Understanding and Self-Organization

What can the speaking lion tell us?

Erich Prem

Austrian Research Institute for Artificial Intelligence

Schottengasse 3, A-1010 Vienna, Austria

ERICH@AI.UNIVIE.AC.AT

Abstract

The current rebirth of self-organizing systems in several distinct domains of research poses new epistemic questions. Self-organizing systems have a tendency to not only behave in an unpredictable way, they are also extremely difficult to analyse. In this paper we discuss three problems with neural networks that are important for self-organization in general. They are related to the proper design of a self-organizing system, to the role of the system engineer, and to the proper explanation of system behaviour. We shall try to present a generally applicable solutions, which is based on a "symbol grounding" neural network architecture. We then discuss the relation of this approach to the measurement problem in physics and point out similarities to existing positions in philosophy. However, it should be noted that our "solution" of the explanation problem may be judged as being a very sceptic one.

Understanding and Self-Organization

What can the speaking lion tell us?

Erich Prem
Austrian Research Institute for Artificial Intelligence[1]
Schottengasse 3, A-1010 Vienna, Austria

ERICH@AI.UNIVIE.AC.AT

1 Introduction

1.1 Understanding and self-organization

Scientists not only search for solutions to problems, they also seek to construct a solid basis upon which their results can be justified and explained. That such an absolutely secure ground of science is impossible to be found within empirical fields is one of the truisms of our time. The recent rebirth of self-organizing systems in many different domains confronts scientists with new epistemic problems. Self-organizing systems have a tendency to not only behave in an unpredictable way, they are also extremely difficult to analyse.

Both—rebirth and the epistemic problems—are manifest in computer science through the discussions that center around artificial neural networks or, as it is called, connectionism. Such "newly" developed techniques are usually accompanied by theoretical arguments around their usefulness and drawbacks. In the case of self-organizing neural networks there has been much discussion about the virtues and possible advantages of emergent properties. In this paper we want to put our finger on three problems with neural networks that are important for self-organization in general. These three problems are related to the proper design of a self-organizing system, to the role of the system engineer, and to the proper explanation of system behaviour. All three problems are just different aspects of one central problem: Only that can be designed which has been understood and only that can be understood which has been designed.

This paper addresses questions concerning understanding the responses of a self-organizing system. We shall try to present a generally applicable solution, one that will also be implementable for neural networks. It is based on a "symbol grounding" architecture which has been originally developed by Georg Dorffner [6]. The increasing importance of autonomous systems which are put in a physical environment (cf. [2]) leads us to assume that these issues will be even more important in the future. However, our approach will also be a sceptic solution of the problem. In the light of the extremely wide applicability of the principle of self-organization, we admit that this paper barely can scratch the surface of epistemic questions dealing with emergence, design, measurement, indeterminacy of translation, semantics, and other problems related to self-organizing systems.

[1] The Austrian Research Institute for Artificial Intelligence is sponsored by the Austrian Federal Ministry for Science and Research.

1.2 Terminology

In this paper we distinguish between physically and non-physically (informationally) self-organizing systems. They differ with respect to their epistemic qualities. A *physically self-organizing system* is open to flow of energy or material. Its behaviour is fully predictable at a low influx of energy by means of laws which describe the behaviour of the system's atomisms ("equilibrium state") [9]. At a higher influx of energy, through the interaction of elements, the behaviour of the system is not fully accounted for by initial and boundary conditions [10]. To successfully predict it, one needs to develop a new way of describing the system. In these descriptions sets of atomic states map on a state in the new descriptive frame (the macro description). This "change of view" means the development of a new observable of the system, which is important for successfully predicting it. We say that a new phenomenon has emerged.

In an *informationally self-organizing system* no physical emergence takes place, since the system is not necessarily open to flow of energy. Taking artificial neural networks as an example (which we will do throughout the paper), such systems share some properties with their physical pendants. This correspondence consists in the fact that (i) different levels of descriptions are necessary to explain the system and (ii) the new observables appear through the interaction of many elements which can be assumed to be atomic. As in the former case the new observables are mappings from the "atomic" state space to a macro description of the system. The main difference to physical self-organization consists in the fact that when simulated on a conventional computer architecture, the number of possible observables of the system is fixed (cf. section 2.1).

It has already been pointed out by von Bertalanffy as soon as in 1950 that the notion of an open, self-organizing system which shows emergent phenomena is related to biological systems [1]. A well-known socio-biological example of self-organization happens in large populations of nest building insects, e.g. termites [9, 10]. A single termite can be described as a simple autonomous agent who reacts upon a specific scent so that material is deposited immediately. If many insects interact through similar deposit of building and odor emitting material, this simple mechanism results in the construction of an emergent, seemingly planful behaviour—the construction of a termite dome. Investigating the system by only considering individual termites cannot reveal this emergent phemonenon. The adequate description levels of this behaviour are the emergent phenomenon and the interaction of the individuals. It is important to realize that emergence only appears because of many interacting termites. Macro-desriptions like "The insects are constructing and arch." are important in describing and predicting the system of termites, whereas the micro-descriptions of single termites do not possess this predictive value (for a detailled discussion see [10, p.10]).

2 System Design Problems

2.1 Conventional systems, rule-based systems

A conventional system can be designed to behave in a well-defined way, because the internal dynamics of the system are highly constrained. No self-steering interaction between the programmed entities takes place, therefore there is no emergent behavior of the system (for a more mathematical formulation see e.g. [11, pp.31–39]). A mechanical example for such a system is the motor of a car, where the parts of the engine constrain the turbulances of exploding gas to guide the forces in order to operate the wheeles. Another example is a rule-based system in AI where the system designer tries to capture the mapping of a problem space to a space of solutions by means of a partition of the former. This can be done, because the computer program ensures that the rules are mapped on the physical state space of the computer system in a way that the system is kept at a low energy level. Therefore, no physically relevant interactions between the rules take place and no emergent properties of the system arise. Thus, the rules specify the physical state transitions and the transitions in the problem-solving state space.

The conventional approach to designing a solution with such systems consists in classifying "situations", i.e. input data, and connecting them to "actions". In immediately grounded connectionist systems or autonomous systems the classification refers to the data which arrives from the sensoric input. Such an input can be regarded as a measurement of the physical environment, like e.g. a video-camera recording. The system designer can classify the data according to what is happening at her own sensory "device" by means of the concepts which he has aquired. Being in possession of a relevant conceptual framework enables the designer to develop the rules and to *understand* what the rules do. When later explaining an action of the system, the programmer can give "causal" explanations in the sense that a specific set of data items "means" this or that and consequently leads to the action in question (cf. [19]). Explaining and understanding the system therefore crucially depend on the fact that the system actually *uses* the same conceptual framework as we do.

For several reasons (pointed out e.g. by [12]) this implies severe limitations, even in the case of a machine-learning system. The influence of the designer restricts the system not only in the way that the mapping from situations to actions is being designed. The more profound problem is the pregiven conceptual bsis of the designer. This problem has originally been one of the main arguments for artificially neural systems and is now extended through the importance of autonomously acting systems in physical environments. We shall call it *the designer problem.*

2.2 Artificial neural networks

As opposed to conventional systems, in artificial neural networks the interaction of many simple computing units results in an emergent phenomenon, in this case a useful computation. The overall behaviour of the system is a consequence of the interaction of units, so that the appropriate level of describing and explaining the system is the level at which the behaviour emerges. With respect to section 1.2

it must be noted that the emergent character of neural networks is a function of the system taken into account. If one views the computer together with the neural network simulator as the system in question, than no *new* states emerge. It is only with respect to the designed entities that we can talk of "emergent properties".

A great part of the neural network literature only deals with the appropriate design of these entities. The question which confronts the designer is: "How can a network be designed, so that it shows the desired emergent behaviour". This question is dual to: "What sort of behaviour will emerge from that system?" and can therefore be labelled *the inverse emergence problem.* This problem can be translated to our termite example as: Which behaviour should a single termite have, so that instead of the dome the insects build a bridge?

In neural networks there cerainly exists an important influence in designing the architecture (number of units, learning parameters and rules, etc.). But it can be said that this is not comparable to the aforementioned strong influence of the designer. The hope at least is that the *relevant* system behviour only emerges from the interaction of many units. Therefore, the decisions made through the network do not directly depend on the conceptual framework of the net designer, especially if the input to the system is "immediately grounded" in the physical environment [3] (and not of the symbolic sort as e.g. in NETtalk [22]). The designer's strategy consists in enabling the system to self-organize to a satisfactory degree.

One of the well-known drawbacks of this technique is the *explanation problem.* (See e.g. [13] or [5].) It is now very difficult to explain the system behaviour within our own conceptual framework, i.e. with words containing personal experiences, sensual impressions etc. and not purely mathematical terms. We shall refer to this way of explaining a network as "understanding". One of the procedures that has been suggested and often used for this is to investigate the hidden units' representation. This can be done by activating input units for which a predication e.g. "there-is-a-dog" can be given. A statistical relation between this input and an activation of a hidden unit can lead to assertions of the kind: "Hidden unit 7 is activated for dogs." However, this approach often does not work, because the hidden units simply do not represent something that would be easily expressible in natural language (within our conceptual framework). The more complicated the network architecture the more implausible this approach becomes. For systems with a manifold of inputs (e.g. many and/or complex sensors with many states that receive data directly from the physical environment) it becomes impossible to trace statistical relations between the parts of a self-organizing system and its emergent properties.

An approach to this problem is not just an epistemic need of the philosophically eager scientist, but a technical imperative. Without even a rudimentary approach to predicting the system, to evaluate it there can be no confidence in its results. It is also for reasons of maintanance that the explanation problem must be addressed. Note that even if we would be satisfied with a merely mathematical solution to the problem, a physically self-organizing system simply develops *new* observables. Therefore, understanding of such systems without the endeavour of applying difficult techniques of analysis is not feasible. The "explanation"-problem thus is a principal limitation of any self-organizing system. It coexists with a dualtity between design-

based limitations and understandability. In the next section we will show that there are possibilities to overcome these problems in a systematic way.

3 What do we need?

3.1 Law-like rules and concepts

The rules in a conventional system not only support its design but also guarantee our understanding of the system. Therefore, if we want to explain a self-organizing system, we are searching for rule-like descriptions of the system behaviour. Since we hope that we did not design these rules (because of the designer problem), we must find another way to guarantee that the rule by which we describe the system actually is followed by the system.

Consequently, our first step will consist in a technique of attributing rule-like descriptions to the self-organizing system. The basis of this description must involve an understandable description of the situation in which the system finds itself. To understand what these situations for the system are, we need a possibility to make the system use one of our conecpts without actually using it for decisions. This paradox requirement can be overcome by forcing the system to translate between the "concepts" it "uses" and the ones which we (the system observers) easily understand. What we need is a mapping of the system's states that result from self-organization to the concepts which we understand. The principle of the system (the COSYC-architecture) can be gathered from figure 1.

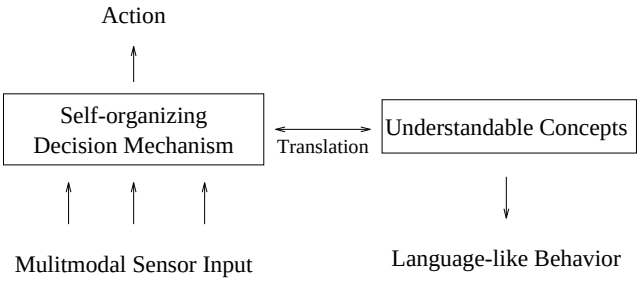

Figure 1: Systematic of the COSYC-Architecture.

This architecture is supposed to achieve the following. It is able to "mimic" the designer's response to a given situation. When it sees a flower, it responds with "flower!". In order to make the system answer with concepts which are understandable there will have to be a training process through which the system builds its translation function. (See figure 2.) We do *not* assume to know this translation function. We only make the system respond in a plausible way to the situations in which it finds itself. Moreover, by assuming that the system is self-organizing and dealing with immediately grounded data, we assume that no analysis of the translation function is possible. Strictly speaking, we do not even have to know the state space of the decision making, acting system as long as the words are consistent with what we expect them to be.

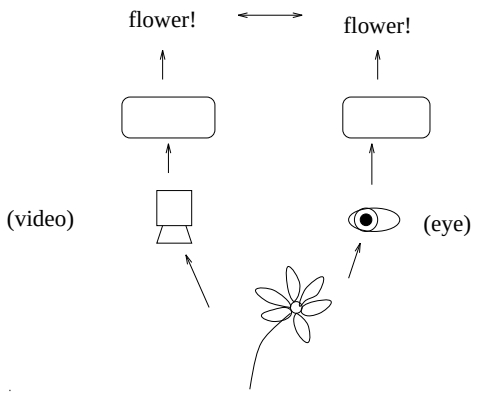

Figure 2: Supervising the system.

3.2 Automated generation of descriptions

Such a mapping does not necessarily have to be constructed by the system. In our termite example scientists have found an explanation of the system through careful investigation. The termites, however, form a relatively simple system because of the low dynamics and the easy to observe interacting units. Note that this approach to understanding the system is not limited to neural networks, but can be regarded as generally applicable to self-organizing systems. All we need is a consistent language-like behaviour that is achieved through a translation mechanism.

The point which we would like to make here is that such a system can be explained by using *its own* descriptions of the situations. We can construct predictive rules based on the system's own situation descriptions. When it senses a "dog", it reacts in a specific way. These attribution of situations shall form the basis of our understanding the self-organizing mechanism. But can we be sure that this is a useful description of what the system does?

3.3 Objections

The difference between (i) investigating the termites and (ii) the mimicry of a conventional conceptual frame seems to be a very profound one. In the case of the termites (i) the concepts which explain the behaviour *really* describe what happens. In (ii) the concepts are just classifications of the system's states that do not *really* describe what is going on in the system.

What is obviously meant with such an objection is that in (i) we have grasped "nature" with our concepts and describe what is causally happening and could not otherwise be explained satisfactory. In (ii) the system's translation is just a mapping on a set of words that could well be *wrong*, i.e. not what the system *really* does. The system could, e.g. if we taught it wrong usages of words, lie to us and truly do something else. Can this be a serious doubt?

Of course, we must assume that the system has learned to use the words it learned correctly. Its description of situations should be comparable with the way we classify them and the actions which the system undertakes. This question, we propose, must be formulated in the following way: Which fact can guarantee that

the system *means* (through parroting) what I mean with this same word? The answer is that if the system is truly complex (i.e. not simply analyzable), then the only fact that does guarantee this can be found in the system's behaviour. If it behaves as if it would use my terms correctly, we not only can but must say that it does.

This problem has been intensively discussed within a philosophical debate around the later works of Ludwig Wittgenstein [23]. In Kripke's interpretation [8], the main achievement of Wittgenstein was to have shown that there is no other guarantee for what somebody *means* with a word but some outer criterion to be found in his or her behaviour. According to this position it is useless to search for facts within the speaker which could guarantee what is meant with a word. And it is also impossible to ensure that the word will always be used correctly in the future.

With respect to our problem here this position holds that as long as the system correctly uses my *words* we have to say that the system means the same as we do. In this sense do we *understand* the situations in which the system is. Additionally, as long as the system behaves according to rules which are based on these situations we have properly designed the state space which is necessary to describe the system. We can then say to understand the system's actions in terms of our rules. For the sake of clarity, we shall reformulate this position in section 5. Let us first present a neural network architecture which is capable of implementing the necessary translation function for the case of an informationally self-organizing system.

4 The COSYC-Architecure

The neural network model of Dorffner [6] can be expanded to cope with multi-modal sensory input. A detailed technical description will be given elsewhere [20]. For the purpose of this paper it suffices to know that the categorization of input data happens without the influence of the designer. (For a theoretical argument related to this specific architecture cf. [18].)

Figure 3 shows that the model can have several sensoric inputs, which are supposed to be immediately grounded, i.e. directly connected to physical sensors. For each input channel unsupervised (informational) self-organization within a layer of connectionist units (called a *C-layer* [7]) takes place. A second unsupervised categorization is where *concept formation* happens. Again, layers of connectionist units ensure that the system is able to categorize the manifoldness of inputs without being taught which concepts to form. (Nevertheless, this process *can* be supported by the designer.) Another set of layers (*SY-layers*) learns to map these C-layer representations onto single, discrete states. These discrete states can be interpreted as *symbols* or *words*. Thus the system learns (i) to categorize the input data and eventually use this categorization for acting in the world and (ii) translates its conceptualization to symbolic states which are interpretable as words (if taught and coded appropriately).

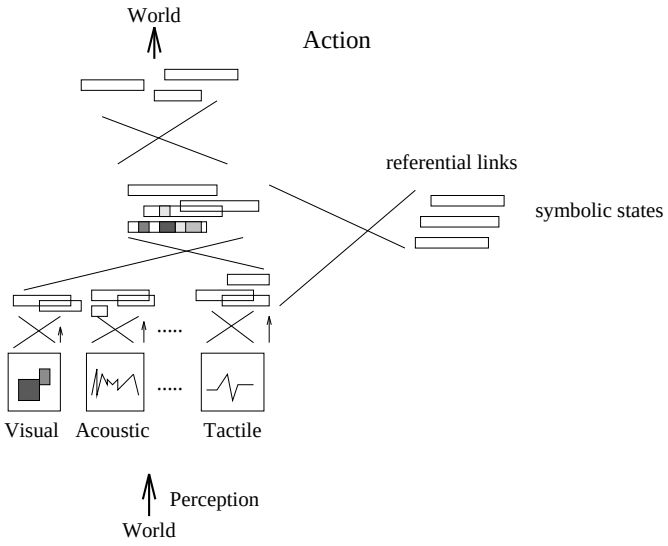

Figure 3: Understanding what a net does.

5 Understanding and Measurement

One possible interpretation of our approach is to compare it with a measuring device and a physical process (figure 4). What we have introduced above is a system that responds with "flower!" to describe its own state. Now imagine a physical system and a meter, e.g. a thermometer that delivers numbers. We usually assume to understand the physical system if we can use the measurement in a rule which predicts another aspect of the physical system (of course, another measurement). If we fail to make a correct prediction, we can either blame our rule, blame the meter or blame the explanation of the meter. Note that the system being measured can have (and presumably always has) additional states, which are not captured by the meter. But unless this does not result in wrong predictions, we are not only satisfied with our description, we say that we have *understood* the system. We can even say that physical events correspond to changes of state which are only specified by the evaluation of observables on states [21]. This means—as long as our predictions are correct and we have no need to introduce or to assume the existence of additional states—that all there is about physical systems (i.e. all there *is*) is that for which an observable exists.

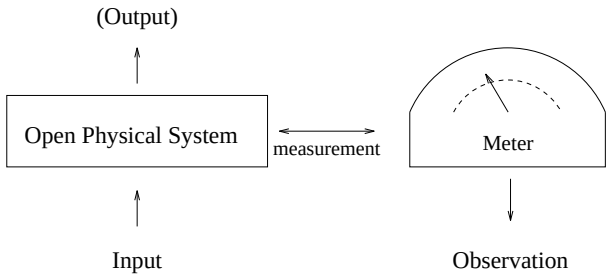

Figure 4: A physical measurement.

Compared to our neural network example it now becomes clear in which sense we understand what the network does: As long as its own descriptions of the situations in which the system finds itself are consistent with what we expect these descriptions to be like, there is no reason whatsoever to say that we do not *really* understand what is going on in the network. In the COSYC-architecture all there is to explain the neural network (the left part of the system in figure 3 with inputs, categorization and action) is what COSYC's translation function (the right part in figure 3) tells us about it. This forms the *relevant set of states* of the connectionist system.

Assume now that we have a physically self-organizing system which can develop new interesting states. Imagine further that, since we cannot predict what the relevant states will be, we invent a procedure to build new measuring devices which capture the newly developed observable (cf. [15, pp.105–108], [21, p.90]). This new meter maps the states of the system to numbers or symbols. In a way, this would be a better way of explaining the system, since we now have a direct meter for interesting phenomena. However, it is not clear how we can understand the new measurement. Note, that we do not know what the meter is actually measuring, we can only see its output, which consists in "meaningless" numbers or symbols. Trying to find out what an unknown meter measures results in extensive experiments where one compares the results of the measurement with well-known situations measured by other devices or sensory organs.

It has been pointed out by Pattee [16, 17] that all this are typical features of measuring devices, since we do not know the internal dynamic constraints of the meter. Nevertheless, we can understand new measurements, because someone can explain them to us. I can, for example, be told that a bat can use its subsonic sensors to detect objects in the night. We understand the bat's measuring device through a comparison with our eyes and think that subsonic "hearing" is in a way like "seeing".

This sort of translation is exactly what happens in the proposed architecure. The system is constructing a meter and automatically "explaining" it by using words the way we would do. In the case of the simulated neural network the meter construction is not a principal problem, because the state space of the computer is fixed. Therefore, we can theoretically construct a meter of the system's state space onto any desired scale. This cannot be done easily in the case of a physically self-organizing system [4]. Although one such self-building measuring device has been described and built by Gordon Pask [14], the applicability of the proposed method in physically self-organizing systems depends on physical methods for the construction of the translation function from "internal states" to symbols.

6 Understanding and rules

There is a second level at which the question of understanding what a system does can be answerded. In what we said above, we tried to reduce the problem to the question of properly *discovering* the state space of the system, i.e. the situations which are important to predict its beahviour. This was supposed to help us to discover the predicting *rules*. However, understanding can also mean to recognize the purposiveness of a specific behaviour. In this case the *rules* which the system

follows are themselves proven to be useful, meaningful, or goal-achieving. It has been previously pointed out by the author [19] that neural networks in general, but especially together with their appearance in autonomous systems or "artificial life" models tend to be explained in a teleological i.e. goal-oriented way.

In the case of our self-organizing system, however, we cannot be sure about the goals which the system will try to achieve. Because it is a technical system, we assume that it has been designed to fulfill one of our goals. This is why it may be a useful technical system.

Would the proposed mimicry be enough if the goals of the system would be different? Wittgenstein would deny this, since we can only understand what a system says and does, because of our socially shared way of living, our common "Lebensform." *If a lion could speak, we could not understand him.* [23, p.568]

7 Acknowledgements

The author wishes to thank Robert Trappl for having made this research possible as well as Georg Dorffner and Felix Annerl for continuous discussions.

References

[1] Bertalanffy L.von: The Theory of Open Systems in Physics and Biology, *Science*, Vol.111 (1950), pp.23-9, 1950.

[2] Braitenberg V.: *Vehicles. Experiments in Synthetic Psychology*, MIT Press, Cambridge, MA, 1984.

[3] Brooks R.A.: Elephants Don't Play Chess, in Maes P.(ed.), *Designing Autonomous Agents*, MIT Press, Cambridge, MA, Bradford Books, pp. 3-16, 1990.

[4] Cariani P.: Implications from Structural Evolution: Semantic Adaptation, in Caudill M.(ed.), *Proceedings of the International Joint Conference on Neural Networks (Winter Meeting), Washington D.C.*, Lawrence Erlbaum, Hillsdale, NJ, pp.47-50, 1990.

[5] Diederich J.: Explanation and Neural Computation, GMD, Nr.458, 1990.

[6] Dorffner G.: On Redefining Symbols and Reuniting Connectionism with Cognitively Plausible Symbol Manipulations, Oesterreichisches Forschungsinstitut fuer Artificial Intelligence, Wien, TR-92-13, 1992.

[7] Dorffner G.: A Step Toward Sub-Symbolic Language Models without Linguistic Representations, in Reilly R., Sharkey N. (eds.), *Connectionist Approaches to Language Processing*, Vol.I, Lawrence Erlbaum, Hove, 1992.

[8] Kripke S.: *Wittgenstein on Rules and Private Language*, Harvard University Press, Cambridge, MA, 1982.

[9] Kugler P.N., Turvey M.T.: *Information natural law and the self-assembly of rhythmic movement*, Lawrence Erlbaum, Hillsdale, NJ, 1987.

[10] Kugler P.N., Turvey M.T.: Self-Organization, Flow Fields, and Information, in Kugler P.N., Turvey M.T. (eds.), *Selforganization in biological work spaces*, North-Holland, Amsterdam, pp.1-33, 1989.

[11] Lange O.: *Wholes and Parts. A General Theory of System Behaviour.*, Pergamon, PWN-Polish Scientific Publishers, Oxford (Warszawa 1962), 1965.

[12] Lischka C.: Ueber die Blindheit des Wissensingenieurs, die Geworfenheit kognitiver Systeme und anderes ..., *KI 4, p.15-19*, 1987.

[13] Partridge D.: What's Wrong With Neural Architectures, Computer Science Dept., Univ. of Exeter, Research Report 142, 1987.

[14] Pask G.: Physical Analogues to the Growth of a Concept, *Mechanization of Thought Processes: Proc.of a Symposium*, National Physical Laboratories, November 1958, HMSO, London, 1958.

[15] Pask G.: *An Approach to Cybernetics*, Hutchinson, London, 1961.

[16] Pattee H.H.: Universal Principles of Measurement and Language Functions in Evolving Systems, in Casti J.L., Karlqvist A.(eds.), *Complexity, Language and Life: Mathematical Approaches*, Springer, Berlin, 1986.

[17] Pattee H.H.: The Measurement Problem in Physics, Computation, and Brain Theories, in Carvallo M.E.(ed.), *Nature, Cognition and System II - Current Systems-Scientific Research on Natural and Cognitive Systems, Vol.2: On Complementarity and Beyond*, Kluwer, Dordrecht, p.197-192, 1992.

[18] Perlis D.: How Can a Program Mean?, in *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, Morgan Kaufmann, Los Altos, CA, pp.163-166, 1987.

[19] Prem E.: Aspects of Rules and Connectionism, in Trappl R.(ed.), *Cybernetics and Systems '92*, World Scientific Publishing, Singapore, pp.1343-1350, 1992.

[20] Prem E.: COSYC: An Architecture for Connectionist Symbolic Cognition, TR Austrian Research Institute for Artificial Intelligence, to appear Spring 1993.

[21] Rosen R.: *Fundamentals of Measurement and Representation of Natural Systems*, North-Holland, New York, 1978.

[22] Sejnowski T.J., Rosenberg C.R.: Parallel Networks that Learn to Pronounce English Text, *Complex Systems*, 1 (1987), 145-168, 1987.

[23] Wittgenstein L.: *Philosophische Untersuchungen*, Suhrkamp, Frankfurt/Main, Werkausgabe Bd.1, 1986, 1953.