Philosophica 57 (1996, 1) pp. 131-172

THE DEVELOPMENT OF SCIENTIFIC CONCEPTS AND THEIR EMBODIMENT IN THE REPRESENTATIONAL ACTIVITIES OF COGNITIVE SYSTEMS. NEURAL REPRESENTATION SPACES, THEORY SPACES, AND PARADIGMATIC SHIFTS.

Markus F. Peschl

ABSTRACT

The goal of this paper is to show, how scientific concepts and processes are embedded into representational activities of cognitive systems. The most obvious fact is that science is conducted by cognitive systems whose neural systems enable them to represent and successfully interact with the world. In the course of this paper it will turn out that concepts from computational neuroscience and artificial life provide interesting insights into the problem of knowledge representation and, as a consequence, into the understanding of science and of what is referred to as the context of discovery. It will be shown how the dynamics of theories and scientific concepts can be described by the dynamics going on in the neural representation space (activation- and weight space). Furthermore, concepts from genetic algorithms and their combination with artificial neural networks could give a cognitively founded explanation for paradigmatic shifts.

1. *Introduction*

It seems to be an accepted fact that *science* is the result of *cognitive activities*. The production of theories, conducting experiments, "having new ideas", developing and "inventing" new perspectives on a well known phenomenon, verifying a hypothesis, deducing implications, applying certain methods, neglecting certain results, etc. are not some abstract and detached "scientific processes"; they are deeply cognitive capacities which are closely tied to the activities of cognitive systems, their representational capabilities, their interactions with the environment,

their interactions with each other, and their ability to produce artifacts, to generate and use language, and to manipulate the environmental dynamics. Recent publications in philosophy of science (e.g., Brakel 1994; Churchland 1989, 1991, 1995; Giere 1992, 1994) represent a first step toward taking seriously this connection between cognitive and scientific processes.

Nevertheless, many approaches in philosophy of science (e.g., Logical Positivism, Popper's philosophy of science, etc.) did not include (or even explicitly exclude) cognitive processes and the activities of a cognitive system into their theories and investigations. Their focus is on the "context of justification"; i.e., these approaches are using methods and tools from logic in order to deduce theories, they are trying to verify or falsify already existing theories, etc. In any case, the really interesting part in the scientific process – "discovering", constructing, or developing a new theory – is more or less neglected. The reasons for this are manifold: the process of discovery is said to be a more or less irrational process and, thus, cannot be included in a theory about scientific knowledge; the psychological, neuroscientific, or cognitive processes being involved in the context of developing new theories are said to be too complex and, thus, cannot be understood. In other words, the "context of discovery" is still somewhat shrouded in mystery for most traditional philosophers of science. They prefer to stay in their logical analyses, in their abstract and detached description of scientific processes (i.e., scientific theories are abstract and objective descriptions of the environmental dynamics or complex systems of logical sentences), despite the fact that the *cognitive* process of discovering or constructing new knowledge and theories is the really interesting and fascinating activity in science. However, as mentioned above, there is an increasing interest in these not so formal processes of science. Psychological as well as social studies of science are only a first step. The foundation of all these processes are cognitive activities of cognitive systems. The focus of interests has to shift from formal, social, or psychological investigations and descriptions of science to the "roots" of the scientific process: the investigation of cognitive systems, their ability to represent the world and to interact with it, and to construct new knowledge and theories. From this basis social, cultural, and scientific structures, and the dynamics of theories will appear in another light.

The situation of traditional philosophy of science can be compared

to the development in cognitive science and artificial intelligence: for a long time there was the hope that intelligent behaviour could be understood, generated, and simulated by applying logic, formal systems, symbol manipulation, propositional approaches (e.g., Newell and Simon 1976; Newell 1980, Fodor 1980, 1989; Winston 1992, and many others). The concepts of a formal representation of the environment and logical operations on these representations were in the foreground. The problem of learning and acquiring new knowledge was approached in the context of this formal, (deductive) pseudo-inductive, and logical framework ("machine learning") – the results were rather disappointing and did not at all match the observations of human intelligent behaviour or learning behaviour. Recent developments in the fields of cognitive science (Posner 1989; Green 1996) and artificial life (e.g., Langton 1989, 1994, 1995, and many others) have revealed, however, that so-called lower cognitive processes, such as primary and sensorimotor processing, neural processing on any level of complexity, neural learning mechanisms. sensory systems, etc. are at least as important for generating, understanding, and simulating so-called higher cognitive activities.

With the advent of *neural computation* (e.g., Arbib 1995; Anderson 1988; Anderson et al. 1991; Churchland et al. 1989, 1990, 1992; Hertz et al. 1991; Rumelhart et al. 1986; Schwartz 1990; Sejnowski et al. 1988; Varela et al. 1991, and many others) artificial life, dynamic systems (Gelder 1995; Port 1995), etc. an epistemological as well as methodological – almost paradigmatic – shift has occurred in the cognitive science community. A new understanding of *knowledge representation* and *cognition* is the result of this process, which is still developing. The emphasis is on a more dynamic and not so rigid and formal view of cognition, of knowledge, and processing. It is based on the assumption that cognitive activities have their foundation in the neural and biological substratum and dynamics (and not in logical formulas). In this paper philosophy of science (and its traditional understanding of scientific processes) will be confronted with these new perspectives, methods and theories about cognition, knowledge representation, and cognitive processes. In contrast to artificial intelligence, which has been influenced and which is based on principles stemming from philosophy of science and logic, I am trying to show that recent developments in cognitive science, neural computation, and artificial life have a crucial impact on epistemological concepts, such as knowledge representation. As a consequence,

they will change our understanding of the process of science, of (scientific) theories, and of developing and constructing new theories.

The goal of this paper is to sketch these rather new concepts of (neural) representation and to make explicit their implications for philosophy of science and epistemology. (Scientific) Theories turn out to be only one form of representation which is embedded in the more general and flexible neural representation system of a cognitive system. Consequently, the development of scientific theories follows a similar dynamics as neural representations. This view has important implications for an alternative understanding of developing and constructing new theories. Construction processes in conceptual representation spaces (being neurally realized as activation and weight space) turn out to be more important than formal systems, complex deductions, or accurate mappings of the environment.

2. *Cognitive Science and (Philosophy of) Science*

One of the goals of this paper is to show that there exists a *close relationship* between philosophy of science and recent developments in cognitive science (see section 1). There is the obvious connection that science is done by cognitive systems; hence, cognitive science could perhaps contribute its models and theories to the investigation of the process of science. On a more fundamental level one can find at least two links which connect these two fields: an *epistemological* and a *methodological* link.

2.1 Epistemological Link

It seems that both cognitive systems and science have a rather similar goal: the *representation of the world*. Both are interested in an adequate representation, description, explanation, prediction, and manipulation of the environmental dynamics.

Any cognitive system is a living system. In order to survive, it is necessary to maintain a state of homeostasis (Maturana 1980). From an abstract and system theoretic perspective, the process of life can be characterized as a sequence of transient equilibria – energy from the environment is necessary to maintain these equilibria. In order to achieve this goal the cognitive system has to have some *knowledge* about its

environment so that it can look for sources of energy and avoid inadequate or dangerous environmental states or situations. In other words, it is *necessary to represent the environment* in some way, in order to survive in this environment. Phylogenetic and ontogenetic processes have brought about a wide range of more or less complex representation systems. The *nervous system* has turned out to be an extremely successful, flexible, and complex representational substratum which can be found in most complex organisms.

As shown in Peschl (1994a) (see also Peschl 1993, 1994b), a (recurrent) neural representation system can be understood as a *transformation system* which transforms the sensory input with respect to the current internal state into behavioral output[1]. The current input selects a (n internal representational) state out of the space of possible successor states. This space is predetermined by the current internal state and the neural architecture. The important point to note is that it is *not* the goal of the neural representation system to map the environment as accurately as possible (to representational states), but to *generate functionally fitting behaviour* (see also Glasersfeld 1995; Roth 1994). Hence, representation in neural systems is a *strategy to survive* by externalizing behaviour, rather than by depicting the environment.

The "goal" of evolutionary as well as ontogenetic processes (i.e., learning, development, etc.) is to generate and provide these representational structures (in the form of a specific neural architecture), which are capable of generating adequate behaviour (being necessary for the organism's survival). In other words, the representational system aims at *manipulating* the organism's internal and external environmental dynamics in order to achieve desired[2] states.

From a constructivist perspective (e.g., Foerster 1973; Glasersfeld 1984, 1995; Steier 1991; Varela 1991; Watzlawick 1984, and many others) scientific theories have similar goals: they are not understood as "objective descriptions" of the environment (which is impossible anyway from an epistemological and constructivist perspective), but as strategies for successfully coping with the environment. In other words, the aspect of manipulating and predicting the environmental dynamics are the central features of scientific theories[3]. Think, for instance, of modern particle physics: an incredible effort is pushed into the development of huge particle accelerators in order to manipulate the environmental dynamics in such a way that a predicted effect can be "seen". Another example is

modern genetics: the goal is to manipulate the genetic material in a desired way in order to produce a certain protein structure or organism. Of course, some kind of knowledge has to be developed about the genetic structure, its biochemistry, etc. However, it is only so far of interest as it provides *tools* or strategies for successfully manipulating the genetic material or predicting its effects. At best, the descriptive or explanatory aspect of scientific theories is captured in "if-then" rules in the following sense: if the environment is in a certain condition (either through its own dynamics or by penetration in an experiment), then the effect x is very likely.

Like in cognitive systems, the goal of scientific theories is to (a) find out, (b) describe, (c) predict, and (d) make use of *functional relationships* and *regularities* which are found in or constructed from the environment. What is referred to as "objective scientific description or explanation" is only a by-product which has its status as "true knowledge" only because of its success in predicting and manipulating the environment in a superior manner. From an epistemological and constructivist perspective the difference between so-called scientific theories and so-called common sense knowledge seems to get blurred; there seem to remain only quantitative differences concerning the generality, accuracy in predicting the environmental dynamics, consistency, elegance, etc. Both are structures which can be used to generate behaviour functionally fitting into the environment. As history of science as well as our own experience show, there is no way to tell that there do not exist other knowledge, representational, or theoretic structures which are capable of generating the same or even "more fitting" behaviour.

Hence, the epistemological link between scientific theories and neurally represented knowledge is closer, than most philosophers of science (want to) assume. From the perspective that scientific theories are also a product of cognitive and neural representational processes it is no wonder that this artificial gap between scientific and common sense concepts and knowledge begins to collapse the more we understand cognitive systems and their (neurally based) representational capabilities.

2.2 Methodological Link

The second link between cognitive science and (philosophy of) science is based on the epistemological assumptions from section 2.1 and concerns

methodological questions. As neural processes are the foundation of any representational process, approaches, in which neural systems, evolutionary processes, living systems, etc. are *simulated*, are an important link for understanding epistemological issues. What makes these approaches interesting is the fact that they offer interesting methods for explaining and understanding the dynamics of neural or evolutionary systems. Their theories and methods do not only provide rich details in the generated data, but also provide a *conceptual* framework and models for cognitive processes on various levels of complexity. As will be shown, these methods have also crucial implications for the epistemological realm: they give us new insights in the representational dynamics and the representational relationship to the environment. And this is the point, where it becomes interesting for our original problem of trying to understand the process of science from a cognitive perspective. (Explanatory and simulation) Methods from the fields of neural computation/connectionism, genetic algorithms (e.g., Belew 1990; Goldberg 1989; Holland 1975; Mitchel 1994, and may others), artificial life, etc. offer a *conceptual* level of explanation (e.g., representational state spaces) which is of interest for both the scientific and the cognitive domain.

In the course of this paper it will turn out that these concepts lead to an alternative understanding of scientific theories and how they are embedded and generated by the neural representation system. The claim is that the method of simulating cognitive systems provides conceptual tools which are not only relevant for the understanding of cognitive systems and epistemological questions, but also for philosophy of science.

It has hopefully become clear by now that cognitive science and the study of cognitive systems (as computational systems) in general can offer new perspectives and insights into understanding the process of science. Science is brought back to its "roots"; i.e., science is not some abstract and detached process, but it is conducted by cognitive systems and it is based on the *representational capabilities* and *dynamics* of one or a group of cognitive systems. The goal of this paper is to sketch the foundations and implications of such a "radically cognitive account of science" in the light of recent developments in cognitive science (e.g., neural computation, genetic algorithms, artificial life, etc.).

3. *Scientific Theories, Representation, Neural Systems, and Representational Spaces*

3.1 Structural Similarities between Scientific and Cognitive Processes

Comparing the process of science with the activities of cognitive systems one can see that both are trying to achieve rather similar goals:

(i) First of all, both are interested in *regularities* in the environment. I.e., science and cognitive systems are looking for environmental *patterns* which are occurring on a regular basis in the spatial and/or temporal domain. Before any knowledge or a theory is constructed, a cognitive system (perhaps in the context of a scientific investigation) discovers that certain phenomena in the environment happen according to some repeatable patterns or rules. These "primary regularities" are extracted by neurally realized feature detectors on various levels of abstraction. In the scientific domain these feature detectors can be compared to a theory-laden view of the world extracting these features of an observed phenomenon which seem to be relevant for the theory;

(ii) In a second step the highly inductive neural machinery isolates correlations and states in the environment, which seem to be relevant for the observed regularities. Scientific as well as cognitive processes are based on the following assumptions: (a) there is some "hidden reality" which is not directly accessible by our sensory systems (see also Kosso 1992). (b) Furthermore, (hidden) mechanisms in this "hidden reality" are responsible for the regularities which can be observed in the accessible macro-domain. In other words, these regularities are emergent phenomena of processes occurring in the "hidden domain". Looking a bit closer, one can say that these emergent phenomena are emerging in the moment of *transduction*; i.e., the micro-processes of both the environment and the particular sensor interact and the result is a neural signal leading to a certain primary representation (= "observation") in the brain. Figure 1 sketches the relationship between these domains. Science aims at constructing abstract representations or mechanisms which fit into the observed phenomena by revealing one possible relationship between the hidden and observable domain. The criterion for a "successful theory" is a mechanism which predicts or manipulates the phenomenon in an expected manner.

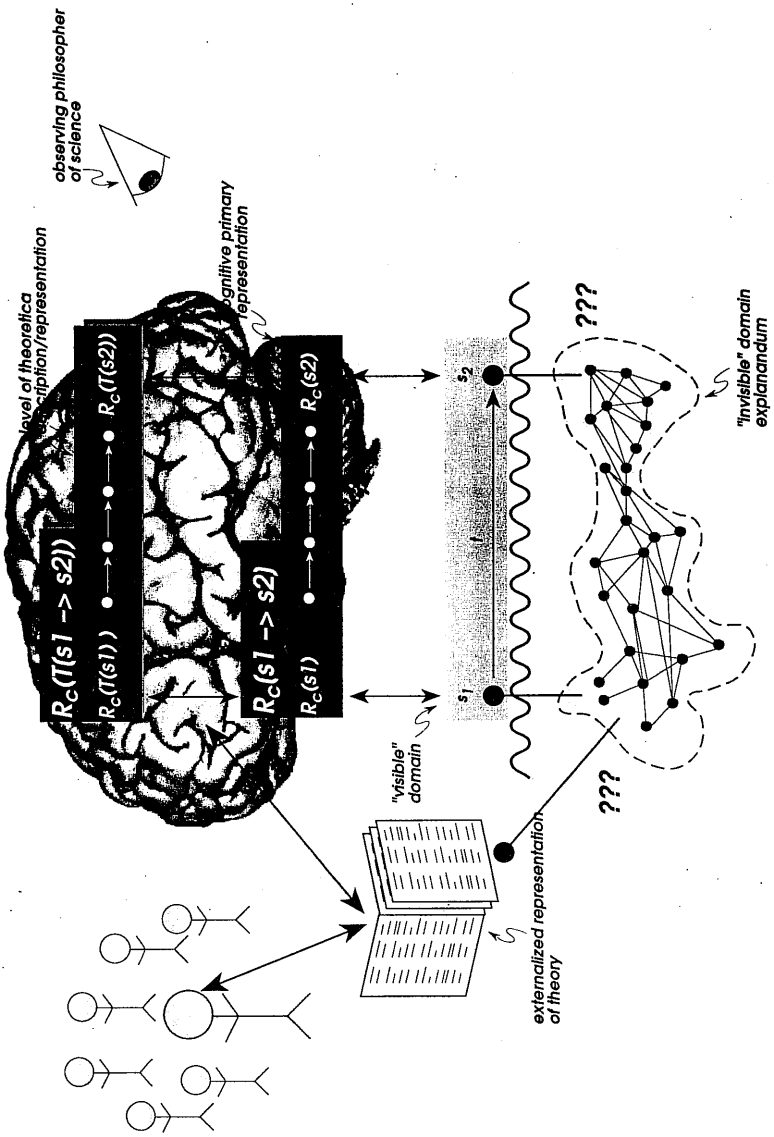


Figure 1. The relationship between the "hidden domain", the observable domain, and its representations in the brain (i.e., common sense or primary representation and representation of the theory [representing the environmental phenomenon]).

Note, that at least *two steps of constructions* are involved in this process:
(a) *constructing the correlations*: we have to keep in mind that the regularities which are extracted by the cognitive system are *system relative*; i.e., they are a result of an active process of construction which has its substratum in the neural architecture. These regularities do not explicitly "lie around" out there in the environment. Of course, the environmental dynamics follows some kind of regular pattern, but the regularities, which are extracted by the cognitive system are primarily regularities *with respect to* the representational system. In other words, the structure of the representation system constructs regularities according to its own regularities which fit into the constraints of the environmental dynamics. This process applies to both the cognitive as well as the scientific domain[4];
(b) *constructing a theory about the "hidden reality"*: as a result of the inaccessibility of the "hidden reality" the cognitive system has to *construct* a (common sense or scientific) *theory* about the mechanisms which govern this hidden domain and which lead to the observed phenomena and regularities. In other words, this representation has to account for the regularities by providing (theoretical or abstract) mechanisms which are capable of explaining, predicting, and/or generating the environmental phenomenon. This knowledge (e.g., models, abstract mechanisms, etc.) has to fit into the dynamics of the environment like a key fits into a lock (cf. the concept of *functional fitness*, Glasersfeld 1984, 1995).

The most simple form of such a representation is the model of a black box; of course, it is not a very powerful model, as it describes only the input-output relations of an observed system or phenomenon. However, in most cases such a model is the starting point for constructing more complex mechanisms, rules, dynamics, internal relationships, etc. which account for the observed behaviour. The strategy of "opening up the observed system" by means of more or less sophisticated experiments is not only applied in the domain of science, but also in everyday life. The results of this strategy are twofold: first of all, the investigating cognitive system gets some hints as to how the externally observed behaviour is generated by finding out more about the relationships and interactions between the internal subsystems of the system. Secondly, the observer realizes that with each internal subsystem a new black box is associated which has to be opened and explained. It depends on the cognitive system's question, the problem, the scientific sophistication, etc.,

at which level this (almost infinite) reductionist process stops.

In any case, there has to be *constructed* some model, knowledge, or theory (in the most general sense) in order to fill the black box, which is encountered, whenever we are interacting with the environment. Keep in mind, that the resulting theories are the result of an *active process of construction* rather than of a passive mapping. The only criterion, which has to be fulfilled by everyday as well as scientific theories, is that they are *consistent* with the environmental structures; i.e., that they fit into the environment. One of the implications is, of course, that there is more than one theory which meets this criterion. As long as this knowledge or theory can be *used* in a beneficial way for the survival (in the most general sense) of the organism or a group of organisms, it is a *functionally fitting* or adequate theory about an aspect of the environment.

(iii) The ultimate goal of all these (construction) processes is to *make use of these representations, knowledge, or theories*; in other words, to apply the more or less complex and abstract everyday or scientific models and theories in order to *predict* and/or to *manipulate* the environmental dynamics. This does not only apply to common sense knowledge, but also to scientific theories/knowledge which claim to "objectively describe" the environment. In each of these theories there is a "behavioral aspect". Most of them do not so much focus on describing the environmental dynamics, but on questions like "what happens, if..." – i.e., they are interested either in predicting the environmental dynamics under a given condition or in actively manipulating the dynamics of the environment by applying the knowledge about its internal states, relationships, and state transitions[5].

Note, that (in both the cognitive and the scientific domain) knowledge or theories are never developed just per se or just for mapping or depicting the environment. All efforts of learning, adaptation, evolution, or developing common sense knowledge or representations as well as scientific theories *finally aim at externalizing* some kind of *behaviour* which is beneficial for the organism[6]. The important thing, I want to point at in this paper, is that the traditional notion of representation or theory suggests to somehow map an aspect of the environment to some representational substratum – from a epistemological, neuroscientific, as well as philosophy of science perspective this understanding seems to be misleading, however. It can be shown that neither knowledge being represented in neural structures, nor in scientific theories primarily repre-

sent or map the world, but rather have to be seen as *strategies for successfully coping* and coupling with the world (Peschl 1994a, 1994b). I.e., there is neuroscientific as well as epistemological evidence that it is not possible to have a direct access to the environment. Due to non-linear transduction processes and the recurrent neural architecture a stable referential relationship between the environment and its representation has to be given up as well. Thus, it is by no means clear, what is meant by "mapping" or "describing" or "representing" (in the classical referential sense) the world. We lack an objective criterion for how "near" or accurate the mapping or theoretical description is to the real world, as any cognitive systems is *always* and *only* confronted with neurally constructed representations of the world – they are the only verification criterion! At best, "negative statements" can be made about the environment: i.e., in the case when a theory or representation hurts an environmental boundary condition and, thus, does not functionally fit.

It seems to be only due to our already neurally constructed view and experience of the world that, in order to behave adequately, a pictorial (Kosslyn et al. 1977, 1990, 1994), linguistic (Fodor 1975, 1981), or referential representation of the environment is thought to be necessary.

As can be seen in figure 2, cognitive and scientific processes do not only have similar goals, but also follow *structurally similar dynamics*. From an abstract and epistemological perspective both are organized as dynamic feedback systems interacting with the environment. Two feedback/recurrent loops and dynamics have to be differentiated: (i) the *internal feedback* represents the internal dynamics of knowledge or theories (i.e., the dynamics emerging from constructing, changing, etc. neural structures or theories). (ii) The second feedback dynamics concerns the *external interactions* with the environment: as a result of the internal representational dynamics the cognitive system externalizes behaviour and causes changes in the environment which are detected by the sensory system which, in turn, perturbate the representational dynamics. Similarly, a scientific theory "externalizes behaviour" by conducting an experiment. In this process the theory or knowledge is tested in the environment. On the input side the results of this experiment are measured and cause a confirmation or the need for change in the representational structure (= theory).

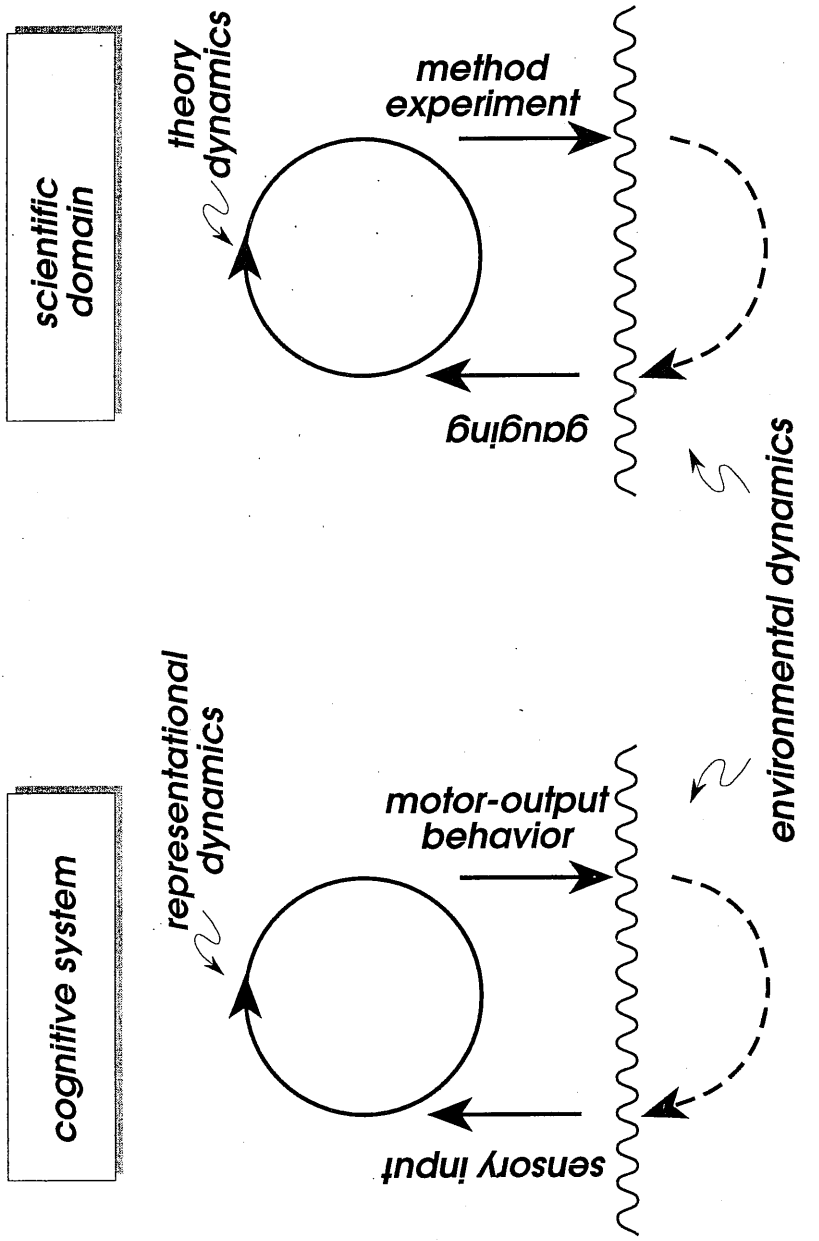


Figure 2. Structural similarities between the feedback processes in science and cognitive systems.

Of course, there is a close relationship and interaction between the internal and external feedback loops. The internal dynamics is responsible for *generating hypotheses* by continuously *adapting, constructing*, and *changing* the representational structures (i.e., neural architecture and, thus, the knowledge, theories, etc. being represented in the neural substratum). Furthermore, the internal dynamics is responsible for externalizing this knowledge in the form of behaviour, experiments, applying methods, controlling motor devices, etc. On the other hand the internal dynamics is driven in part by the signals, inputs, stimuli, etc. entering the cognitive system or theoretical domain from the environment via the sensory system or via gauges (and their interpretation). As is shown in Peschl (1994a), the environmental input does *not* determine the internal dynamics, but only has *modulatory influence* on it.

In order to test and/or verify the fitness of internally constructed knowledge or theory, behaviour is externalized according to the "instructions" from the representational structure. The organism's behavioral/motor output is the result of the internal representational dynamics/knowledge and *modulates* the environmental dynamics. In the scientific realm the behavioral output can be compared to the process of conducting an experiment, of applying a method, and/or of penetrating the environmental dynamics with some machinery. In any case the environmental dynamics is influenced in some way. These changes in the environmental states lead to a change on the organism's sensory surface or in the gauges. In this moment the new environmental state is transformed into a (neural, numerical, etc.) representation of itself ("transduction"). In the representational realm the results of these externalizations can be checked and verified, whether a desired stated has been reached or not. The success or failure of this verification process indicates how well the knowledge or theory fits into the environmental dynamics. In other words, if it failed to modulate the environment in a desired way, it is necessary to make changes in the representational structure or theory. Sections 4 and 5 will explain these processes in more detail for the cognitive as well as scientific realm.

In any case the success or failure of the behavioral externalization ("empirical experiment") determines the level of functional fitness of the representational structure or theory of the particular aspect of the environment. After the representation or theory has been changed, the cycle described above and in figure 2 starts again: behaviour is externalized

according to the new knowledge/theory, the environment is modulated, etc. Although figure 2 suggests that the processes occurring in the scientific and the cognitive realm are two different and detached systems, one has to keep in mind that scientific activities, such as constructing new theories, conducting experiments, receiving input from the environment, etc. are all *embedded* into the feedback dynamics being depicted on the left side of the figure. Perhaps that is why these processes are so similar.

3.2 Functional Fitness and Representation in Neurally Based Cognitive Systems

3.2.1 Deductive vs. Inductive Processes

The two cycles being depicted in figure 2 can be divided into two alternating subprocesses: the *deductive* part of externalizing behaviour and the more *inductive* part of constructing, adapting, and changing the representational structures (e.g., theories, knowledge, etc.). From an abstract and epistemological perspective no new knowledge is developed in the deductive process of externalizing behaviour. The only thing which happens is that already (implicitly) existing knowledge structures are applied for generating behaviour. In terms of scientific processes this means that implications, predictions, methodological instructions, etc. are deduced from a theory. In other words, certain states of the space of (theoretic) possibilities, which is implicitly predetermined by the theory describing it, are made explicit by deduction[7]. In terms of cognitive/neural systems this means that the internal representational dynamics, which is determined by the neural architecture, the current internal state, and the current input, selects a state out of its predetermined and prestructured space of possible (representational) states. As has been mentioned, the externalized behaviour is a subset of the space of possible representational states (i.e., many different internal representational states can lead to a single behavioral action; see also Peschl 1994a, 1994b for further details).

In any case, behaviour – be it in the form of an experiment or in behavioral actions of a cognitive system – always has to be interpreted as a result of a deduction in the internal representational dynamics. One could say that it is an externalization of a fraction of the organism's or theory's knowledge or representational structure (in a certain environmen-

tal context). Contrary to the deductive character of behavioral externalizations, new knowledge is constructed or existing knowledge is adapted or changed by inductive processes in cognitive systems and scientific activities. From an epistemological perspective this seems to be the more interesting process in science as well as in cognitive systems. In philosophy of science this aspect is referred to as "context of discovery"; in the context of investigating cognitive systems these processes of constructing new knowledge or changing representational structures are referred to as learning, adaptation, or evolutionary dynamics. As will be shown in the following sections, it seems that rather new theories and methods from cognitive science, which are investigating these inductive processes, could shed some light on the still mysterious process of developing new (scientific) theories.

3.2.2 Neural Representation and Transformation

For that reason a brief overview of representational mechanisms and processes in neural systems is given in the following paragraphs. As has been mentioned, any neural system can be understood as a non-linear (recurrent) transformation system which transforms an input into an output. Theoretically this transformation could be described by a recurrent function (and/or set of differential equations). Approaches in connectionism or neural computation describe this transformation by a computational neural structure or architecture which tries to model natural neural systems on a very abstract level. Thus, the behavioral and representational dynamics of a neural system can be simulated by a computer on an abstract level. From an epistemological perspective this is a very interesting process, as it becomes feasible to study representational processes and principles of neural systems in great detail and on various levels of complexity[8].

3.2.3 Representational Spaces and Substrata

Computational neuroscience provides a theoretical and explanatory framework which is of interest not only for the study of neural dynamics, but also for a better understanding of representational issues in neural systems. The core idea of this framework is based on the concept of a *state space* (which originally has been used by cybernetics and system theory;

e.g., Ashby 1964; Wiener 1948; Heiden 1992; Port 1995, and many others). In other words, the dynamics of (spreading) activations, the dynamics and changes in the synaptic weights, or the evolutionary dynamics can be explained and simulated by making use of states and state transitions in state spaces. I cannot go into details here – the focus of the following paragraphs will be only the epistemological implications for the problem of representation in neural systems.

The concept of representation changes radically with the introduction of neural activation spaces (see also Churchland 1989, 1995; Churchland et al. 1992; Peschl 1992a, 1992b, and many others). It seems that we have to give up the notion of linguistically transparent representations (Clark 1989) and replace it by the concept of *distributed representation* (Hinton et al. 1986; Rumelhart et al 1986; Gelder 1992; Elman 1991); furthermore, there is evidence (Peschl 1994a) that the concept of a *referential representation* (i.e., a representational state stands for a certain phenomenon in a stable way) has to be abandoned, as well. Especially the second issue brings about the necessity for an alternative concept of representation. As will be shown in sections 4ff, this new perspective has crucial implications even on the understanding of scientific processes.

Generally speaking, three (four) representational substrata (state spaces) can be found in neural systems. Of course, there is strong interaction between these representational dynamics going on:

(i) *activation space*: let's assume that a neural system consists of n neurons/units. Each neuron can assume a certain activation value. A state space describing the state or pattern of activations can be constructing by appointing the activation values of each neuron one dimension. Hence, a n-dimensional activation space is created, where a certain state of activations (= a pattern of activations in the neural system at a certain time t) can be described as a single point.

From an epistemological perspective, such a state in activation space can be interpreted as the current *representational state*. However, this state does not represent an environmental phenomenon in the traditional sense: first of all many different neural activations contribute to the pattern of activations. So, it is impossible to find a single representational substratum (such as a symbol). Secondly, as most neural systems have a *recurrent architecture*, the internal representational state is not only determined by the current environmental input (which is supposed to be represented in the traditional view), but also by the previous internal

state. The current input can only select from a set/space of possible successor states. However, this set of successor states is predetermined by the neural architecture and by the current internal state. So, there is no way of guaranteeing a stable referential relationship between repraesentandum and repraesentans.

As an implication of these facts, what is represented in the neural activation space can be characterized as follows: the current state or pattern of activations is *not* a stable depiction or mapping of the (transduced) environmental state. Rather, it represents a state which relates the current external input and the previous internal state to each other with the goal of generating functionally fitting behaviour.

(i-a) *trajectories in activation space* are (temporal) sequences of patterns of activations which have representational character (Horgan et al. 1996). In many cases a recurrent neural system "falls" into stable states, such as fixed point attractors, cyclic attractors, or chaotic attractors (Hertz et al. 1991). These stabilities sometimes can play the role of representations. However, the same problems with a stable referential relationship apply as in point (i).

(ii) *weight space*: the synaptic architecture of the weighs are responsible for the dynamics of activations. In other words, they control the flow and spreading of activations in the neural system. Thus, they play a rather important role in representational issues: they are responsible for generating patterns of activations (see (i)) and, thus, behaviour. Abstractly speaking, the synaptic weights determine the space of possible state transitions (and, thus, behavioral externalizations) in the activation space. The current input and the current internal state only instantiate/select one of these predetermined states and state transitions[9]. Hence, the whole representational as well as behavioral dynamics is *embodied* in the synaptic weights. This implies that, what an external observer refers to as "knowledge" of an organism is represented in the synaptic weights and architecture.

Of course, these weights are changing over time as well. This is referred to as ontogenetic adaptation or "learning". Section 4 will discuss these processes and their relevance for the development of scientific theories in detail. What is important at this point is that the weights and their dynamics can be represented in a m-dimensional weight space[10]. I.e., a certain configuration of weights is a single point in weight space. The dynamics of learning can be represented as a *sequence of points*

forming a *trajectory* through weight space. It is important to keep in mind that a certain point in weight space determines the whole structure and the dynamics of the activation space. Hence, whenever the point moves in weight space (= "learning") the dynamics changes in the activation space. This is exactly what we observe: we see that the behaviour of the organism changes and say that it must have "learned" something. From this observation we imply that its knowledge, representation or theory about the environment must have changed.

Of course, there is a close interaction between the dynamics in the weight space and activation space: the success or failure of the externalized behaviour (being the result of the current configuration in weight space) determines whether and how it is necessary to learn/adapt. The resulting changes in the weight space lead to changes in the behavioral dynamics which - hopefully - functionally fit into the environment.

(iii) *genetic space*: the genetic code is the basic representational entity for any cognitive system. It determines the basic features of the body structure as well as of the representational structure. It has to be clear that the expression of the genetic code does not lead directly to these structures - a very complex process of *development* is involved (Edelman 1988; Berger et al. 1992; Chiba et al. 1988; Jessel 1991; Lawrence 1992; Cangelosi et al. 1994). I.e., there does not take palace a 1:1 mapping from the genetic code to the mature organism. Rather, the body and representational structure develops in a complex process of interactions between the genetic code, the environment, and the body structures, which already have been produced and expressed. This implies that the genetic material has to be understood - similarly as the neural representational substratum - as representing a strategy for *generating* "behaviour" in the form of an organism (which itself has to generate adequate behaviour) in a process of development and interaction with the environment.

Again, via the criterion of success and failure (= reproduction) an interaction and feedback between the neural and the genetic representational substratum and dynamics is established. The goal is not to map the environment, but to develop representational/genetic structures which are capable of generating functionally fitting organisms. The field of artificial life (e.g., Langton 1989, 1995; Steels 1996; Meyer 1991, and many others), genetic algorithms (e.g., Mitchel 1994; Belew 1990; Holland 1975, and many others), and of studying the interaction between

evolutionary processes and neural dynamics (e.g., Belew 1990, 1992; Cangelosi et al. 1994; Harp et al. 1989; Hinton 1987; Miller et al. 1989; Nolfi et al. 1990, 1991, and many others) give new insights in this complex interplay between phylogenetic and ontogenetic dynamics. In the following sections these (computational) concepts will be applied to achieve an alternative view of scientific processes, and how they are embedded in cognitive activities.

3.2.4 Functional Fitness and Neural Representation

In the context of neural as well as genetic representation the concept of *functional fitness* plays an important role. From section 3.2.3 above we learned that there is empirical evidence that the concept of referential representations (such as propositions) has to be abandoned. The aspect of *generating behaviour* (or functioning organisms) seems to be the main task of representational structures, rather than depicting or mapping the environment. Constructivist approaches (e.g., Glasersfeld 1984, 1995; Maturana et al. 1980, and many others) refer to this concept as adequate or *viable behaviour*.

In other words, the externalized behaviour has to functionally fit into the structures of the (internal and external) environment. As we have seen above, the behaviour is the result of the internal neural dynamics being itself determined by the synaptic weight configuration, the internal state, and the input. From this perspective it can be seen that the knowledge being embodied in the synaptic architecture can be interpreted as a kind of theory or *strategy* for generating functionally fitting behaviour in the context of the organism's task to survive. The goal is to manipulate the organism's internal and external environment in such a way that it is beneficial for the cognitive system's survival and reproduction. The behaviour has to fit into the external and internal environmental constraints.

The goal of any representational dynamics can*not* be to create an accurate "picture" of the environment. Rather, constructive and adaptive processes, such as neural plasticity or evolutionary dynamics, have to change the neural architecture in such a way that it is capable of generating viable behaviour. Hence, it is not surprising that we do not have real success in trying to find referential representations in any of the representational substrata having been discussed in section 3.2.3[11]. A categorical

error seems to be involved in these investigations: How can we expect to find referential representational structures (such as symbols) in the substratum which is responsible for generating exactly these structures? In other words, the representational mechanisms and substrata being responsible for generating stable referential representations (e.g., neural activities) are confused with their results (e.g., propositions, mental images, etc.).

One of the consequences of such a view is that knowledge – and this applies to scientific theories as well – (i) is always hypothetical, (ii) is in a continuous flow, and (iii) characterizes the environment only to the extent what it is *not*[12]. *Furthermore, (iv) knowledge is always system-relative* and (v) there can exist two or more (competing) theories or strategies which equally well fit into the same environmental constraints.

4. *Acquiring New Concepts, Learning, the Context of Discovery, and Moving Around in Theory Space*

4.1 Learning and Dynamics in the Neural Representation Space

What an external observer refers to as "learning" or acquiring new knowledge can be explained, as we have seen above, as adaptation and construction processes occurring in the neural substratum. The changes in the synaptic weights lead to a change in the dynamics of spreading activations which, in turn, lead to a change in the organism's behavioral dynamics which is interpreted as a change in knowledge or as the construction of a new theory or representation by an observer.

From empirical evidence as well as simulation experiments it is known that "learning" can be interpreted as a more or less directed *search process* in the weight space on a conceptual level . Abstractly speaking, a point is moving around in weight space. This process is physically realized as changes in the synaptic weights. There exists a wide variety of "learning algorithms and mechanisms" which are based in one way or the other on Hebb's (1949) concepts of learning. Long term potentiation (LTP), Long term depression (LTD) (Brown et al. 1990; Churchland et al. 1992; Gazzaniga 1995; Dudai 1989; Singer 1990), or connectionist learning algorithms (e.g., Hertz et al 1991; Rumelhart 1989, and many others are only instantiations of Hebb's

principles. From an epistemological perspective, the basic principle can be summarized as follows: these physically realized relationships (i.e., synaptic configurations) which lead to successful behaviours are reinforced, whereas synaptic configurations which are responsible for generating inadequate behaviour are changed or suppressed.

Changing the synaptic weights is an *inductive process* in which "new" knowledge or theories is/are generated on a hypothetical basis. These strategies for generating behaviour are used and applied under the assumption "as if they were true or fitting". Only in the (deductive) process of externalizing them it becomes evident, whether the theories or knowledge being represented in the current synaptic configuration are/is viable or not (see the feedback loop in figure 2). The success or failure of the externalization leads to an internally or externally determined error[13] which has to be minimized in the following learning steps. From this perspective learning turns out to be a search process in which an error has to be minimized.

A minimal error means a (n epistemologically) stable relationship with the environment in the context of the organism's task to survive. This stability is physically realized as a stable homeostasis. In this cycle of alternately inductive and deductive processes a physical structure is developing which is capable of generating functionally fitting behaviour. From an observer's perspective, one could say that this structure "represents" knowledge about the environment and has developed in a process of construction and adaptation. However, one should not make the error, to assume that there is some kind of convergence toward a "true" or "ultimate" knowledge about the environment. As has been discussed, this knowledge or theory is always *system-relative*; i.e., the success or failure does not only depend on how well the externalized behaviour fits into the environment, but also on the internal organization of the whole organism which determines what is successful and what is a failure. In other words, the structure of the organism implicitly defines the premises under which knowledge can be successful or not. Thus, different organisms (even of a single species) will have more or less different criteria for successful, functionally fitting, viable, or adequate knowledge, theories, or behaviour. In the scientific realm this is known as the phenomenon that – under different background assumptions and methods – different theories will be adequate or "true".

Of course, this does *not* imply – as it is often done by many critics

of the constructivist approach – that the resulting constructs or theories are completely arbitrary. Rather, they have to be understood as the result of a process which aims at constructing (transformation) mechanisms which are capable of coping with the *constraints* of the internal and external environment. The ontogenetic constructs are constrained by (a) the external environment, (b) the organization of the internal environment, and (c) by the genetically determined space of possible constructs (i.e., the basic architecture of a specific nervous system allows, despite of the possibility of learning, only a certain space of possible constructs – and this space is implicitly defined by the genetic code). The dynamics of the genetic structure is constrained by the success or failure of the organism (= its reproduction rate), by the environmental resources, by the organism's body and representation architecture, as well as by the inherent mechanisms of genetic expression.

4.2 Dynamics in Theory Space

What can we learn from the concepts having been discussed above for the process of science and, more specifically, for the process of acquiring new theoretic concepts? First of all, we have to keep in mind that the process of constructing new theories is a deeply cognitive process. In other words, it is rooted in the processes and the representational dynamics having been described above and in section 3.2. From this perspective it is not surprising that similar concepts can be applied to this still mysterious process of "discovery" in science. Discovery is not so much characterized as discovering new features or regularities in the environment; rather, it is the discovery and construction of new relationships and strategies for coping successfully and effectively with the environment – these processes are only occurring inside the neural representation system.

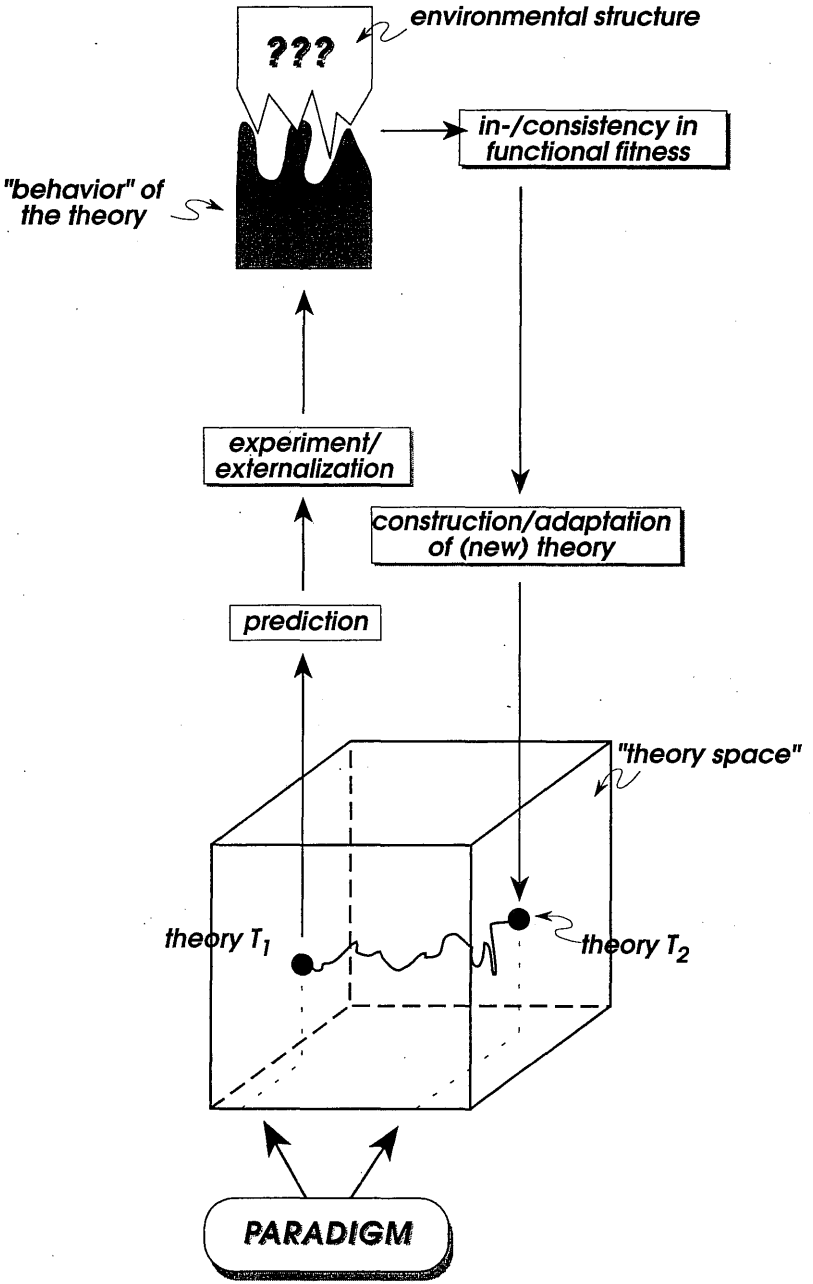


Figure 3. The dynamics in *theory space* and its interaction with the environment.

Scientific theories are represented in the same way as any other knowledge in the neural substratum. Therefore, they can be interpreted as certain states in a state space. To represent a certain theory T_i means to be in a certain state in synaptic weight space. To assume a certain weight configuration implies a set of behavioral strategies which can be externalized in/to the environment (in certain internal and external contexts). Figure 3 shows the situation for scientific theories on a more abstract level. The (implicitly assumed) scientific paradigm (in the sense) of Kuhn (1970) gives rise to a space of possible theories. Each point in this theory space instantiates a certain theory T_i. This space is embedded in the larger synaptic weight space. Thus, moving around in the synaptic weight space has an effect on the state of the theory space.

The process of constructing new theories or changing/adapting already existing theories is based on the same neural processes as any learning process (see above) – it can be characterized as search process in a representational state space. In figure 3 this cyclic process of developing theories is depicted in detail: the theory space is embedded in the neural representation space. A certain point in this (high dimensional) space represents a certain theory T_i. Similarly as in the common sense domain, this neural configuration leads to (i) a prediction which, in turn, can be (ii) *externalized* in the form of an *experiment*. The experiment can be compared to the behaviour of a cognitive system. I.e., some kind of direct or indirect motor action modulates, or perturbates the environmental dynamics. In other words, the theory represents knowledge or a strategy about how to penetrate the environment in such a way that a desired or predicted state is achieved. The theory also determines the methods which are applied to the environment. In the case of cognitive systems the "method" are the motor systems; in the scientific realm these motor systems are extended by more or less complex tools and/or machines which perturbate the environment according to the theory's rules and instructions.

As can be seen in figure 3, the resulting "theory's behaviour", which is externalized in an experiment, fits more or less into the structure of the environment. The level of functional fitness is determined by the *success* of penetrating the environment in a certain (desired) way. If the environmental dynamics does not "respond" in the desired or predicted way, this indicates that this particular (configuration of the) theory has failed (or is falsified), and that it is necessary to change, adapt or completely recon-

struct it. The goal is to *reduce the inconsistencies* between the theoretical descriptions, predictions, and the actual environmental dynamics. Looking at this process the other way around, theories as well as any other (successful) representational structure can be described as results of a process which aims at establishing consistency between environmental and body constraints. The knowledge or theory is the mediating substratum which is responsible for generating functionally fitting behaviour in the process of interaction between the organism and the environment. This consistency is achieved by a continuous process of adaptation and construction of functionally fitting strategies and behaviours. These construction and adaptation processes are realized by the neural dynamics having been shown in section 4.1. It can be described as an optimization process searching for an adequate transformation mechanism which is realized as a weight configuration in the synaptic weight space. In other words, a point moves around in theory space – each point instantiates a particular theory and moving points in theory space represent the transition form one theory to another.

The goal of these processes is to *extract relevant regularities* from the environmental dynamics. The representation of these system relative regularities are the foundation for any externalization of behaviour or experiments. They are used for making predictions – and, in most cases, predictions enhance the chances for survival or, at least, simplify life, if they prove to be successful in the "real world". From this perspective, neural construction and adaptation processes are the heart of any scientific *inductive* process in which a new theory is created or an already existing theory is adapted or changed. Furthermore, it turns out that, as can be seen in figure 3, the "creation" or construction of new theories or knowledge does not bring forth "really new" knowledge. Rather, the context of discovery can be described as a *search* process in an already predetermined space of possible theories. This space is predetermined by the paradigm (Kuhn 1970) which has been chosen by the cognitive system. The goal of this search process is to *optimize* the fit and the level of consistency within the boundaries of this paradigm[14]. Most research which is done in modern natural sciences turns out to be optimizing sets of parameters, methods, experimental set-ups, etc. leading to a better fit and consistency between predicted and actual phenomena[15].

This may sound a bit disappointing and provocative in the context of the epistemological and social status which scientific knowledge or theo-

ries normally claims to have. As an implication of the "cognitive view" as well as of a historic view of scientific processes, the notion of ultimate, objective, or true knowledge has to be seriously questioned. As has been mentioned already, I suggest to replace it by the concepts of system relativity, functional fitness, and viability. Scientific knowledge, nevertheless, remains at the peak of what we can know about our environment. However, it will always remain hypothetical, system relative, in steady flow, and does not describe or map the environment, but rather provides strategies for successfully coping, modulating, and manipulating the environmental dynamics.

In the picture about embedding scientific processes in neural dynamics, which has been presented above, a couple of questions remain unanswered: What happens, if the search in theory space is not successful or unsatisfactory? What happens, if the cognitive system's goals and desires change? Who defines the (semantics of the) dimensions of the theory space? Which role does the paradigm play, and what happens, if the scientist (alias cognitive system) changes the paradigm or "invents" or constructs a new paradigm? These questions will be addressed in the following section.

5. *Paradigmatic Shifts and Theory Spaces*

In the previous section we have seen that a paradigm (Kuhn 1970) gives rise to a space of possible theories (= "theory space") which is searched in the process of "normal science". In this context we are facing the problem of what happens to the theory space when a paradigmatic shift occurs and, furthermore, which (cognitive) mechanisms can be found to explain paradigmatic shifts. As is well known from the history of science, in most cases a paradigmatic shift leads to completely new concepts, perspectives, and categories of how to view and understand a certain environmental phenomenon. In constructivist terminology this means that a completely new key is constructed in order to fit into the lock/environment[16]. How do these new categories, terminologies, and theories emerge? In order to approach this problem from a cognitive perspective, let's have a look at the interactions between evolutionary processes and ontogenetic dynamics in the neural representational substratum in a first step.

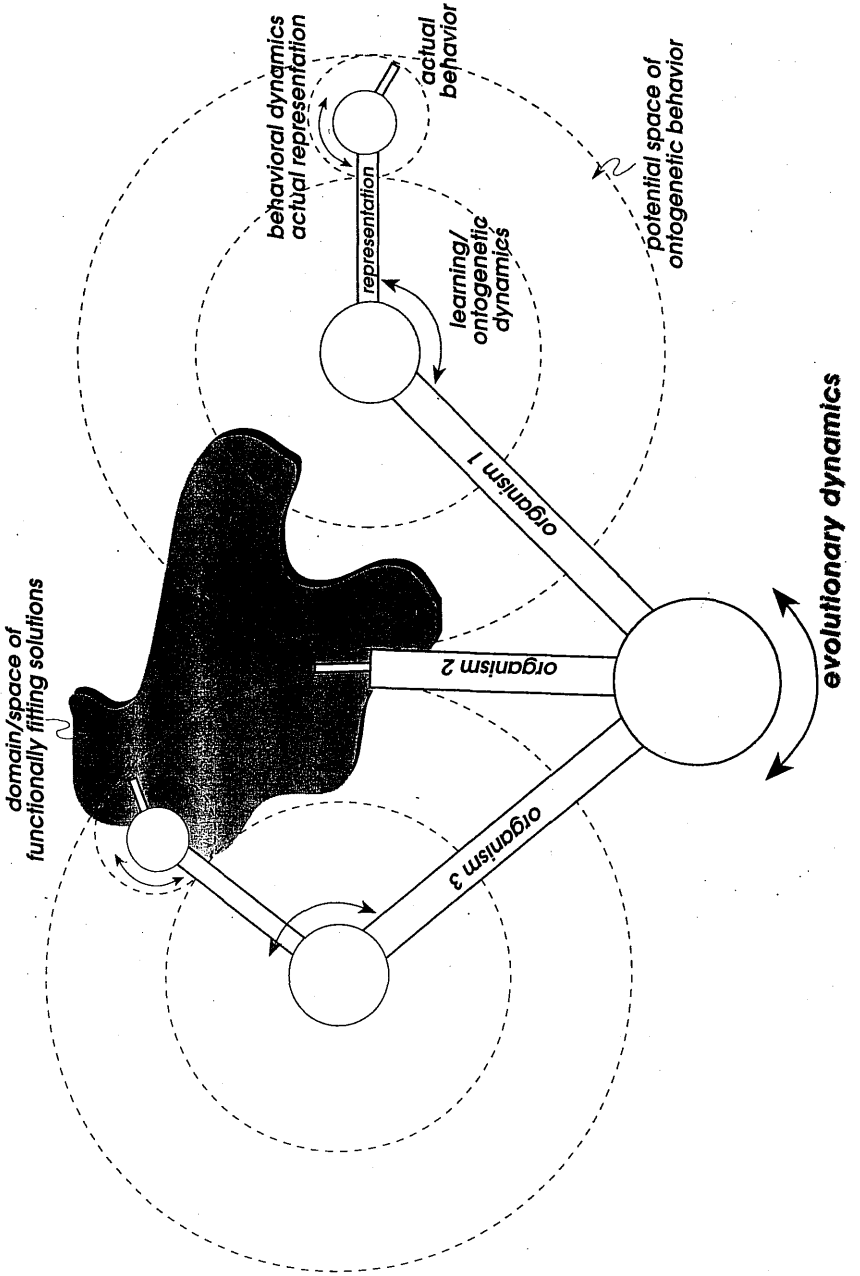
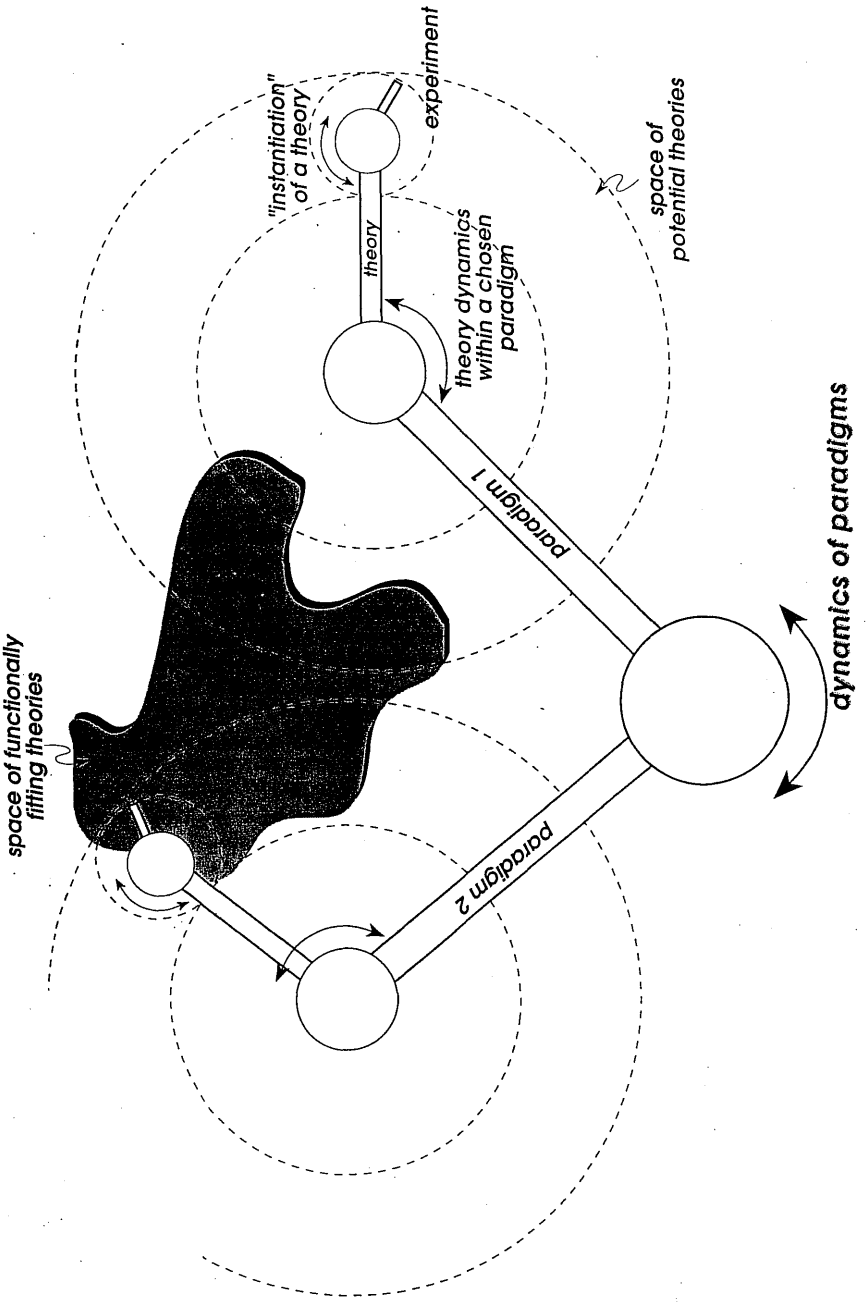


Figure 4. Comparing the evolutionary and ontogenetic dynamics of cognitive systems and scientific processes in the "robot-arm analogy"


experiment
"instantiation" of a theory
theory
theory dynamics within a chosen paradigm
space of potential theories
paradigm 1
dynamics of paradigms
space of functionally fitting theories
paradigm 2

The left side of figure 4 shows the "robot-arm analogy" (Belew 1990) for cognitive systems. It is an analogy which demonstrates the interaction between phylogenetic/evolutionary and ontogenetic dynamics. This robot-arm has three degrees of freedom: (i) in the *evolutionary dynamics* a particular genetic code is instantiated and expressed. In the process of interaction with the environment an organism develops. This organism has a (neural) representation system at its disposal. (ii) The second degree of freedom consists in the representational dynamics having been discussed in section 4.1 (i.e., learning, neural plasticity, search in weight space, etc.). The current state of the representation/weight space gives rise to a structure in activation space. (iii) The dynamics of spreading activations instantiates states in the activation space and leads to the externalization of behaviour. The success or failure of this behaviour (= level of functional fitness of the behaviour and of the representational structure) causes changes in the representational dynamics. Over more generations the success or failure of the basic architecture of the representation system, the resulting behaviour, as well as the genetically encoded basic body structures, developmental instructions, and learning/adaptation mechanisms cause a genetic drift ("evolution"). More abstractly speaking, the genetic code changes over time and gives rise to a newly structured representational space (= synaptic weight space, potential space of possible representational configurations of an organism) and, thus, to a new set of behavioral strategies. In the course of ontogenesis this space is searched, as has been described in section 4.1. A particular state in the weight space gives rise to a space of potential representations and behavioral strategies (= activation space)[17]. The goal of the phylo- and ontogenetic dynamics is to turn the robot's arm in such a way that its tips find the region of functionally fitting solutions, behaviours, or knowledge.

What can these interactions between evolutionary and ontogenetic representational dynamics teach us for our problem of paradigmatic shifts? The right side of figure 4 shows how the dynamics of paradigms and theories fits into this picture. In section 4.2 it has been shown that "normal science" (in the sense of Kuhn (1970)) can be characterized as a search and optimization process in theory space. It is embedded in the adaptation, learning, and construction processes of the whole neural representation system. The goal is to find consistency between the environment and the theories which are generated (i.e., moving points in

theory space) within the context and boundaries of the chosen paradigm. In this terminology a *paradigmatic shift* can be described as the construction of a whole *new theory space*. It consists of different dimensions, new and different semantics in the dimensions, and different representational and behavioral dynamics. This newly constructed theory space, of course, is embedded in the neural substratum and represents a whole new space of potential representational constructs, relationships, and behavioral patterns. In order to test this potential space of new theories, this new theory space has to be explored, as described in section 4.2.

Looking at examples from history of science, one can see that the introduction of new paradigms often brought about some kind of surprise about the new way of looking at and structuring well known phenomena. It is the "irrational" and unexpected character which makes paradigmatic shifts so interesting – whereas in normal science most results are rather predictable and the theories being responsible for them have to undergo only minor adaptations[18]. Contrary to already established paradigms, newly constructed and unexplored paradigms are based on completely new concepts, basic assumptions, terminologies, and methods in most cases. This "irrational" character suggests that the (cognitive) processes being involved in generating paradigmatic shifts might have *evolutionary character*: a new paradigm is brought forth in a trial-&-error manner. It is even more hypothetical than the generation of a new theory in the context of an already established theory space/paradigm. This is due to the fact that at the moment of the introduction of a new paradigm it means only to suggest and generate a completely hypothetical framework and space of potential theories. Hence, there is relatively high risk involved in this process. It can be compared to the process of expressing a gene which has undergone some kind of mutation. It is completely unclear, whether the resulting organism and its potential representational structures and behaviours will be capable of surviving. Similarly, at the moment of the conception of a new paradigm, a totally new potential theory space is created which has to be explored by the process of "normal science" – it is not at all clear, whether this space of potential theories will be successful or not.

The mechanisms being involved in paradigmatic shifts can be compared to *evolutionary operators* which are applied to cognitive/representational structures. The introduction of completely new and unexpected categories, making use of metaphors, combining aspects from different

theories, etc. have a lot in common with random mutations, cross over operators, etc. By applying these operators, a completely new theory space is established – as can be seen in figure 4 (right), the goal is to rotate the robot arm into the region of functionally fitting theories. From this perspective it is also clear that two or more different theories can account for the same phenomenon. I.e., the same area of functionally fitting solutions can be reached with different robot-arm configurations (= different conceptual systems). In other words, a phenomenon is approached from two or more different sides or angles. As the goal is not to create an image or 1:1 mapping of the environment, but to construct consistencies in the form of functionally fitting behaviour, it is no contradiction that two or more theories can account for the same phenomenon by making use of different representational categories. In any case, the interaction between evolutionary mechanisms and ontogenetic representational dynamics could shed some light on the mysterious phenomenon of paradigmatic shifts in science. Evolutionary operators act as "paradigm generators"; each of these paradigms establishes a space of potential theories which has to be searched according to the rules having been outlined in sections 4.2 and 4.1.

Critics of such an evolutionary perspective of growth and development of scientific knowledge (e.g., Thagard 1988) are right in stating that a purely blind search for new scientific concepts is not an adequate model. That is why the focus of this paper is not only on phylogenetic processes, but also on ontogenetic learning/adaptation (see section 4). The important point is the *interaction* between the rather directed ontogenetic and neurally based learning, adaptation, and construction processes and the "blind" phylogenetic processes. Evolutionary variation "blindly" brings forth a completely hypothetical space of knowledge/strategies (paradigm) which is explored in a directed manner in the course of ontogenetic development. In this process of exploration the new paradigm will prove its in-/adequacy very soon.

Contrary to Thagard's view(1988) that "the biological roots of the human information processing system are not directly relevant to the task of developing a model for the growth of scientific knowledge" (p 105), the presented concepts suggest that so-called scientific processes are not at all abstract processes occurring in a detached system called science. Rather, they are embedded in and results of the activities and dynamics of one or a group of neural systems and, thus, follow a similar dynamics.

6. *Implications and Conclusion*

The approach being suggested in this paper does not aim at replacing traditional theories and concepts in philosophy of science. Rather, the goal is to give (the explanation and theory about) science a new foundation – scientific processes are embedded in cognitive processes. Starting from this very basic assumption, it might be possible to view scientific processes and traditional approaches in philosophy of science from a new (cognitive) perspective which might lead to a reformulation of these theories. I am aware that there is still a very long way to go, until satisfactory theories about the cognitive foundation of science will be available. Hence, the goal of this paper was *not* to suggest such a detailed theory, but rather to show, how basic principles and concepts which have emerged in last decade in the fields of cognitive science, computational neuroscience, and artificial life can be used as *tools* for enabling the construction of such a theory.

Both cognitive systems and science have the representation of the world as their most important task. As has been shown, neither neural nor scientific representation aims at depicting or mapping the world. Of course, the obvious goal of science is to describe and explain environmental phenomena – however, as has been discussed in the context of the constructivist concepts, the epistemological status of theories (and any kind of knowledge, in general) is constructive rather than descriptive. Hence, theories, even if they have "descriptive character", are the result of complex processes of construction, which, in principle, can not produce more than functionally fitting representations of the world. The goal is to *construct strategies* which are capable of adequately predicting and coping with the environmental dynamics. Even if it seems that our cognitive or nervous system provides us with "pictures of the environment", it can be shown that these pictures are the result of *active processes of construction* being embodied in the architecture of the neural representation system. The goal of these construction processes is not to reconstruct the environmental structure as accurately as possible, but to provide the organism with relevant information and representations for generating adequate behaviour, making reasonable decisions, etc. The representations or theories only have to fit into the environment. This constructive character of representations becomes even more obvious in the case of scientific theories. In most cases they do not speak about entities which

can be perceived by our sensory systems – theoretic entities are constructs about hidden mechanisms[19] which fit (as explanations or predictions) into the perceivable environmental dynamics.

In this sense knowledge and/or theories become *tools* which are used for predicting, controlling, and manipulating the environment. In the common sense domain we use representational entities, such as concepts, symbols, language, etc., as means for manipulating and influencing the environmental dynamics and/or the (representational) dynamics of other cognitive systems. Although theoretic or scientific entities, such as concepts in physics (particles, waves, force fields, etc.), biology, or psychology (dynamics of propositions, "mind", etc.), have never been explicitly "seen" or felt, they turn out to be extremely powerful and useful tools in the domains of predicting, manipulating, and explaining environmental phenomena:

As has been shown, concepts from computational neuroscience and artificial life provide a conceptual framework enabling the embedding of scientific into cognitive processes. The dynamics of theories is realized in the dynamics occurring in the synaptic weight space and the genetic space. From this perspective the "context of discovery" and the construction of new scientific concepts can be operationalized in the sense that cognitive and neural mechanisms act as explanatory vehicles which account for the still mysterious process of discovering and constructing new scientific knowledge.

As an implication, scientific knowledge becomes some kind of "truth-tool", which is not necessarily structurally equivalent or homomorphic with the environment. From studying neural systems we can learn that representation has to be understood as a *strategy for coping with the environment* in the context of the organism's task to survive. This is achieved by a mutual process of adaptation and construction leading to changes in the representational structure which enable the generation of (hopefully) adequate behaviour. The resulting representational structure does not have anything to do with an "objective" description of the environment. The same can be applied in the scientific domain: why do we expect from tools or strategies for manipulating the environment that they map or represent the environment in an iso-/homomorphic and "objective" way? Do we expect from a hammer or a tooth that it is structurally equivalent with a nail or with the entities the tooth chews? Both represent knowledge in the sense that they are results of constructive

processes which are based on evolutionary and cognitive dynamics. In both cases the goal is to cope successfully with some environmental phenomenon or problem. Similarly, scientific theories are not so much descriptions of an environmental phenomenon, but answers to questions of how to deal with this phenomenon in the form of functionally fitting solutions.

University of Vienna

NOTES

1. To be more precise, a new internal state is generated. The behavioral output is a subset of this internal state. I.e., a subset of all neurons which constitute the internal state is connected to motor systems and, thus, control the organism's behavioral dynamics.
2. These environmental states are "desired" in the sense as they are necessary for the organism's survival. Of course, the manipulation of environmental states includes also the organism's internal environment, such as blood pressure, body temperature, etc.
3. In its most extreme form the idea of successfully manipulating the environment – without being really interested in understanding what is happening – can be found in the development of modern *technologies*. The goal is not so much an adequate description of the environment, but a *functioning system*.
4. In the scientific domain this can be seen even more clearly: a variety of *different theories* exists for a single phenomenon. Most of these theories describe the environmental regularities quite well with respect to their assumptions and underlying theoretical framework.
5. This behavioral or manipulative aspect can be found in its minimal form whenever an *experiment* is conducted. One aspect of the environment is actively pushed into a certain state in order to produce a certain phenomenon – it does not matter whether this is the process of letting fall an apple, of accelerating electrons, or of putting an subject into a psychological lab and an experimental setting.
6. And this applies to any organism and to almost any form of knowledge.
7. This process has been investigated in great detail by *traditional* philosophy of science.

8. In many cases this would not be possible in empirical experiments (for practical, methodological, and/or ethical reasons)
9. This process of instantiating a certain state or state transition can be compared to the process of *deducing* certain predictions, behaviours, or experiments from theories; see also section 3.2.1 and the right parts of figure 2.
10. Under the assumption that the neural networks consists of *m* synaptic weighs.
11. See also the discussions about the representational capabilities of artificial and natural neural systems.
12. It characterizes the environment only *negatively*; i.e., it is the result of the "collisions" with the environment in the process of interacting and adapting with/to it.
13. In many cases this *error* is defined by the state of the organism's dis-/-equilibrium or homeostasis.
14. This process could be compared to Kuhn's concept of puzzle solving (Kuhn 1970).
15. Think, for instance, of *serial experiments* in biology, physics, in the process of developing almost any theory.
16. Whereas in the search process (of normal science) described in section 4.2 only minor changes are made to the key.
17. The spaces of potential behaviours or representations are marked by dotted circles/regions in figure 4.
18. I.e., their main claims and the basic assumptions and categories, on which they are based, remain unquestioned and *untouched.*
19. I.e., these mechanisms are not perceivable by human sensory systems. They are hidden in the sense that they seem to be responsible for the directly observable phenomena, but can be accessed only via sophisticated instruments which detect and transform these hidden dynamics into the perceivable domain according to the theories on which they are based.

REFERENCES

Anderson J.A., A. Pellionisz, and E. Rosenfeld (Eds.) (1991), *Neurocomputing 2.* Directions of research. Cambridge, MA: MIT Press.

Anderson J.A. and E. Rosenfeld (Eds.) (1988), *Neurocomputing. Foundations of research.* Cambridge, MA: MIT Press.

Arbib M.A. (Ed.) (1995), *The handbook of brain theory and neural net-*

works. Cambridge, MA: MIT Press.

Ashby R. W. (1964), *An introduction to cybernetics*. London: Methuen.

Belew R.K. (1990), Evolution, learning, and culture: computational metaphors for adaptive algorithms. *Complex Systems* **4**, pp. 11–49.

Belew R.K., J. McInerney, and N.N. Schraudolph (1992), Evolving networks: using the genetic algorithm with connectionist learning. In C.G. Langton, C. Taylor, J.D. Farmer, and S. Rasmussen (Eds.), *Artificial Life II*, Redwood City, CA, pp. 511–547. Addison-Wesley.

Berger P. and M. Singer (1992), *Dealing with genes: the language of heredity*. Mill Valley, CA: University Science Books.

Brakel J.v. (1994), Cognitive scientism of science. *Psycoloquy (electronic journal) 5*(20). (filename: scientific-cognition.3.vanbrakel).

Brown T.H., A.H. Ganong, E.W. Kariss, and C.L. Keenan (1990), Hebbian synapses: biophysical mechanisms and algorithms. *Annual Review of Neuroscience* **13**, pp. 475–511.

Cangelosi A., D. Parisi, and S. Nolfi (1994), Cell division and migration in a genotype for neural networks. *Network: computation in neural systems* **5**(4), pp. 497–516.

Chiba A., D. Shepherd, and R.K. Murphey (1988), Synaptic rearrangement during postembryotic development in the cricket. *Science* **240**, pp. 901–905.

Churchland P.M. (Ed.) (1989), *A neurocomputational perspective – the nature of mind and the structure of science*. Cambridge, MA: MIT Press.

Churchland P.M. (1991), A deeper unity: some Feyerabendian themes in neurocomputational form. In G. Munevar (Ed.), *Beyond reason: essays on the philosophy of Paul Feyerabend*, pp. 1–23. Dordrecht, Boston: Kluwer Academic Publishers. (reprinted in R.N.Giere (ed.), Cognitive models of science, Minnesota Studies in the Philosophy of Science XV, 1992).

Churchland P.M. (1995), *The engine of reason, the seat of the soul. A philosophical journey into the brain*. Cambridge, MA: MIT Press.

Churchland P.S., C. Koch, and T.J. Sejnowski (1990), What is computational neuroscience? In E.L. Schwartz (Ed.), *Computational neuroscience*. Cambridge, MA: MIT Press.

Churchland P.S. and T.J. Sejnowski (1989), Neural representation and neural computation, In A.M. Galaburda (Ed.), *From reading to neurons*, pp. 217–250. Cambridge, MA: MIT Press.

Churchland P.S. and T.J. Sejnowski (1992), *The computational brain*. Cambridge, MA: MIT Press.

Clark A. (1989), *Microcognition: philosophy, cognitive science, and paral-*

lel distributed processing. Cambridge, MA: MIT Press.

Dudai Y. (1989), *The neurobiology of memory: concept, findings, trends*. New York: Oxford University Press.

Edelman G.M. (1988), *Topobiology: an introduction to molecular embryology*. New York: Basic Books.

Elman J.L. (1991), Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning* 7(2/3), 195–225.

Fodor J.A. (1975), *The language of thought*. New York: Crowell.

Fodor J.A. (1981), *Representations: philosophical essays on the foundations of cognitive science*. Cambridge, MA: MIT Press.

Foerster H.v. (1973), On constructing a reality. In W.F. E. Preiser (Ed.), *Environmental design research*, Volume 2. Stroudsburg, PA: Hutchinson & Ross. (reprinted in P.Watzlawick (ed.), The invented reality, Norton, pp 41–61, 1984).

Gazzaniga M.S. (Ed.) (1995), *The cognitive neurosciences*. Cambridge, MA: MIT Press.

Gelder T.v. (1992), Defining "distributed representation". *Connection Science* 4(3/4), pp. 175–191.

Gelder T.v. and R. Port (1995), It's about time: an overview of the dynamical approach to cognition. In R. Port and T.v. Gelder (Eds.), *Mind as motion*. Cambridge, MA: MIT Press.

Giere R.N. (Ed.) (1992), *Cognitive models of science*, Volume XV of *Minnesota studies in the philosophy of science*. Minneapolis: University of Minnesota Press.

Giere R.N. (1994), The cognitive structure of scientific theories. *Philosophy of Science* **61**, pp. 276–296.

Glasersfeld E.v. (1984), An introduction to radical constructivism. In P. Watzlawick (Ed.), *The invented reality*, pp. 17–40. New York: Norton.

Glasersfeld E.v. (1995), *Radical constructivism: a way of knowing and learning*. London: Falmer Press.

Goldberg D.E. (1989), *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley.

Green D.W. et al. (1996), *Cognitive science. An introduction*. Cambridge, MA: B. Blackwell.

Harp S., T. Samad, and A. Guha (1989), Towards the genetic synthesis of neural networks. In J.D. Schaffer (Ed.), *Proceedings of the Third International Conference on Genetic Algorithms*, San Mateo, CA. M. Kaufmann Pub.

Hebb D.O. (1949), *The organization of behaviour; a neuropsychological theory*. New York: Wiley.

Heiden U.a.d. (1992), Selbstorganisation in dynamischen Systemen. In W. Krohn and G. Küppers (Eds.), *Emergenz: die Entstehung von Ordnung, Organisation und Bedeutung*, pp. 57–88. Frankfurt/M.: Suhrkamp.

Hertz J., A. Krogh, and R.G. Palmer (1991), *Introduction to the theory of neural computation*, Volume 1 of *Santa Fe Institute studies in the sciences of complexity. Lecture notes*. Redwood City, CA: Addison-Wesley.

Hinton G.E. (1987), Connectionist learning procedures. Technical Report CMU-CS-87-115, Carnegie-Mellon University, Pittburgh, PA.

Hinton G.E., J.L. McClelland, and D.E. Rumelhart (1986), Distributed representations. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing: explorations in the microstructure of cognition. Foundations*, Volume I, pp. 77–109. Cambridge, MA: MIT Press.

Hinton G.E. and S.J. Nowlan (1987), How learning can guide evolution. *Complex Systems* **1**, pp. 495–502.

Hinton G.E. and T.J. Sejnowski (1986), Learning and relearning in Boltzman machines. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing: explorations in the microstructure of cognition. Foundations*, Volume I, pp. 282–316. Cambridge, MA: MIT Press.

Holland J.H. (1975), *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor: University of Michigan Press.

Horgan T. and J. Tienson (1996). *Connectionism and the philosophy of psychology*. Cambridge, MA: MIT Press.

Jessel T.M. (1991), Neuronal survival and synapse formation. In E.R. Kandel, J.H. Schwartz, and T.M. Jessel (Eds.), *Principles of neural science* (3rd ed.)., pp. 929–944. New York: Elsevier.

Kosslyn S.M. (1990), Mental imagery. In D.N. Osherson and H. Lasnik (Eds.), *An invitation to cognitive science*, Volume 2, pp. 73–97. Cambridge, MA: MIT Press.

Kosslyn S.M. (1994), *Image and brain. The resolution of the imagery debate*. Cambridge, MA: MIT Press.

Kosslyn S.M. and J.R. Pomerantz (1977), Imagery, propositions, and the form of internal representations. *Cognitive Psychology* **9**, pp. 52–76.

Kosso P. (1992), *Reading the book of nature*. Cambridge: Cambridge University Press.

Kuhn T.S. (1970), *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Langton C.G. (Ed.) (1989), *Artificial Life*, Redwood City, CA. Addison-Wesley.

Langton C.G. (Ed.) (1994), *Artificial Life III*, Redwood City, CA. Addison-Wesley.

Langton C.G. (Ed.) (1995), *Artificial Life. An Introduction*. Cambridge, MA: MIT Press.

Lawrence P.A. (1992), *The making of a fly. The genetics of animal design*. London; Boston: B. Blackwell.

Maturana H.R. and F.J. Varela (Eds.) (1980), *Autopoiesis and cognition: the realization of the living*, Volume 42 of *Boston studies in the philosophy of science*. Dordrecht; Boston: D.Reidel Pub. Co.

Meyer J.A. and S.W. Wilson (Eds.) (1991), *From animals to animats: proceedings of the First International Conference on Simulation of Adaptive Behaviour (SAB '90)*, Cambridge, MA. MIT Press.

Miller G., P. Todd, and S. Hedge (1989), Designing neural networks using genetic algorithms. In J.D. Schaffer (Ed.), *Proceedings of the Third International Conference on Genetic Algorithms*, San Mateo, CA. M. Kaufmann Pub.

Mitchell M. and S. Forrest (1994), Genetic algorithms and artificial life. *Artificial Life* **1**(3), pp. 267–291.

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

Newell A., P.S. Rosenbloom, and J.E. Laird (1989), Symbolic architectures for cognition. In M.I. Posner (Ed.), *Foundations of cognitive science*, pp. 93–131. Cambridge, MA: MIT Press.

Newell A. and H.A. Simon (1976), Computer science as empirical inquiry: symbols and search. *Communications of the Assoc. for Computing Machinery (ACM)* **19**(3), pp. 113–126. (reprinted in M.Boden (ed.), The Philosophy of Artificial Intelligence, Oxford University Press, 1990).

Nolfi S., J.L. Elman, and D. Parisi (1990), Learning and evolution in neural networks. Technical Report CRL-9019, CRL, Univ. of California, San Diego.

Nolfi S. and D. Parisi (1991), Growing neural networks. Technical Report PCIA-91-18, Inst. of Psychology, C.N.R., Rome.

Peschl M.F. (1992a), Construction, representation, and embodiment of knowledge, meaning, and symbols in neural structures. Towards an alternative understanding of knowledge representation and philosophy of science. *Connection Science* **4**(3&4), pp. 327–338.

Peschl M.F. (1992b), Embodiment of knowledge in natural and artificial neural structures. Suggestions for a cognitive foundation of philosophy of science from a computational neuroepistemology perspective.

Methodologica **11**, pp. 7–34.

Peschl M.F. (1993), Knowledge representation in cognitive systems and science. In search of a new foundation for philosophy of science from a neurocomputational and evolutionary perspective of cognition. *Journal of Social and Evolutionary Systems* **16**, pp. 181–213.

Peschl M.F. (1994a), Autonomy vs. environmental dependency in neural knowledge representation. In R. Brooks and P. Maes (Eds.), *Artificial Life IV*, Cambridge, MA, pp. 417–423. MIT Press.

Peschl M.F. (1994b), *Repräsentation und Konstruktion. Kognitions- und neuroinformtische Konzepte als Grundlage einer naturalisierten Epistemologie und Wissenschaftstheorie.* Braunschweig/Wiesbaden: Vieweg.

Port R. and T.v. Gelder (Eds.) (1995), *Mind as motion: explorations in the dynamics of cognition.* Cambridge, MA: MIT Press.

Posner M.I. (Ed.) (1989), *Foundations of cognitive science.* Cambridge, MA: MIT Press.

Roth G. (1994), *Das Gehirn und seine Wirklichkeit. Kognitive Neurobiologie und ihre philosophischen Konsequenzen.* Frankfurt/M.: Suhrkamp.

Rumelhart D.E. (1989), The architecture of mind: a connectionist approach. In M.I. Posner (Ed.), *Foundations of cognitive science*, pp. 133–159. Cambridge, MA: MIT Press.

Rumelhart D.E., G.E. Hinton, and R.J. Williams (1986), Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing: explorations in the microstructure of cognition. Foundations*, Volume I, pp. 318–361. Cambridge, MA: MIT Press.

Rumelhart D.E. and J.L. McClelland (Eds.) (1986), *Parallel Distributed Processing: explorations in the microstructure of cognition. Foundations*, Volume I. Cambridge, MA: MIT Press.

Rumelhart D.E., P. Smolensky, J.L. McClelland, and G.E. Hinton (1986), Schemata and sequential thought processes in PDP models. In J.L. McClelland and D.E. Rumelhart (Eds.), *Parallel Distributed Processing: explorations in the microstructure of cognition. Psychological and biological models*, Volume II, pp. 7–57. Cambridge, MA: MIT Press.

Schwartz E.L. (Ed.) (1990), *Computational neuroscience.* Cambridge, MA: MIT Press.

Sejnowski T.J., C. Koch, and P.S. Churchland (1988), Computational neuroscience. *Science* **241**(4871), pp. 1299–1306.

Singer W. (1990), Search for coherence: a basic principle of cortical self-organization. *Concepts in Neuroscience* **1**, pp. 1–26.

Steels L. (1996), The origins of intelligence. In *Proceedings of the Carlo*

Erba Foundation Meeting on Artificial Life, Milano. Foundazione Carlo Erba.

Steier F. (Ed.) (1991), *Research and reflexivity*. London; Newbury Park, CA: SAGE Publishers.

Thagard P. (1988), *Computational philosophy of science*. Cambridge, MA: MIT Press.

Varela F.J., E. Thompson, and E. Rosch (1991), *The embodied mind: cognitive science and human experience*. Cambridge, MA: MIT Press.

Watzlawick P. (Ed.) (1984), *The invented reality*. New York: Norton.

Wiener N. (1948), *Cybernetics; or, Control and communication in the animal and the machine*. New York: Wiley.

Winston P.H. (1992), *Artificial Intelligence* (3rd ed.). Reading, MA: Addison-Wesley.