Taking connectionism seriously:
the vague promise of subsymbolism and an alternative.

Paul F.M.J.Verschure
AI lab, Institute for Informatics, University of Zürich,
Winterthurerstrasse 190, CH- 8057 Zürich,
Switzerland.
e-mail: verschur@ifi.unizh.ch

Abstract

Connectionism is drawing much attention as a new paradigm for cognitive science. An important objective of connectionism has become the definition of a subsymbolic bridge between the mind and the brain.

By analyzing an important example of this subsymbolic approach, NETtalk, I will show that this type of connectionism does not fulfil its promises and is applying new techniques in a symbolic approach.

It is shown that connectionist models can only become part of such a new approach when they are embedded in an alternative conceptual framework where the emphasis is not placed upon what knowledge a system must posses to be able to accomplish a task but on how a system can develop this knowledge through its interaction with the environment.

Introduction

Connectionism has been gaining much attention in cognitive science. On of the reasons is that problems of the traditional cognitivistic approach, like the need for noise and fault tolerance and the capability to generalize, are solvable with connectionist, brain-like, techniques.

This proposal makes the problem of complete reduction (PCR) (Haugeland, 1978), or of how a symbolic description of cognition can be reduced to a non-symbolic one, again highly relevant.

In the traditional cognitivistic view cognition is seen as formal symbol manipulation. The basic steps of this approach can be defined as: "1, Characterize the situation in terms of identifiable objects with well defined properties. 2, Find general rules that apply to situations in terms of those objects and properties. 3, Apply the rules to the situation of concern, drawing conclusions about what should be done." (Winograd and Flores, 1986, p.15).

The physical symbol system hypothesis (Newell, 1980) can be taken as the most influential formulation of this approach. The hypothesis states that a physical symbol system (PSS) constitutes the necessary and sufficient conditions for general intelligence. A PSS consists of a set of actions and is embedded in a world that consists of discrete states; objects and their relations. Moreover, a PSS has a "body of knowledge" that specifies the relations between the events in the world and the actions of the system, we can also refer to this body of knowledge as a world model built up with symbolic representations. The actions of the system, either in the world, or internal inferences, are organized around the goals of the system according to the principle of rationality: roughly a system will use its knowledge to reach its goals. An important implication of this conceptualization of cognition is that it can (and must) be modelled at the abstract level of symbol manipulation. The specifics of the implementation are, therefore, of no importance. PCR is no longer an issue since the non symbolic level of brain dynamics is not taken to be very relevant in explaining cognition.

The hypothesis of physical symbol systems is often seen as the only plausible model for general intelligence which has no serious competitors (e.g. Pylyshyn, 1989). Despite this claim this paradigm also confronts some serious problems. One of these problems is the symbol grounding problem (Harnad, 1990), or the question of how symbols acquire their meaning. In the cognitivistic tradition the meaning of symbols is taken as given (Newell, 1981), which implies that cognitivism has to resort to a nativistic position: that the "body of knowledge" is just present from the start on. Moreover, one has to assume that the system possesses very reliable transduction functions that allow the coupling between events and objects in the world and their internal symbolic representation. These assumption have been criticized on several grounds. For instance, the genome does not have the coding capacity to represent this body of knowledge (Edelman, 1987), or it still needs to be explained how during evolution this "body of knowledge" could have been acquired (Piaget in Piatelli-Palmarini, 1980). Moreover, practical applications developed within this paradigm, for instance robot control architectures, have not been

very successful (see Malcolm et al., 1989, for an overview).

A related issue is the frame of reference problem (FOR) (Clancey, 1992) which conceptualizes the relation between the designer, the observer, and the system. The designer of a system develops this system out of his/her domain ontology (i.e. a categorization of the task domain into events, objects, and relations). The consequence of this is that the knowledge on which the system is based is grounded in the experience of the designer and that this domain ontology is static.

An alternative position towards explaining cognition can be found in traditional connectionism (e.g. Rosenblatt, 1958). Here the hypothesis of formal symbol manipulation was rejected in favour of theories that take the dynamics of the brain into account. The appropriate tool here was not logic but statistics. It is assumed that by interacting in its environment an organism, which does not possess prior knowledge of this environment, develops preferences for specific responses to certain stimuli. The evolving associations between stimuli and responses are directly related to the development of distinct connection patterns in its nervous system. The classical example of this approach is the perceptron proposed by Rosenblatt (1958). Also in this case PCR is dissolved since the intentional level of symbol manipulation is not taken to be relevant in explaining behavior.

When we compare the solutions of PCR of both approaches they have two contradictory positions. While cognitivism emphasizes the importance of a formal symbol manipulating mind traditional connectionism underlines the importance of the dynamical brain. This contrast can be seen as a mind-brain dilemma (Verschure, 1992). Subsymbolic connectionism has an alternative position towards this dilemma.

Smolensky (1988) tried to define a theoretical framework for connectionism where he assumes that cognition, as described within classical cognitivism, is an *emergent* property of the interaction of a large number of units which are subsymbolic. His proposal is based on developments in the present main stream of connectionist research (e.g. Rumelhart and McClelland, 1986).

Smolensky assumes that in a connectionist model symbols are encoded by the 'complex patterns of activity over many units. Each unit participates in many such patterns ... The interactions between individual units are simple, but these units do not have conceptual semantics: they are subconceptual' (Smolensky , 1988, p. 6).

The subsymbolic description of cognition at the level of units is supposed to be, in principle, reducible to brain processes. The limited knowledge we have of the brain is here seen as the only barrier we have to take to complete this subsymbolic reduction of cognition.

Subsymbolic connectionism offers a new perspective on the relation between the mind and the brain. It assumes that both levels can be joined up by specifying "bridging principles" between the cognitivistic symbol manipulating mind and the dynamic brain. If this approach can show how PCR can be solved without rejecting one of the levels of description involved it can indeed be taken as progress.

To evaluate this claim of subsymbolic connectionism I will first analyze its paradigmatic example, NETtalk. This analysis will show that subsymbolic connectionism does not fulfil its promise to solve the mind-brain dilemma, but still constitutes, in essence, a symbolic approach. Next I will sketch an alternative framework which does allow a solution to this dilemma. Central to this alternative position is that in order to understand cognition the focus should not be on a predefined "body of knowledge", but on how this can be acquired through the system-environment interaction.

NETtalk: the example of subsymbolic reduction

NETtalk, the famous 'parallel network that learns to read aloud' by Sejnowski and Rosenberg (1986, 1987) is put forward by Smolensky, and others, as *the* example of subsymbolic reduction.

With NETtalk Sejnowski and Rosenberg have successfully built a model that could pronounce English words. Although they acknowledge the differences between the architecture of NETtalk and the brain they assume that NETtalk can teach us how information (in this case letter to phoneme mappings) is represented in 'large populations of neurons'.

The input layer of NETtalk consists of 7 identical groups of 29 units each. The letters of the alphabet plus 3 extra features representing word boundary and punctuation are coded in every group by a special unit. The hidden layer of NETtalk has no pre-assigned interpretation but is necessary to accomplish the mapping between the input- and the output layer. Every unit of the output layer represents one of 23 articulatory features or one of 3 features representing stress and syllable boundaries. The network learns, by means of back propagation, to associate the letter coded for by the active unit of the fourth group of the input layer with a specific set of articulatory features represented by a specific pattern of active output units. The other 6 groups of the input layer provide a context. The coupled activation patterns of the input- and output layer are determined by the designers of the system.

NETtalk is able to learn to correctly pronounce 95% of the presented words after training with 50000 words. It could correctly generalize to new cases in 78% of the test words.

Sejnowski and Rosenberg next tried to determine the features coded by the hidden units of NETtalk by clustering input patterns that lead to the same activation patterns of these elements. This cluster analysis of NETtalk showed that the activity

	Vowels:	Consonants:		Vowels:	Consonants:
Tensed	9	0	**Voiced**	**1**	**21**
Medium	8	0	**Unvoiced**	**1**	**12**
High	**6**	**1**	Fricative	0	9
Central 1	**5**	**1**	Palatal	0	8
Front 1	**5**	**1**	**Velar**	**1**	**8**
Front 2	**5**	**2**	Labial	0	7
Central 2	4	0	Stop	0	7
Low	4	0	Affricative	0	6
Back 1	2	0	Alveolar	0	6
Back 2	2	0	Nasal	0	6
			Dental	0	5
			Liquid	0	4
			Glide	**1**	**3**
			Glottal	0	1

Table 1: the frequency of occurrence of the articulatory features in coding vowels and consonants.

patterns of the hidden units could be understood as separating two main features: vowels and consonants. These results where considered to be an important proof of the power of subsymbolic computing: the emergence of a 'symbolic' separation of the letter to phoneme mapping in vowels and consonants.

A closer analysis of the letter to phoneme mapping the network has to learn shows, however, that the patterns presented to the network can beforehand be separated into two global categories: vowels and consonants. To illustrate this in Table 1 the 24 articulatory features represented by the units of the output layer are shown with their frequency of being involved in coding a vowel or a consonant. Articulatory features that are used to code both vowels and consonants are printed in bold face.

Table 1 shows that the features that are used to code about 95% of the vowels only code about 5% of the consonants and vice versa. Only 8 of the 24 features show an overlap and are used for coding vowels *and* consonants. Notice, however, that this overlap is always rather limited. For instance the feature "Unvoiced" is used 12 times in encoding a consonant and only once in encoding a vowel. Because every input letter is related to *a number* of articulatory features it can unambiguously be coded as a vowel or a consonant. Only one of the 51 symbols learned is completely defined by features related to the opposite class (the letter c as pronounced in logic is completely defined by articulatory features which mostly code vowels). See Verschure (1992) for an elaborate analysis.

NETtalk is put forward as a clear example of a model possessing subsymbolic representations. In this analysis it is shown, however, that the subsymbolic reduction given by NETtalk of the pronunciation of English words, expressed in the separation of vowels and consonants, is put in by the designers of the system. The vowels are always translated to a set of articulatory features of which we know beforehand that they distinguish vowels from consonants. Therefore, it is not surprising that NETtalk learns to discriminate them from the category of patterns coding consonants. The trick of subsymbolic reduction seems to lie in the transformation from the symbols (in this case articulatory features) to the actual activation patterns that NETtalk learns. This transformation, which conserves symbolic regularities (a vowel-consonant distinction), is made by the designers: Sejnowski and Rosenberg and not by NETtalk. Therefore, the claim that NETtalk started out *without* 'considerable "innate" knowledge in the form of input and output representations that were chosen by the experimenters' (Sejnowski & Rosenberg, 1987, p.158) does not relate to the reality behind the model.

The analysis of NETtalk suggests that subsymbolic reduction seems to boil down to a circularity consisting of the following steps: 1, The designer of the system defines basic symbolic properties in which a certain task can be described (in NETtalk articulatory features and characters); the knowledge the system must have to accomplish the task is defined. 2, These properties get translated to regularities of activation patterns presented to a connectionist model (in the case of NETtalk this is expressed in which letter should be pronounced with which set of pronunciation features). 3, The connectionist model learns to separate the patterns on their differences and groups them together on their regularities. These separations and groupings get expressed in the dynamics of the network, for instance in the activation of the hidden layer or in a specific distribution of the weights. 4, The regularities expressed in the dynamics of the network, which are completely determined by the regularities put in by the designers of the system, are symbolically interpreted by the designer (in the case of NETtalk as a vowel/consonant distinction). Steps 1 to 3 show a strong similarity to the ones of the cognitivistic approach listed earlier. It can be shown (Verschure, 1992) that this hypothesis concerning the circularity of subsymbolic reduction can easily be generalized to other connectionist models which have 'emergent' properties and

models that rely on completely distributed representations.

Subsymbolism seems to be based on a misconception of the epistemological status of the representations of the model (Verschure, 1990). Knowledge that is put in by the designers, that relates to their domain ontology (the symbolic categorization of the domain consisting of characters and phonetic features), is erroneously interpreted as an emergent property of the model. This provides another example of the seriousness of the FOR problem.

The solution of the mind-brain dilemma that subsymbolic connectionism offers remains a vague promise.

From symbols to dynamics

The analysis of NETtalk showed that subsymbolic connectionism can be seen as a new methodology in a well known theoretical framework: cognitivism. The initial ambition of connectionism to form an alternative paradigm for cognitive science is not fulfilled. It seems useful to reevaluate the role that connectionism can play in cognitive science.

To reassess the ambition of connectionism it is useful to first evaluate the nature of connectionist models. Connectionist models are dynamical structures with a brain-like flavor, but they can also be applied to model other phenomena like the immune system or auto catalysis (e.g. Farmer, 1990). This implies that these models are neutral to any interpretation and cannot by themselves constitute a new paradigm.

In defining an alternative conceptual framework the FOR problem can be taken as a starting point. To understand behavior it is important not to confuse the different perspectives involved. If the design of a system is based on an external domain ontology (from the designer or observer) and its behavior is interpreted as if it were related to the experience of the artefact we are suddenly confronted with the symbol grounding problem. Because, it is not recognized that the representations of the system are founded in this external domain ontology. In this respect the symbol grounding problem can be seen as an artifact of a symbolic approach which ignores the FOR problem.

It is obvious that intelligence is related to knowledge. The point is, however, that this knowledge should from the start on be grounded in the experience of the system and not in that of the designer or observer. Moreover, symbolic descriptions of behavioral regularities can be taken as being part of an observer ontology. But there is no reason to automatically assume that the behavior of the system is produced by internal symbolic processes that mirror these regularities.

Given the above mentioned problems there is no reason to subscribe immediately to the assumptions made by cognitivism. In our own work, which relates to the emerging field of "New AI" (Brooks, 1991), a different set of assumptions is made. The first assumption is that cognition can only be modelled using autonomous agents (see also Brooks, 1991): systems that have realistic sensors and effectors with which they interact with the world. Next, these systems do not assume highly reliable transduction functions that take care of the perception of, for instance, a letter, but they span the whole domain from sensing to acting. This allows the development of representations that are grounded in the experience of the system. The behavior of the system is not separable from its environment. It is the result of the ongoing interaction between the two and not a distinct property of one of these elements (e.g. Ashby, 1960) Furthermore, a different set of assumptions about the world is made: First, the real world is constantly changing, only partially knowable, and only partially predictable. Therefore there cannot be a predefined body of knowledge that approximates the properties of the real world (see also Agre & Chapman, 1988; Suchman, 1987). Second, the world does not consist of a collection of events. The notion event is completely connected to the interaction between a system and the world. This last point will be further dealt with in the discussion section.

While cognitivism assumes that there is a "body of knowledge" to be able to explain behavior and postpones the question of learning (Haugeland, 1985) this proposal takes the opposite strategy. The central theme is how a system can acquire knowledge from its interaction with the world: how does adaptation take place and what are its prerequisites. Moreover, all processes, internal and external, are in principle dynamic. The observed behavior can, however, be described in symbolic terms.

Starting with the assumptions outlined above we have developed a design methodology for autonomous agents, distributed adaptive control (DAC) (Pfeifer & Verschure, 1992; Verschure et al., 1992) which is based on a model for classical conditioning (Verschure & Coolen, 1991). The basic properties of the system are related to a value scheme, which is taken to be defined by the genetic setup of the system (Edelman, 1987). The value scheme defines the properties of the sensors and effectors and some initial sense-act relations (reflexes). The value scheme allows a coarse adaptation to the environment, for instance, when there is a collision to the left turn to the right. The system is also equiped with a more sophisticated sensor: an inverse range finder which represents, in essence, time to contact. The states of this more sophisticated sensor are gradually integrated into the basic reflexes of the system due to the system environment interaction. This integration process, which is based on a Hebbian learning mechanism, will lead to a fine tuned adaptation to the specifics of the environment. In Figure 1 the set up of the control architecture is depicted. The three sensors project their state onto specific neural fields.

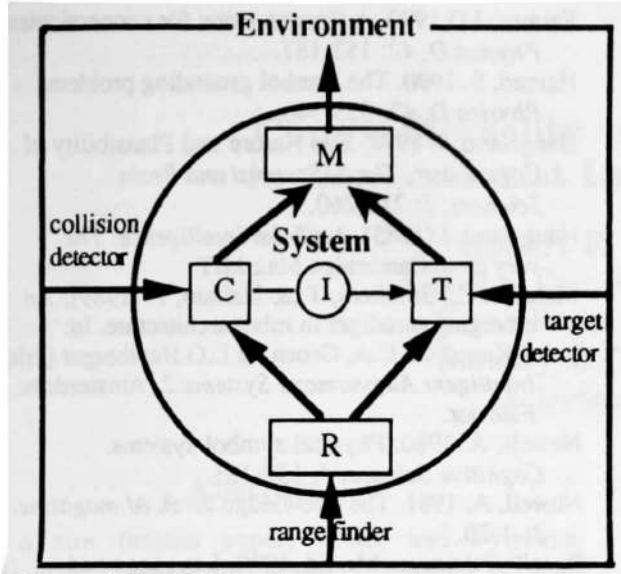

Figure 1: The DAC architecture and its relation to the environment.

Activation of units in the fields that relate to the collision detector (C) and the target detector (T) will automatically trigger an action, avoid or approach respectively. The basic reflexes can be described as: "collision left -> turn right" and symmetric for the other side, and "target left -> turn left" and symmetric for the other side. The default action of the system is to move forward. These actions are represented by motor units in field M. The connections between C, T and M are prewired. These connections implement the reflexes by connecting the related sensing and acting components. Since the change of the connections between C and T and the range finder field (R) is based on a Hebbian learning rule any state in R which occurs congruently with an action will be associated with the activation in C and/or T that triggered this action. Over time specific prototypical states of R will develop that will trigger specific actions. Next to these four fields a special inhibitory unit, I, is defined that regulates the interaction between avoid and approach actions: activation in C will inhibit the output from T.

We showed (Verschure et al., 1992) that a system based on these properties can develop emergent behaviors like wall following in an environment where targets are placed behind holes in walls. This regularity leads the system to associate being parallel to a wall with approach actions. Over time this behavior was generalized to any situation were the system was next to a wall. It follows a wall wiggling along switching between approach and avoidance actions.

This wall following behavior can be described in symbolic terms like a strategy or rule which is based on the representation of a wall and the action to follow it. The properties of the control architecture, however, indicate that such a rule is not present in the system. This behavior will only emerge when a specific regularity is present in the system-environment interaction (e.g. targets behind holes in walls). This indicates that although behavioral regularities might be based on special internal regularities of a system this does not have to be the case. Moreover, this emergent behavior is only present from the point of view of the observer. The system can only act on its immediate sensory states, while wall following behavior is displayed over several time steps consisting of many actions. This behavior that for an observer looks very structured can only be explained when it is decomposed into the actions that constitute it. This decomposition, however, shows that the system is acting like it always would, whether it is following walls or doing something else, that for an observer might look not that well organized.

Discussion

The analysis of subsymbolic connectionism has shown that it is in fact applying new techniques in a well known conceptual framework: cognitivism. Therefore, it does not provide a new perspective on the mind brain dilemma. It was argued that to assess the role connectionism could play in cognitive science it is of importance to find an alternative conceptual scheme in which it can be applied. The reason for this is not to find a justification for doing connectionism, but to address the mind-brain dilemma.This alternative framework can be found in the developing field of "New AI". The contrast between the two approaches now becomes that assumptions of cognitivism, which lead to the symbol grounding problem, become central research issues. One of this issues is, for instance, what is the role and nature of knowledge in adaptive behavior.

In doing this it becomes clear that the issue of emergence should also be viewed from the perspective of FOR. Emergent behavior then relates to an observer who specifies a specific time and or spatial frame in which "interesting" behavior is displayed by the system. This emergent behavior is not a property of the system but of the interaction with the environment. The chunk of action that an observer can call wall following is related to a set of actions that become a connected whole in the frame of reference of the observer. To explain this behavior it should be viewed from the perspective of the system. Which in the case of the presented example means that what is wall following from the observer perspective can only be explained from the system's perspective as a sequence of approach or avoid actions given the immediate sensory and the internal states.

This perspective on behavior gives a different status to notions that are taken for granted in the symbolic paradigm. For instance, the latter assumes that the world consists of objects and events which are somehow mirrored by the internal representations of the system. In our case we see that the notion of event and object is defined from the perspective of the system where an event always

relates to actions. For instance, initially an action can only be triggered by one of the basic reflexes defined by the value scheme. Due to the learning mechanism this can be transferred to range finder states. What will now become a situation in which a specific action will be triggered cannot be predicted but depends on the specifics of the system environment interaction. Only from the perspective of the learning history of the system the notion event can be defined.

An important issue is how this proposal will scale up to the phenomena traditionally studied in cognitive science like reasoning and language. The central question is, however, whether we should see this issue as a conflict between two approaches. From the perspective of FOR we can see that the accounts offered by traditional approaches can be taken as observer characterizations of behavioral regularities. Which would mean that it is possible to describe some parts of behavior, like language, in terms of discrete elements that we call symbols. From the systems perspective linguistic behavior is still behavior built up out of many actions.

The mind brain dilemma can be addressed from the presented perspective. Supposedly conflicting paradigms in fact provide a different perspective on the phenomenon of behavior. With this we can overcome the isolated position of the study of the mind as a special science and focus on the initial ambition behind cognitive science to develop a fruitful interaction between the behavioral and the neurosciences.

Acknowledgements

The research reported in this paper was partly sponsored by grant 21-30269.90 of the Swiss National Science Foundation to Rolf Pfeifer. The author thanks Rolf Pfeifer and Thomas Wehrle for valuable discussions.

References

Agre, P.E.; and Chapman, D. 1987. Pengi: An implementation of a theory of activity. *AAAI-87*, Seattle, WA: 268-272.

Ashby, W. R. 1960. *Design for a brain: The origin of adaptive behavior*. New York: Wiley.

Brooks, R.A. 1991. Intelligence without reason. *IJCAI-91, Proceedings of the twelfth international conference on artificial intelligence, vol 1:* 569-595.

Clancey, W.J. 1992 The frame of reference problem in the design of intelligent machines. In: K.V.Lehn Ed.: *Architectures for intelligence. Proc. 22nd Carnegie Symposium on Cognition.* pp. 357-423, Hillsdale, N.J.: Erlbaum.

Edelman, G.M. 1987. *Neural Darwinism: The theory of neuronal group selection.* New York: Basic Books.

Farmer, J.D.1990. A Rosetta stone for connectionism, *Physica D, 42*: 153-187.

Harnad, S. 1990. The symbol grounding problem. *Physica D, 42*: 335-346.

Haugeland, J. 1978, The Nature and Plausibility of Cognitivism, *The Behavioral and Brain Sciences, 2*: 215-260.

Haugeland, J.(1985). *Artificial Intelligence: The very idea.* Cambridge Ma.: MIT.

Malcolm,C., Smithers, T.,& Hallam, J. (1989). An emerging paradigm in robot architecture. In: T.Kanade, F.C.A. Groen, & L.O.Herzberger (Eds.): *Intelligent Autonomous Systems 2*, Amsterdam: Elsevier.

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

Newell, A. 1981. The knowledge level. *AI magazine, 2*: 1-20.

Piatelli-Palmarini, M., Ed. 1980. *Language and Learning. The debate between Jean Piaget and Noam Chomsky.* Cambridge Ma.: Harvard University Press.

Pylyshyn, Z. 1989. Computing in cognitive science. In: M.I.Posner Ed. *Foundations of cognitive science.* pp. 51-91. Cambridge Ma.: MIT.

Rosenblatt, F.1958 The Perceptron: a probabilistic model for information storage in the brain. *Psychological Review, 65*: 386-408.

Rumelhart, D.E., McClelland, J.L. and the PDP research group. 1986. *Parallel Distributed Processing; explorations in the microstructure of cognition, volume 1: foundations.* Cambridge: MIT press.

Sejnowski, T. & Rosenberg, C. 1986. NETtalk: a Parallel network that learns to read aloud. The Johns Hopkins University Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01.

Sejnowski, T. & Rosenberg, C.1987 Parallel Networks that Learn to Pronounce English Text. *Complex Systems, 1*: 145-168.

Smolensky, P. 1988 On the Proper Treatment of Connectionism.*Behavioral and Brain Sciences, 11*: 1-73.

Suchman, L.A. 1987. *Plans and situated actions.* Cambridge University Press.

Verschure, P.F.M.J.1990 Smolensky's Theory of Mind, *Behavioral and Brain Sciences, 13*: 407.

Verschure, P.F.M.J. 1992 Connectionist Explanation: taking positions in the mind-brain dilemma. *submitted.*

Verschure, P.F.M.J., & Coolen, A.C.C. 1991 Adaptive Fields: Distributed representations of classically conditioned associations. *Network, 2:* 189-206.

Verschure, P.F.M.J, Kröse, B.J.A., & Pfeifer, R. 1992. Distributed Adaptive Control: The self-organization of structured behavior. *Robotics and Autonomous Agents*, In Press.

Winograd, T., & Flores, F. (1986). *Understanding Computers and Cognition: A new foundation for design.* Reading Ma.: Addison Wesley.