL to the symbol grounding problem in AI. I will focus primarily on the concepts of formality and syntactic interpretability in relation to the symbol–matter and symbol grounding problems. My specific concern is to build a bridge that leads from AL and the biological realm to a conception of symbolic activity in AI-inspired cognitive science that is more dynamical than the conception found in the classical formalist paradigm.

SEARLE'S ARGUMENT THAT ''SYNTAX IS NOT INTRINSIC TO PHYSICS''

Claus Emmeche (1992), in an endnote in his book on AL, observes that John Searle's (1990, 1992) recent argument against the classic symbol processing paradigm in cognitive science, is relevant also to AL. Emmeche does not explore this suggestion in the book; I will take it as my starting point here.[2]

The classic symbol processing paradigm in cognitive science holds that mental processes involve symbolic computations in an internal language of thought and that the brain is accordingly a digital computer or physical symbol system. Searle's earlier (and notorious) argument against this paradigm is his so-called ''Chinese Room Argument'' (Searle, 1980). This argument addressed the relation between the syntactic and semantic features of mental states. In contrast, his new argument addresses the relation between the physical and syntactic features of the brain. In a nutshell the new argument is that syntactic features are essentially observer-relative and so it is not possible for a physical system, such as the brain, to be intrinsically syntactic; consequently, the view that the brain is a ''syntactic engine'' is incoherent.

The core of Searle's argument has four steps, from which he then draws two further conclusions:

1. Computation is defined syntactically in terms of symbol manipulations.
2. Syntactic features are defined formally (roughly, in terms of certain shapes).
3. The formal and hence syntactic features of physical systems are always specified relative to some mapping or assignment function provided by an observer outside the system. (''`Syntax' is an observer-relative notion.'')

[2] Emmeche does refer to a paper of his in preparation, ''The idea of biological computation,'' where I imagine he discusses the issue further. But I have not seen this paper, and I do not know whether it has appeared in print. I discussed Searle's argument in relation to AL in an earlier and considerably abridged version of this article (Thompson 1995).

4. Therefore, physical systems are syntactic and hence computational only relative to the mapping or assignment function, not intrinsically. (''Syntax is not intrinsic to physics.'')

On the basis of (1)–(4) Searle then argues:

5. One cannot *discover* that a physical system (such as the brain) is intrinsically computational, though one can *assign* to it a computational interpretation.
6. Syntactic features cannot function causally in the production of behavior. (''Syntax has no causal powers.'')

It is important to see how this argument both differs from and tries to strike deeper than the Chinese Room Argument. The target of the Chinese Room Argument is the Strong AI position that the mind is a computer program, and the argument proceeds by trying to show that mental content cannot be gotten simply from the syntactic operations of computer programs. In contrast, the target of the new argument is the position, dubbed ''Cognitivism'' by Searle, that the *brain* is a digital computer or physical symbol system,[3] and the argument proceeds by trying to show that ''syntax'' is an essentially observer-relative notion and therefore it is incoherent to suppose that a physical system could be intrinsically syntactic. In Searle's formulation, the Chinese Room Argument tries to show that ''semantics is not intrinsic to syntax,'' but the new argument tries to show that ''syntax is not intrinsic to physics.'' Thus the argument tries to strike deeper because it does not challenge merely the idea that mental processes are computational; it challenges the very coherence of the idea of a physical symbol system. If Searle were right about the relation between physics and syntax, the implications for both AI and AL would be considerable: The hypothesis that the brain is a syntactic engine would be incoherent; consequently, the view that mental processes are computational would not even be able to get off the ground (unless one is a dualist). And Strong AL too would be untenable, for it would be incoherent to suppose that there could be an observer-independent computational medium in which to realize the principles definitive of life.

The crux of the issue as Searle sees it is whether there can be syntactic features that are intrinsic to a physical system rather than based on some outside assignment or interpretation. Thus for Searle the issue is an ontological one: Are syntactic properties intrinsic or observer-relative? And because he insists that there is a principled distinction between ontology and epistemology—questions about what something is must be treated separately from questions about how we know or determine what something is—Searle rejects attempts to answer the ontological issue about syntax on the basis of

[3] Searle uses the term ''digital computer'' but I think ''physical symbol system'' is a more precise designation for the kind of computational system he has in mind (see Newell 1980).

considerations about what fixes the syntactic *interpretation* of a physical system. His treatment of syntax thus runs parallel to his treatment of semantics and intentionality. Searle insists that there is a distinction between the ''intrinsic intentionality'' of mental states and the ''derived intentionality'' of, say, spoken utterances and written marks: As a matter of ontology, the intentionality of a mental state is internal to it and does not depend on how the state is used or interpreted (Searle, 1983). Thus ontological issues about the intentional content of mental states must be treated separately from epistemological issues about what fixes the semantic interpretation of mental states. Where the parallel breaks down, however, is that for Searle there is no such thing as ''intrinsic syntax''—syntactic properties do depend entirely on how the states of a system are used or interpreted.

Before taking up this issue in detail it is worth continuing for a moment to consider how Searle's position fits into his overall philosophy of mind. Consider the question, ''what exactly is the nature of the observer or agent who treats certain physical phenomena as syntactical''? Here two tenets of Searle's philosophy of mind become relevant. The first is the aforementioned distinction between intrinsic intentionality and derived intentionality. The second is Searle's view that, for a state to be an intentional or mental state, it is necessary that it be at least potentially accessible to consciousness; hence there cannot be mental states that are ''in principle inaccessible to consciousness'' (Searle, 1992). Now if we combine these two tenets with Searle's claim that syntax is observer-relative, we arrive at the claim that physical systems have syntactic properties only in relation to observers who are conscious intentional agents. Although Searle does not himself make this connection explicit, it is implicit in the following passage:

> In Turing's human computer [a person consciously applying an algorithm to solve an arithmetical problem] there really is a program level intrinsic to the system, and it is functioning causally at that level to convert input to output. This is because the human is consciously following the rules for doing a certain computation, and this causally explains his performance. But when we program the mechanical computer to perform the same computation, the assignment of a computational interpretation is now relative to us, the outside homunculi. There is no intentional causation intrinsic to the system. The human computer is consciously following rules, and this fact explains his behavior, but the mechanical computer is not literally following any rules. It is designed to behave exactly as if it were following rules; so for practical, commercial purposes it does not matter that it is not actually following any rules. It could not be following rules because it has no intentional content intrinsic to the system that is functioning causally to produce behavior. Now cognitivism tells us that the brain functions like the commercial computer and that this causes cognition. But without a homunculus, both commercial computer and brain have only patterns, and the patterns have no causal powers in addition to those of the implementing medium. So it seems there is no way cognitivism *could* give a causal account of cognition. (Searle, 1992, p. 216)

Searle's position thus turns out to be in a sense the inverse of Daniel C. Dennett's (1978, 1987). Dennett holds that no intentionality is intrinsic; all

intentionality is relative to a stance we adopt toward a system, the ''intentional stance.'' Adopting the intentional stance serves to reveal ''real patterns'' (Dennett, 1991) in the system's behavior, but the explanation of how these patterns are generated resides at the design level, the level at which a complex system is a ''syntactic engine.'' For Searle, on the other hand, physical systems can be syntactic engines only in relation to the conscious intentional agents who can assign syntactic interpretations to physical phenomena.

Although Searle's position on syntax is thus linked to his general position in the philosophy of mind, his claim that ''syntax is not intrinsic to physics'' can be rebutted without entering into the philosophical debates about intentionality and consciousness. First, one can sidestep Searle's problematic division between ontology and epistemology by holding that, as a matter of scientific practice, questions about what something is cannot be treated separately from questions about how we find out what something is. In the present context, this would mean that the issue about what syntactic properties are cannot be treated separately from the issue of how a system gets syntactically interpreted. On this view, the parallel between syntax and semantics would be complete: Just as when one discovers that a system is semantically interpretable there need be no further question about whether it is ''really'' semantic (see Haugeland, 1985), so too when one discovers that a system is syntactically interpretable there need be no further question about whether it is ''really'' syntactic.

Of course, ''really'' is not necessarily equivalent to ''intrinsically.'' Something can be really *F* without being intrinsically *F;* it might be *F* only relationally. There is therefore an interesting question that remains about syntactic interpretation: Does what fixes the syntactic interpretability of a system imply that syntax is *assigned* to the system? Or does it imply that syntax is *discovered* in the system?

These questions cannot be answered *a priori;* instead, one must look at the methodological and empirical justifications given on behalf of the syntactic level in computational endeavors such as AI and AL. The next two sections will be devoted to this task and they will reveal that there is a straightforward refutation of Searle's claim that ''syntax is not intrinsic to physics.'' The line of argument to be pursued will have some interesting consequences, however. It will provide little support to the view that Searle is actually arguing against—the cognitivist or symbol processing paradigm; and it will raise some questions about the cogency of the Strong AL thesis. But it will also turn out to provide a new line of support for the so-called ''autonomous systems'' research program within AL, biology, and connectionist cognitive science (Varela, 1979; Varela & Bourgine, 1992a, 1992b). And it is from this research program that a more dynamical, rather than purely formalist, conception of symbolic activity might be forthcoming.

SYNTAX AND THE BRAIN

We can begin by looking at the standard cognitivist arguments for considering the brain to be, at a certain level of description, a physical symbol system. Two main arguments have been given. The first appeals to *intentionality*, specifically to the intentional or semantic properties of mental states that are implicated in the generation of behavior and that are presumed to be in some way realized in the brain. The second appeals to *complexity*, specifically to the organizational complexity of the brain and nervous system.

The first argument has been extensively presented by Fodor and Pylyshyn (1988). According to Fodor and Pylyshyn, the symbol processing model of mental processes rests on two ideas. The first idea is that one can construct languages in which semantic features correspond systematically (within certain well-known limits) to syntactic features. The second idea is that one can devise machines that have the function of operating on symbols, but whose operations are sensitive only to the syntactic structure (physical form) of the symbols. Fodor and Pylyshyn describe how the two ideas fit together:

> If, in principle, syntactic relations can be made to parallel semantic relations, and if, in principle, you can have a mechanism whose operations on formulas are sensitive to their syntax, then it may be possible to construct a *syntactically* driven machine whose state transitions satisfy *semantical* criteria of coherence. Such a machine would be just what's required for a mechanical model of the semantical coherence of thought; correspondingly, the idea that the brain *is* such a machine is the foundational hypothesis of Classical cognitive science. (1988, p. 30)

As they go on to discuss, this ''foundational hypothesis'' implies that the symbol structures in the classical model of mental processes correspond to real physical structures in the brain. What Searle's argument challenges is the assumption that there could be any non-observer-relative fact of the matter about this correspondence. Hence what needs to be considered is how the syntactic features of computational states are typically grounded within the cognitivist or symbol processing framework.

In his book *Computation and Cognition,* Pylyshyn (1984) discusses how computational explanations must be relativized to the mappings provided by two interpretation functions, the *semantic function* (SF) and the *instantiation function* (IF). The semantic function maps from articulated functional states onto some domain of intended interpretation (e.g., positive integers). (It is worth noting parenthetically that this function is required for something to be a computation because a computation is a rule-governed process defined over *semantically* interpretable items (Fodor, 1980; Pylyshyn, 1984). Hence the first premise of Searle's argument, that computation is defined only *syntactically,* is not strictly speaking correct within the context of the symbol processing approach.) The instantiation function maps from physical states to computational states; more precisely, it specifies the equivalence classes of physical states that count as syntactically distinct computational states.

This is the interpretation function relevant to syntactic interpretability, for its purpose is to indicate how syntactic states are physically realized.

Unfortunately Pylyshyn's treatment of the instantiation function does not provide much clarification about syntactic interpretability. Here is how he describes the IF:

> By mapping from physical to computational states, such a function provides a way of interpreting a sequence of nomologically governed physical state changes as a *computation,* and therefore of viewing a physical object as a *computer*. (1984, p. 56)

This description does not make clear just what is required of an instantiation function for it to be the case that the syntactically distinct states figuring in the interpretation are not simply assigned to the system, but discovered in it.

The problem becomes more pressing when one remembers that computational states are said to be independent of any particular material medium and hence multiply realizable. As Pylyshyn goes on to observe, the physical realizations of a given computational function are essentially open-ended: Computational sequences ''can be realized in devices operating in any imaginable media—mechanical (as in Charles Babbage's Analytical Engine), hydraulic, acoustical, optical, or organic—even a group of pigeons trained to peck as a Turing machine!'' (1984, p. 57). Such multiple realizability has typically been considered as a point in favor of applying the computational framework to the mind, for it provides a model of how there need be only weak token-token identities between mental states and physical states (rather than strict type-type identities), and consequently how psychology can be autonomous in relation to neuroscience. Searle counters, however, that the multiple realizability of computational states is simply a sign that the states are not intrinsic to the system, but depend on an interpretation from outside. He thinks that a distinction should be drawn between devices whose functions are multiply realizable (e.g., thermostats) but are nevertheless defined in terms of the production of the same physical effects (e.g., regulating temperature), and devices whose multiple realizability is due to the relevant properties being purely formal or syntactic (e.g., Turing machines). For Searle, it is this difference in types of multiple realizability that explains why, for example, nobody would suppose it possible to make a thermostat out of pigeons even though it might be possible to train pigeons to peck as some simple Turing machine.

This distinction between what might be called *functional* multiple realizability, on the one hand, and *formal* multiple realizability, on the other, indicates that there is another concept that is revelant here. This is the concept of *digitality.* One reason that is typically given for why physical symbol systems can be multiply realizable is that they are *digital* systems. There is disagreement among philosophers and cognitive scientists about how to define a digital system (Haugeland, 1981; Lewis, 1971; Pylyshyn, 1984), but

for our purposes a digital system is one whose states belong to a finite number of types that are perfectly definite—that is, for any given type, a state is either of that type or it is not, and variation among the states that belong to a given type is insignificant. In physical symbol systems, the insignificant variations occur at the physical level (the level of the so-called ''physical machine'') and the perfectly definite types correspond to the syntactic states of the system (the states at, say, either the level of the ''logical machine'' or that of the ''abstract machine''). Thus syntactic properties can correspond to arbitrarily many physical properties and syntactic state-transitions can involve arbitrarily many physical causal laws. Hence physical symbol systems are as a class independent of any particular material medium.

One rather large shortcoming of Searle's argument is that he does not address this relation between digitality and formal multiple realizability. But certain remarks that he makes do suggest how he would view the relation (p. 210). In these remarks Searle appears to admit that digitality is not observer-relative, but he claims that although the states of a digital system might more naturally support a syntactic characterization by an outside observer, they are nevertheless syntactic *only* in relation to such a characterization.

To address this claim we need to turn to the second line of reasoning given by the cognitivist for supposing that the brain is a physical symbol system. This argument is the one based on appealing to the organizational complexity of the brain.

Ray Jackendoff (1987) gives a concise presentation of the argument from organizational complexity in his book, *Consciousness and the Computational Mind.* Jackendoff argues that computational levels of description are required for systems whose components interact combinatorially. In a computer, the state of each component (e.g., binary switch, flip-flop—hence at the level of the logical machine) is independent of the states of the other components; consequently, the activity of larger components in the machine is not a sum or average of the activity of the component parts, but depends rather on the state of each component and their particular combinatorial properties. It is the complexity of the combinatorial properties that is ultimately responsible for the medium-independence of computational systems: ''any device with the same combinatorial properties will have the same set of possible states and will go through the same sequence of states. Since the combinatorial properties are formal rather than physical, we will have arrived at a computational description'' (1987, p. 30).

Jackendoff claims that a comparable case obtains in the organizational complexity of the brain, though not in other biological organs, such as the stomach. The activity of neuronal groups is not the sum or average of the component neuronal activities, but depends rather on how the neurons interact combinatorially. Such combinatorial properties are both formal and structurally grounded in the brain; hence it is in virtue of its organizational complexity that the brain is a syntactic engine.

The original inspiration and foundation for this approach to the brain is of course the seminal paper by Warren McCulloch and Walter Pitts (1943) entitled ''A Logical Calculus of the Ideas Immanent in Nervous Activity.'' By taking the neuron as the functional unit of the nervous system and by treating neuronal activity as binary, digital, and synchronous (neurons as threshold devices that could be either active or inactive, and that all change state at the same discrete time-steps), McCulloch and Pitts were able to show formally that the organizational complexity of various neural nets is sufficient for the computation of the logical operations of the propositional calculus. This result has been foundational for the entire field of cognitive science because it shows how the operation of the nervous system might at some level be appropriately described using mathematical logic and hence how the brain could be seen at that level to be a symbolic machine.

Appealing to organizational complexity provides the beginnings of a more satisfactory approach to the issues about syntactic interpretability. Suppose, for example, that McCulloch and Pitts's idealized neural nets reflected the real functional architecture of the brain. It would then be the case that certain equivalence classes of neuronal states would be computational states in virtue of the *roles* that neurons play in nervous system activity. Of course it would still be true that in constructing a computational *model* of the brain *we* would have to map the all-or-none activity of neurons onto, say, 0's and 1's in binary notation, but the model would nonetheless be nonarbitrarily grounded in an intrinsic (non-observer relative) feature of the brain, namely, the role that neuronal activity plays in the operation of the nervous system and the generation of behavior. Searle seems to consider this line of argument, but he simply dismisses it by repeating his assertion that ''syntax is not intrinsic to physics,'' which of course begs the entire question.

My invocation of McCulloch and Pitts is *not* intended as a claim about how computational processes are actually supported by brain processes. On the contrary, it is well known that real neurons are not simple binary switches, and although neurons are the fundamental *anatomical* units of the nervous system, the basic *functional* units are probably relatively invariant patterns of activity in neuronal assemblies (Maturana & Varela, 1980; Grossberg, 1980; Edelman, 1987; Freeman & Skarda, 1985; Singer, 1993). Rather, by invoking McCulloch and Pitts I mean to illustrate the point that, contrary to Searle, there is nothing incoherent in supposing that computational processes may be thus grounded in the organizational complexity of the nervous system. It is for this reason that I said the argument from organizational complexity provides the beginnings of an approach to syntactic interpretability: It answers the charge raised by Searle, while leaving open the substantive (and ultimately more interesting) empirical issues about the syntactic interpretability of complex systems such as the brain.

The line of argument I have been pursuing does have theoretical implications for these empirical issues. We have seen that to ground claims about syntactic interpretability it is not enough to appeal simply to an abstract map-

ping from the syntactic to the physical, such as Pylyshyn's instantiation function; we must also appeal to the organizational complexity of the realizing system. For this reason I think that Fodor and Pylyshyn (1988) are wrong when they claim that it is mistaken for philosophers such as Dennett (1978, 1987) and connectionists such as Smolensky (1988) to appeal to complexity as a way of distinguishing cognitive from noncognitive systems. On the contrary, it is the organizational complexity of certain systems that warrants the hypothesis that there are additional syntactic (and semantic) levels of description for their behavior.

One consequence of this point is that the purely top-down approach that characterizes the symbol processing treatment of syntactic interpretability is unsatisfactory. What I mean by ''purely top-down'' is an approach in which hypotheses about the *brain* are made entirely on the basis of hypotheses about *mental representations.* For example, Fodor and Pylyshyn write:

> . . . the symbol structures in a Classical model are assumed to correspond to real physical structures in the brain and the *combinatorial structure* of a representation is supposed to have a counterpart in structural relations among physical properties of the brain. For example, the relation ''part of,'' which holds between a relatively simple symbol and a more complex one, is assumed to correspond to some physical relation among brain states. . . . (1988, p. 13)

Such a top-down approach cries out for lower-level constraints. Fodor and Pylyshyn simply take for granted the idea that there is some appropriate instantiation function from syntactic structures to physical structures in the brain, even though, to use their example, no one has the foggiest idea how the syntactic relation ''part of'' is supposed to correspond to some physical relation among brain states. If the claim that the brain is a ''syntactic engine'' is to be taken seriously, then we must try to find out what type of syntactic engine it is. To do this we need not only top-down hypotheses based on models of mental representations, we also need detailed models of how syntactic features can, in the first instance, emerge as a consequence of a system's organizational complexity, and in the second instance, act to constrain the behavior of such a system. Nowhere is the necessity of such models provided for within the symbol processing approach.

In contrast the new research area known as emergent computation holds considerable promise (Forrest, 1990, 1991). Emergent computation is the study of complex systems having three general features: (i) they are composed of a collection of agents each of which follows explicit instructions; (ii) the agents interact according to the instructions and thereby generate implicit, emergent global patterns; and (iii) there is an interpretation function that maps the global patterns onto computations (Forrest, 1990). Within such emergent computational systems the low-level agents are themselves devices that have a purely formal specification, but since they are typically rather simple—for example, the on-off cells of a cellular automaton—one can easily imagine biological analogues.

Nevertheless, merely showing how certain kinds of global behavior can be given a computational interpretation is not in itself sufficient to show that their syntactic interpretability is equivalent to the syntax being discovered in the system, rather than being merely assigned to it from outside. The task specified by the problem about syntactic interpretability is to show how (*contra* Searle) syntax can be intrinsic to physics for certain kinds of systems. Consequently, what we need are not merely implicit (unprogrammed) global patterns that naturally support a computational interpretation from outside (as in for example Conway's ''Game of Life'' (Gardner, 1970); rather, what we need are models of how a system can have an internal syntax that is not merely assigned by an outside observer and thereby qualify as a naturally syntactic system. Such a model would have to specify a particular sort of relation between syntax and physics—*one that emerges as a consequence of the system's own operation and yet can shape the system's behavior*.

THE LIVING CELL

One well understood example of a system that has a syntactic level of operation in this sense is the living cell. In fact the cell provides a minimal yet paradigmatic example of a system that is both autonomous (in a sense that will be made precise later) and that has a naturally syntactic level of operation, and so I will dwell on it at some length.

In a cell the syntactic level corresponds of course to the so-called ''genetic code'' whereby genes (lengths of DNA) specify the kinds of proteins a cell can make. More precisely, the ''genetic code'' refers to the rules that prescribe specific amino acids given specific triplets of nucleotide bases in DNA. Protein synthesis is thus said to involve specifications that are written in DNA and then ''decoded'' in a complex process involving molecular transcription (production of mRNA by RNA polymerase and nucleotides) and translation (production of protein by mRNA, ribosome, amino acid, tRNAs and other molecules).

The specification relation between DNA and proteins has a number of features that justify using the syntactic term ''code'' to describe it. First, the code is *quasi-universal:* With the exception of mitochondria, nucleotide triplets always specify the same amino acids regardless of the organism; for example, the triplet AAG (adenine–adenine–guanine) specifies lysine in all organisms from bacteria to humans. Second, the code is *arbitrary* in the sense that, as Maynard Smith puts it, ''It is hard to see why a code in which GGC means glycine and AAG means lysine is either better or worse than one in which the meanings are reversed'' (Maynard Smith, 1986, p. 19). Third, the code is *compositional* in the sense that there are numerous possibilities for nucleotide triplet combinations in a linear array. Fourth, the code is *digital* because what a nucleotide triplet signifies depends on which token it is out of a finite number of types and physical variation within certain

limits among tokens of a given type makes no difference to their belonging to that type (Maynard Smith, 1988).

These four features—quasi-universality, arbitrariness, compositionality, and digitality—warrant describing the genetic code as involving syntactic (formal) relations. But the genetic code does not depend on an interpretation function given by an outside observer; nor is it a set of instructions applied by a control that is external to the cell. Rather, it is a set of specifications that is and must be embedded in the metabolic dynamics that constitutes the cell as an individual.

To give this point the emphasis that it deserves in this context, I wish to make use of Howard Pattee's proposed explanation of the relation between the physical and the syntactic in biological systems (Pattee, 1977). Pattee distinguishes between, on the one hand, the laws of nature, which are universal and closed (holonomic), and on the other hand, structures that conform to the laws of nature yet constrain the motion of matter in ways additional to the laws (nonholonomic auxiliary conditions or constraints). In Pattee's view, the presence of such additional structures in a system can ground the syntactic interpretability of a system: We map the structures onto syntactic properties (e.g., formal rules), which can be described without referring to the physical structures that realize them.

So far we have little more than what is also contained in Pylyshyn's notion of an instantiation function that maps from the physical to the syntactic. But Pattee goes further. Following Von Neumann's (1966) work on self-reproducing automata, he focuses on a certain class of complex system—those that can self-replicate in virtue of containing (among other things) their own description (self-describing systems)—and explicitly raises the key question: ''how can we tell if a self-describing system has its own internal language in any objective sense? How do we know we are not interpreting certain structures as descriptions, only because we recognize them as consistent with rules of one of our own languages?'' And he gives as an answer: ''we must further restrict our model of a complex system to remove the case of the external observer reading a message that is not really in the system itself. This restriction is achieved by requiring *that a complex system must read and write its own messages*'' (Pattee 1977, p. 262).

Again the living cell serves as a paradigm, and thus we return to the genetic code as a naturally syntactic system embedded in the internal operation of the cell. The writing of its own messages corresponds in the cell to the synthesis of the DNA molecules; the reading of its own messages corresponds to the elaborate process of protein formation according to the universal rules of the genetic code and the specific descriptions in the structural DNA. This code is both arbitrary, as previously mentioned, and rate-independent (what a given nucleotide triplet designates is independent of how fast it is written or decoded). But as Pattee goes on to observe, for the code to be read there

must ultimately be a transduction from the rate-independent, ''linguistic mode,'' as he calls it, to the rate-dependent ''dynamical mode.'' The transduction happens when the rate-independent linear array of amino acids folds to become a three-dimensional enzyme. Within the operation of the cell there is thus a transformation from the enzyme as something designated in the genetic code to the enzyme as an operational component. Moreover, as Pattee remarks, this transformation (the protein folding) is not itself described in the linguistic mode; it happens rather according to ''the *laws* of nature under the nonholonomic constraints of the *rule-constructed* polypeptide chain described by the cell's structural DNA'' (1977, p. 263).

We can now appreciate how the genetic code as a naturally syntactic system is and must be embedded in the internal operation of the cell. In general, nucleotide triplets are capable of predictably specifying an amino acid if and only if they are properly embedded in the cell's metabolism, i.e., in the multitude of enzymatic regulations in a complex chemical network. This network has a ''chicken and egg'' character at several levels. First, proteins can arise only from a decoding process, but this process itself cannot happen without proteins. Second, the protein specification and construction processes must be properly situated within the intracellular environment, but this environment is itself a result of those very processes. Finally, the entire cell is an autonomous system of a particular sort: It is what Francisco Varela and Humberto Maturana describe as an *autopoietic* system, that is, a self-producing system defined by an operationally closed network of processes that simultaneously both produce and realize the cell concretely in the physical space of its biochemical components (Varela et al., 1974; Maturana & Varela, 1980; Varela, 1979).

The embedding of the genetic code in the cellular dynamics has important consequences for our understanding of naturally syntactic systems. When we refer to the ''coding relation'' between DNA and proteins we are really choosing to focus on one particular sequence of events in the ongoing metabolic turnover of the cell. We do so by abstracting away from a number of the necessary intervening causal steps and by implicitly invoking a *ceteris paribus* clause to cover the metabolic processes (Varela 1979, p. 75). Thus when we talk about DNA *coding for* proteins we are not referring to a peculiar type of syntactic causal relation; rather we are abbreviating a lengthy but remarkably stable causal sequence of concrete biochemical events. It is precisely the stability and predictability of the entire sequence that grounds treating nucleotide triplets as in effect symbols that stand for amino acids.

It would seem, then, that the genetic code provides a counterexample to Searle's assertion that syntax is essentially observer-relative. I think that this is correct, subject to one important clarification.

Although I am arguing that it is legitimate to view the specification relation between DNA and proteins on a symbolic level, I am *not* arguing that we

can superimpose the software-hardware distinction from symbolic computation onto the living cell. For example, consider this statement from Freemon Dyson's book *Origins of Life:* ''Hardware process information; software embodies information. These two components have their exact analogues in living cells; protein is hardware and nucleic acid is software'' (Dyson, 1985). But protein and nucleic acids are most definitely *not* ''exactly analogous'' to hardware and software. The analogy is misleading for reasons mentioned two paragraphs back: although there is a legitimate sense in which the so-called ''self-description'' of the protein-structure of the cell contained in DNA can be described formally, it must be remembered that the term ''self-description'' is an abbreviated way of referring to relations that must be dynamically embodied. There is no precise analogue to this dynamical embodiment in the case of software.

It can be seen from the points made in the previous two paragraphs that the treatment of syntactic interpretability suggested by the example of DNA in the living cell is hardly conducive to the physical symbol system model of the brain, which is the target of Searle's argument. This model simply takes symbols at face value and treats them as if they were independent (in principle) of the neural processes from which they emerge and by which they are supported. In the case of the living cell, however—which is our best understood example of a system that has a naturally syntactic level of operation—we know that there is no such independence of the symbol vehicles from the dynamical context in which they are embedded. This point has important implications for the thesis of Strong AL and for the AI debate between connectionism and the symbol processing theory.

STRONG AL

According to Christopher Langton, ''the principle assumption made in [Strong] Artificial Life is that the 'logical form' of an organism can be separated from its material basis of construction, and that 'aliveness' will be found to be a property of the former, not of the latter'' (Langton, 1989a, p. 11). The context of this remark is a discussion of machines, and of how, in the development of cybernetics and the mathematical theory of computation, ''The 'logical form' of a machine was separated from its material basis of construction, and it was found that 'machineness' was a property of the former, not of the latter.'' The application of this idea to the corresponding idea in the case of organisms requires only the additional premise that living systems in general are a type or class of machine; the type or class can then be defined by specifying a particular organization (see Maturana & Varela, 1980, for the classic statement of this idea, and Fontana et al., 1994, for a recent reformulation).

In philosophy this general idea is familiar not so much from discussions of life but from discussions of mind, and is known as *functionalism* (Block,

1980). The idea is that the logical form of a mind can be separated from its material basis, and that mentality is a property of the former, not of the latter. As a theory about the nature of the mind, the idea is sometimes known as ''metaphysical functionalism,'' and in its most pure form is the thesis that what makes a state a *mental* state is not anything physical *per se,* but rather simply its *functional role,* that is, the role the state characteristically plays in relation to other states. The functional role of a state is something abstract and formal in the sense that it can be specified as a set of relations without referring to the materiality of the states that happen to embody those relations; and any material that can support the appropriate network of relations will suffice to realize the functional role. Thus the multiple realizability of mind follows from metaphysical functionalism. When metaphysical functionalism is combined with what is sometimes called ''computation-representation functionalism''—the thesis that, in psychological explanation, mental states should be analyzed into component algorithmic processes defined over symbols after the fashion of a computer program—then one arrives at the view that minds are realizable in a purely computational (symbolic) medium (Strong AI).

Adverting to functionalism enables us to specify the sense in which Strong AL claims that materiality and logical form can be separated in an organism. Strong AL would seem to be the computational version of metaphysical functionalism as applied to the biological domain of life instead of the psychological domain of mind. Strong AL holds that what makes a state a ''vital'' state (one involved in, e.g., metabolism, reproduction, etc.) is simply its functional role. Hence the logical form of an organism—the set of functional relations holding among all its component states and processes—can be specified without referring to the organism's material constitution; and the material constitution can be anything as long as it can support the appropriate set of functional relations. Thus multiple realizability follows in the biological domain. Strong AL also holds, however, that the logical form of an organism can be captured entirely in a symbolic description, and so we arrive at the view that life is realizable in a purely computational medium. There is a difference in the type of computational approach taken in Strong AI and Strong AL, however. In computation-representation functionalism, mental processes are supposed to be recursively decomposed in a top-down manner, whereas in AL biological processes are supposed to be recursively generated in a bottom-up manner. But this difference does not affect the main point, which is that the target domain of phenomena in each case (mind or life) is considered to be realizable in a purely computational medium.

In evaluating Strong AL, then, the question that needs to be answered is whether it is indeed possible to abstract the logical form of a living system from its material constitution in the form of a symbolic description that can also be claimed to be a realization of a living system. The discussion in the previous section suggests this may be impossible. Recall Pattee's (1977)

distinction between the rate-independent linguistic mode and the rate-dependent dynamical mode in cellular activity. Two central points connected with this distinction are important here. First, as Pattee emphasizes, the transduction from the first mode to the second (i.e., the protein folding) is not itself linguistically described, but rather is accomplished by the dynamic interaction of the cellular components according to the laws of nature. Second, as Pattee also emphasizes, if the transduction were linguistically described, the speed and precision with which it is accomplished would be considerably reduced. The conclusion at which he accordingly arrives is that ''we would not expect a complete formal description or simulation of a complex system to adapt or function as rapidly or reliably as the partially self-describing, tacit dynamic system it simulates'' (1977, p. 264).

This interdependence of matter and form in the bacterial cell has also been taken up by Emmeche (1992) in a recent paper. He argues that the timing of the processes is crucial in both the transcribing of DNA into mRNA chains and in the synthesizing of enzymes from amino acids by the ribosomes (translation). The timing is regulated by an ''attenuation control system'' that involves both the linguistic mode (protein coding) and the dynamical mode (the physical form of the RNA chain). Hence Emmeche suggests that Pattee may have drawn his linguistic-dynamical distinction too sharply: ''The 'linguistic mode' of the cell (i.e., instructions in the DNA) and the 'dynamic mode' (the workings of the machinery) are so closely connected in the prokaryote cell that the 'logic' that describes the behaviour of the cell is time-dependent and for some part implicitly represented in the machinery that reads the instructions (*pace* Pattee 1977)'' (Emmeche 1992, p. 470). In any case, the point I wish to extract from these considerations is that, because the timing of transcription and translation processes in the cell is crucial, the logical form of the cell as a dynamical system is not atemporal (as is typical of abstract symbolic forms), but rather time-dependent (temporally constituted). For this reason, matter and form may not be separable, even in principle, in biological systems such as the living cell.

The general problem with Strong AL, then, is how it conceives the relation between matter and form in the biological realm. Langton writes that ''Life is a property of *form*, not *matter*, a result of the organization of matter rather than something that inheres in the matter itself'' (1989a, p. 41). What is right with this statement is that life is an emergent phenomenon dependent on processes having a certain form or organization (Varela et al., 1974; Varela, 1979; Maturana & Varela, 1980; Fontana et al., 1994). But what is wrong with it is that, in the biological realm at least, form is something that, as Aristotle emphasized long ago, *does* inhere in the matter itself. Susan Oyama puts this well when she wrties: ''Form emerges in successive interactions. Far from being imposed on matter by some agent, it is a function of the reactivity of matter at many hierarchical levels, and of the responsiveness of those interactions to each other'' (Oyama, 1985, p. 22). Indeed, it is pre-

cisely this conceptualization of the relation between matter and form that is needed to provide a foundation for the familiar idea that, as Langton puts it, ''Neither nucleotides nor amino acids nor any other carbon-chain molecule is alive—yet put them together in the right way, and the dynamic behavior that emerges out of their interactions is what we call life'' (Langton, 1989a, p. 41).

It is worth pointing out that the argument just presented does not prevent applying the idea of multiple realizability per se to living systems. Rather, what it challenges is the stronger notion of pure computational realizability for living systems. In other words, none of the considerations advanced so far rules out the possibility of life being realizable in many different physical media (Searle's notion of physical multiple realizability); what the considerations call into question is the conceptual intelligibility of the idea of purely computational life (Searle's notion of formal multiple realizability as applied to life).

Similar conclusions have been reached by Stevan Harnad in his discussion of symbol grounding and AL (Harnad, 1994). Harnad argues that a computational model in general is a semantically ungrounded symbol system; it becomes grounded only when supplied with an interpretation. Therefore, a computational model of life, or of the organization proper to the living, cannot itself be alive: even if the model turns out to be formally equivalent to a living system (whatever this might mean exactly), it is so only relative to an interpretation, i.e., relative to some (semantic) interpretation function that maps the symbols of the model onto aspects of living processes. An actual living system, on the other hand, though it might involve symbolic processes of various sorts, is not itself an ungrounded symbol system.

Although Harnad and I agree that the Strong AL thesis should be rejected, the point of my argument is different from his. As Harnad clearly indicates in his paper, he is willing to allow that every aspect essential to life could be systematically mirrored in a formal model. I am skeptical of this idea, however. If, as Pattee and Emmeche suggest, the logical form of a living cell has a time-dependent aspect (having to do with rates of reaction, attenuation control mechanisms, etc.), then it is possible that the logical form of a living system—its organization—cannot be completely captured in a purely computational (symbolic) form. I state the point tentatively because a definitive resolution of the issue one way or another requires a fully worked out theory of the organization proper to life. The project of constructing such a theory has certainly begun (Varela et al., 1974; Varela, 1979; Maturana & Varela, 1980; Fontana et al., 1994), but remains to be completed.

A BRIDGE TO ARTIFICIAL INTELLIGENCE

The foregoing discussion has implications also for the debate in AI between connectionism and the symbol processing paradigm. As noted above,

the cognitivist model of the brain as a physical symbol system takes the symbolic level for granted and treats it as if it were independent (in principle) of the neural processes within which it is supposed to be realized. In the case of biological systems such as the cell, however, there is no such independence of the symbolic level from the encompassing dynamical context. Except for the biologically unconstrained, top-down hypotheses based on models of mental representation, there is no reason not to expect the same point to hold for the various types of formal regularities to be found in the brain and nervous system. Of course, in systems that are so organizationally complex it is much harder to specify the processes responsible for these regularities. Nevertheless, this difference in systemic complexity does not justify the indifference that the symbol processing paradigm typically shows to the dynamical context of symbolic activity.

This criticism of the symbol processing approach is nothing new to those engaged in the connectionist research program (Grossberg, 1982; Rummelhart & McClelland, 1986; Smolensky, 1988). Connectionist models do not take symbols at face value and then make entirely top-down hypotheses about the realization of symbol structures in the brain. Instead, symbols are typically treated in the connectionist approach as approximate macrolevel abbreviations of operations whose governing principles reside at a ''subsymbolic'' level (Smolensky, 1988).

This general type of relation between the symbolic and subsymbolic holds for the genetic code in living cells: To describe nucleotide triplets as ''coding'' for amino acids is to abbreviate a lengthy causal sequence of complex intracellular processes whose governing principles reside at a subsymbolic biochemical level. Thus my use of the cell as a minimal paradigm of how syntax can be intrinsic to physics turns out to provide considerations in support of connectionism.

The term ''connectionism'' is of course usually applied to models of subsymbolic principles in neural networks. The goal here is both to understand real, biological neural networks and to solve problems in the theory of machine learning. But J. Doyne Farmer (1990) has argued that the term ''connectionism'' should really be given a much broader signification. He defines a connectionist model as one in which the interactions between the variables at any given time are restricted to a finite list of connections, and the values of the connection strengths and/or the architecture of the connections themselves can change with time. Farmer then shows that this class of (meta) dynamical systems includes not only neural networks, but also classifier systems in AI, immune networks, and autocatalytic chemical reaction networks.

Within connectionism thus broadly construed, the orientation most relevant to the concerns of this paper is the theory of autonomous systems (Varela & Bourgine, 1989a, 1989b). The key distinction here is between *heteronomous* systems, which are defined by input-output functions and an external control, and *autonomous* systems, which are defined by internal processes

of self-organization. Francisco Varela (1979) has attempted to make this characterization more precise by saying that the processes that make up an autonomous system must (1) be related as a network, (2) generate and realize themselves, and (3) constitute the system as a unity in the domain in which those processes exist. Varela summarizes this idea of an autonomous system as a self-constituting network of processes in what he calls the ''Closure Thesis,'' which states that *every autonomous system has operational closure* (1979, p. 58; Varela & Bourgine, 1989a). The term ''closure'' is used here in its algebraic sense: a given domain has closure if all operations defined in the domain remain within the same domain. Thus ''operationally closed'' in this context does not mean that the system is physically and interactionally closed to the outside, but rather that inside and outside are determined by the self-constituting dynamics of the system itself.

Once again the paradigm example of a system that is autonomous in this general sense is the living cell. Intracellular processes generate and realize a metabolic network that has biochemical closure involving a membrane and that constitutes the cell as a biological individual. It is this particular sort of autonomy at the biochemical level that Varela and Maturana call *autopoiesis* (Varela et al., 1974; Maturana & Varela, 1980). Another important example of a system that is claimed to be autonomous (but not autopoietic) is the nervous system. As Maturana and Varela describe it:

> Operationally, the nervous system is a closed network of interacting neurons such that a change of activity in a neuron always leads to a change of activity in other neurons, either directly through synaptic action, or indirectly through the participation of some physical or chemical intervening element. Therefore, the organization of the nervous system as a finite neuronal network is defined by relations of closeness in the neuronal interactions generated in the network. Sensory and effector neurons . . . are not an exception to this because all sensory activity in an organism leads to activity in its effector surfaces, and all effector activity in it leads to changes in its sensory surfaces. (1980, p. 127)

This idea of the nervous system as an autonomous rather than heteronomous system is directly relevant to connectionism and its concern with the dynamical context of symbolic activity. For example, within neural network research systems whose learning is entirely ''supervised'' qualify as heteronomous because changes to the connections in the network are directed by an external training signal, as in the learning algorithm known as ''back-propagation'' (Rummelhart & McClelland, 1986; Rummelhart, 1990). Back-propagation cannot be defined without reference to such a training signal that is outside and independent of the system; thus back-propagation connectionist systems cannot be autonomous. In contrast systems whose learning is ''unsupervised'' capture a key aspect of autonomy because changes to the connections in the network typically depend on cooperative and competitive relations among the nodes without the direction of any external supervisor (Grossberg, 1987; Carpenter & Grossberg, 1990; von der Marsburg, 1990).

Hence it is to this sort of autonomous neural network research that we should look for models of how symbolic processes emerge in the brain and nervous system.

One promising example is the ''adaptive resonance'' neural network theory of Stephen Grossberg, Gail Carpenter, and their colleagues (Grossberg, 1980, 1982, 1987; Carpenter and Grossberg, 1990). Their ART (adaptive resonance theory) and more recent ARTMAP architectures use unsupervised learning principles, but they can also operate in a supervised way when there is feedback from the environment. The architectures embed competitive learning principles in a self-regulating control structure that contains both attentional and orientational subsystems. The cycles of interaction between these two subsystems enable the network to self-organize in real time stable internal configurations or ''recognition categories'' in response to arbitrary sequences of arbitrarily many input patterns without any prior explicit representation of the environment. Grossberg and Carpenter call a set of such stable internal configurations a ''recognition code;'' the symbols (recognition categories) that constitute the code are ''compressed, often digital representations, yet they are formed and stabilized through a process of resonant binding that is distributed across the system'' (Carpenter & Grossberg, 1993). Adaptive resonance theory thus provides one model of how stable formal configurations relevant to perception and action can emerge as a result of distributed subsymbolic processes and then act to shape the adaptive behavior of the system. For this reason, it provides the right sort of model for understanding how syntax can be intrinsic to physics in the case of neural and cognitive processes. (By ''right sort'' I mean one that meets the requirement stated at the end of Section 3 of addressing how formal regularities can emerge as a consequence of a system's autonomous operation while also serving to shape what the system can do.)

The autonomous systems research program in AL and AI thus seems to be in a good position to claim that it may be able to discover the ''laws of qualitative structure'' underlying symbolic activity in complex systems. This idea of a law of qualitative structure was originally invoked in cognitive science by Alan Newell and Herbert Simon (Newell & Simon, 1977) on behalf of their physical symbol system hypothesis. But as we have seen the purely top-down approach this hypothesis takes toward symbol grounding is unsatisfactory. In contrast connectionism and in particular the theory of autonomous systems explicitly consider the dynamical context of symbolic activity, and so hold out the promise of a formal theory of the necessary and sufficient conditions for symbolic activity in natural and artificial dynamical systems.

REFERENCES

Block, N. 1980. What is functionalism? In N. Block (Ed.), *Readings in the philosophy of psychology, volume 1* (pp. 171–184). Cambridge, MA: Harvard Univ. Press.

Carpenter, G., & S. Grossberg. 1990. Self-organizing neural network architectures for real-time adaptive pattern recognition. In Zornetzer, S., J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks.* San Diego: Academic Press. Pp. 455–78.

Carpenter, G., & S. Grossberg. 1993. *Integrating symbolic and neural processing in a self-organizing architecture for pattern recognition and prediction.* Technical Report CAS/CNS-93-002, Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems.

Dennett, D. C. 1978. *Brainstorms.* Cambridge, MA: The MIT Press. A Bradford Book.

Dennett, D. C. 1987. *The international stance.* Cambridge, MA: The MIT Press. A Bradford Book.

Dennett, D. C. 1991. Real patterns. *Journal of Philosophy* **88,** 27–51.

Dyson, F. 1985. *Origins of life.* Cambridge: Cambridge Univ. Press.

Edelman, G. 1987. *Neural Darwinism.* New York: Basic Books.

Emmeche, C. 1992. Life as an abstract phenomenon: is artificial life possible? In F. J. Varela & P. Bourgine (Eds.), *Toward a practice of autonomous systems: Proceedings of the first European conference on artificial life.* Cambridge, MA: The MIT Press. A Bradford Book. Pp. 464–74.

Emmeche, C. 1994. *The garden in the machine: The emerging science of artificial life* (S. Sampson, Trans.). Princeton: Princeton Univ. Press.

Farmer, J. D. 1990. A Rosetta Stone for connectionism. *Physica* D **42,** 153–187. Reprinted in [16].

Fodor, J. 1980. *Representations: Philosophical essays on the foundations of cognitive science.* Cambridge, MA: The MIT Press. A Bradford Book.

Fodor, J., & Z. Pylyshyn. 1988. Connectionism and cognitive architecture: a critical review. *Cognition* **28,** 3–71.

Fontana, W., G. Wagner, & L. W. Buss. 1994. Beyond digital naturalism. *Artificial Life* **1**(1/2), 211–227.

Forrest, S. 1990. Emergent computation, self-organizing, collective, and cooperative phenomena in natural and artificial computing networks. Introduction to the Proceedings of the Ninth Annual CNLS Conference. *Physica* D **42,** 1–11. (Reprinted in Forrest, 1991).

Forrest, S. (Ed.) 1991. *Emergent computation.* Cambridge, MA: The MIT Press.

Freeman, W., & C. Skarda. 1985. Spatial EEG patterns, nonlinear dynamics, and perception: the neo-sherringtonian view. *Brain Research Reviews* **10,** 145–175.

Gardner, M. 1970. Mathematical games: the fantastic combinations of John Conway's new solitaire game of life. *Scientific American* **224,** 120–123.

Grossberg, S. 1980. How does a brain build a cognitive code? *Psychological Review* **87,** 1–55. (Reprinted in Grossberg, 1982).

Grossberg, S. 1982. *Studies of mind and brain: Neural principles of learning, perception, development, cognition and motor control.* Boston Studies in the Philosophy of Science, Vol. 70. Dordrecht: D. Reidel.

Grossberg, S. 1987. Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science* **11,** 23–63.

Harnad, S. 1990. The symbol grounding problem. *Physica* D **42,** 335–346. (Reprinted in Forrest, 1991).

Harnad, S. 1994. Levels of functional equivalence in reverse bioengineering. *Artificial Life* **1**(3), 293–301.

Haugeland, J. 1981. Analog and analog. *Philosophical Topics* **12,** 213–225.

Haugeland, J. 1985. *Artificial intelligence: The very idea.* Cambridge, MA: The MIT Press. A Bradford Book.

Jackendoff, R. 1987. *Consciousness and the computational mind.* Cambridge, MA: The MIT Press. A Bradford Book.

Langton, C. G. 1989a. Artificial life. In Langton, C. G. (Ed.), *Artificial life.* Santa Fe Studies

in the Sciences of Complexity Volume VI. Redwood City, CA: Addison-Wesley. Pp. 1–47.

Langton, C. G. (Ed.) 1989b. *Artificial life.* Santa Fe Studies in the Sciences of Complexity Volume VI. Redwood City, CA: Addison-Wesley.

Langton, C. G., C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.). 1992. *Artificial life II.* Santa Fe Institute Studies in the Sciences of Complexity, proceedings Volume X. Redwood City, CA: Addison-Wesley.

Lewis, D. 1971. Analog and digital. *Nous* V: 321–327.

Maturana, H. R., & F. J. Varela. 1980. *Autopoiesis and cognition: The realization of the living.* Boston studies in the Philosophy of Science, Volume 43. Dordrecht: D. Reidel.

Maynard Smith, J. 1986. *The problems of biology.* Oxford: Oxford Univ. Press.

McCulloch, W. S. & W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 9: 115–133. (Reprinted from *Embodiments of mind*, by W. S. McCulloch, 1965, Cambridge, MA: The MIT Press).

Meyer, J. A., & S. Wilson (Eds.). 1991. *From animals to animats.* Cambridge, MA: The MIT Press. A Bradford Book.

Morris, H. 1991. On the feasibility of computational artificial life: a reply to critics. In Meyer, J. A. & S. Wilson (Eds.). *From animals to animats.* Cambridge, MA: The MIT Press. A Bradford Book. Pp. 40–49.

Newell, A. 1980. Physical symbol systems. *Cognitive Science* **4,** 135–183.

Newell, A., & H. S. Simon. 1977. Computer science as empirical inquiry: symbols and search. *Communications of the Association for Computing Machinery* **19,** 113–126.

Oyama, S. 1985. *The ontogeny of information: Developmental systems and evolution.* Cambridge: Cambridge Univ. Press.

Pattee, H. H. 1977. Dynamic and linguistic modes of complex systems. *International Journal of General Systems Theory* **3,** 259–266.

Pattee, H. H. 1989. Simulations, realizations, and theories of life. In Langton, C. (Ed.), *Artificial life.* Santa Fe Studies in the Sciences of Complexity Volume VI. Redwood City, CA: Addison-Wesley. Pp. 63–75.

Pylyshyn, Z. 1984. *Computation and cognition.* Cambridge, MA: The MIT Press. A Bradford Book.

Rummelhart, D. E. 1990. Brain-style computation: learning and generalization. In Zornetzer, S., J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks.* San Diego: Academic Press. Pp. 405–20.

Rummelhart, D. E., & J. L. McClelland (Eds.) 1986. *Parallel distributed processing: Explorations in the microstructure of cognition. Volume one: Foundations.* Cambridge, MA: The MIT Press.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* **3,** 417–458.

Searle, J. R. 1983. *Intentionality: An essay in the philosophy of mind.* Cambridge: Cambridge Univ. Press.

Searle, J. R. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association* **64,** 21–37.

Searle, 1992. *The rediscovery of the mind.* Cambridge, MA: The MIT Press. A Bradford Book.

Singer, W. 1993. Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* **55,** 349–74.

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11,** 1–74.

Sober, E. 1992. Learning from functionalism—Prospects for strong artificial life. In Langton, C. G., C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.) (1992). *Artificial life II.* Santa Fe Institute Studies in the Sciences of Complexity, Proceedings Volume X. Redwood City, CA: Addison-Wesley. Pp. 749–65.

Thompson, E. 1995. Artificial intelligence, artificial life, and the symbol-matter problem. In R. S. Cohen & M. Marion (Eds.), *Quebec studies in the philosophy of science, volume*

II: Biology, economics and psychology. Boston Studies in the Philosophy of Science. Dordrecht: Kluwer.

Varela, F. J. 1979. *Principles of biological autonomy.* New Jersey: Elsevier, North Holland.

Varela, F. J., H. Maturana, & R. Uribe. 1974. Autopoiesis: the organization of living systems, its characterization and a model. *Biosystems* **5,** 187–195.

Varela, F. J., & P. Bourgine. 1992a. Introduction: Toward a practice of autonomous systems. In Varela, F. J. & P. Bourgine, *Toward a practice of autonomous systems. Proceedings of the first European conference on artificial life.* Cambridge, MA: The MIT Press. A Bradford Book. Pp. xi–xvii.

Varela, F. J., & P. Bourgine. 1992b. *Toward a practice of autonomous systems. Proceedings of the first European conference on artificial life.* Cambridge, MA: The MIT Press. A Bradford Book.

von der Malsburg, C. 1990. Network self-organization. In Zornetzer, S., J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks.* San Diego: Academic Press. Pp. 421–432.

Von Neumann, J. 1966. *The theory of self-reproducing automata* (A. Burks, Ed.). Urbana: IL: University of Illinois Press.