The Construction of 'Reality' in the Robot: Constructivist Perspectives on Situated Artificial Intelligence and Adaptive Robotics

Tom Ziemke

Dept. of Computer Science, University of Skövde

Box 408, S-54128 Skövde, Sweden

tom@ida.his.se

tel +46-500-438330 / fax +46-500-438399

Abstract

This paper discusses different approaches in cognitive science and artificial intelligence research from the perspective of radical constructivism, addressing especially their relation to the biologically based theories of von Uexküll, Piaget as well as Maturana and Varela. In particular recent work in 'New AI' and adaptive robotics on situated and embodied intelligence is examined, and we discuss in detail the role of constructive processes as the basis of situatedness in both robots and living organisms.

Keywords: adaptive robotics, artificial intelligence, embodied cognition, radical constructivism, situatedness

Running head: The Construction of 'Reality' in the Robot

To appear in *Foundations of Science*, special issue on *'Radical Constructivism and the Sciences'* (guest editor: Alexander Riegler), late 2000 or early 2001.

Preliminary (almost final) version 18.3, 000731.

1. Introduction

Let us start with the title of this paper: *"The Construction of 'Reality' in the Robot"* is, as probably many readers noticed immediately, obviously inspired by Piaget's 1937/1954 book *"The Construction of Reality in the Child"*. In that book Piaget presented his theory of how children, in sensorimotor interaction with their environment, develop the concepts of space, time, objects and causality as a basic scaffold or conceptual framework which helps them to build a viable experiential 'reality' that fits environmental constraints. Von Glasersfeld's (1995) summarized his interpretation of Piaget's theories in his formulation of a *radical constructivism* (RC), whose basic principles are as follows:

- Knowledge is not passively received either through the senses or by way of communication;
- knowledge is actively built up by the cognizing subject.
- The function of cognition is adaptive, in the biological sense of the term, tending towards fit or viability;
- cognition serves the subject's organization of the experiential world, not the discovery of an objective ontological reality. (von Glasersfeld, 1995, p. 51)

This notion of RC is, at least at a first glance, largely compatible with much recent research in cognitive science, artificial intelligence (AI) and artificial life which is concerned with adaptive robots or autonomous agents and their construction of internal structures in the course of agent-environment interaction. Typically such learning processes are constrained by some fitness evaluation or feedback in the form of occasional reinforcement. Hence, in RC terms, these robots are actively building up their own knowledge (rather than being programmed) and they typically do so by constructing sensorimotor transformation knowledge rather than an internal mirror of some external reality.

However, this is a fairly recent development in AI research, whose general endeavor roughly might be characterized as the attempt to endow artefacts (computers, robots, etc.) with some of the mental and behavioral capacities of living organisms. In fact, since its inception in the mid-1950 most research in AI and cognitive science has been coined by cognitivism and the computer metaphor for mind, and in particular the objectivist notion of knowledge as recovery of an agent-independent external reality (cf. Stewart, 1996). The turn towards a (more) RC-compatible approach has been paralleled by the development of the notion of cognition as being *situated*. The concept of situatedness has since the mid-1980s been used extensively in the cognitive science and AI literature, in terms such as 'Situated Action' (Suchman, 1987), 'Situated Cognition' (e.g., Clancey, 1997), 'Situated AI' (e.g. Husbands *et al*., 1993), 'Situated Robotics' (e.g., Hallam and Malcolm, 1994), 'Situated Activity' (e.g., Hendriks-Jansen, 1996), and 'Situated Translation' (Risku, 2000). Roughly speaking, the characterization of an agent as 'situated' is usually intended to mean that its behavior and cognitive processes first and foremost are the outcome of a close coupling between agent and environment. Hence, situatedness is nowadays by many cognitive scientists and AI researchers considered a *conditio sine qua non* for any form of 'true' intelligence, natural or artificial. When it comes to the details of how situatedness and agent-environment interaction 'work', however, there are significantly different interpretations.

This paper aims to discuss in detail, from a constructivist perspective, different aspects/notions of situatedness throughout the history of AI research, and in particular the increasing focus on the role of constructive processes as the basis of situatedness. Furthermore, approaches to artificial or 'robotic situatedness' will be evaluated in the context of biologically based constructivist theories of the relevance of agent-environment interaction and constructive processes for 'organismic situatedness'. We start off in Section 2 with a summary and comparison of the constructivist theories of von Uexküll and Piaget, both of which will turn out to be relevant in the discussion of AI, in particular today's research on situated AI and adaptive robotics. Section 3 examines computationalism, symbolic AI and connectionism as well as their respective (in-) compatibilities with a constructivist perspective. Section 4 then discusses in detail the 'New AI' and its focus on situated and embodied

intelligence in robotic agents as well as its use of constructive processes at different levels and time scales. Section 5, finally, puts the 'new', situated AI and adaptive robotics into perspective by discussing its possibilities and limitations in the light of biologically based constructivist theories.

2. Constructivism: Von Uexküll and Piaget

Although Jakob von Uexküll and Jean Piaget probably had no contact with each other's work (cf. von Glasersfeld, 1995) there are a number of interesting similarities in their work. Both of them started off as biologists, and both were strongly inspired by Kant's insight that all knowledge is determined by the knower's subjective ways of perceiving and conceiving. In the introduction to the second edition of his *Critique of Pure Reason* Kant had pointed out:

> Until now one assumed that all cognition had to conform to objects … Henceforth one might try to find out whether we do not get further … if we assume that the objects have to conform to our cognition. (Kant, 1787) [1]

Thus, for example, space and time are, according to Kant, not aspects of an external reality, but they are the fundamental forms human cognition imposes on all experience. Hence, Kant distinguished between an object as it *appears* to us and the thing-in-itself (*'Ding an sich'*) of which we could have no certain knowledge, due to the fact that we can only access/experience it through our senses. Kant is sometimes considered to have had a strong influence on the cognitive science concept of representation. However, von Glasersfeld (1995) points out that this is, at least partly, due to an "unfortunate use" of the term 'representation' introduced by translators of German philosophy.

[1] We here use the Kant translations of von Glasersfeld (1995), who translates the German term 'Erkenntnis' as 'cognition'.

> It may have started earlier, but it became common usage in philosophy with the translation of Kant's *Critique of Pure Reason*. The two German words *Vorstellung* and *Darstellung* were rendered by one and the same English word 'representation'. To speakers of English this implies a reproduction, copy, or other structure that is in some way isomorphic with an original. This condition fits the second German word quite well, but it does not fit the first. *Vorstellung*, which is the word Kant uses throughout his work, should have been translated as 'presentation' … The element of autonomous construction is an essential part of the meaning of *Vorstellung.* If it is lost, one of the most important features of Kant's theory becomes incomprehensible. (von Glasersfeld, 1995, p. 94)

Kant's work strongly influenced both von Uexküll's (1928) and Piaget's (1954) work on the biological and psychological mechanisms underlying the construction of these concepts. In fact, von Uexküll (1928) considered it the "task of biology .. to expand the result of Kant's research" by investigating the role of the body and the relationship between subjects and their objects. Furthermore, both von Uexküll and Piaget were discontent with behaviorist theories which dominated the study of mind and behavior during the first half of the 20th century. According to von Uexküll, the main problem with these approaches was that they overlooked the organism's subjective nature, which integrates the organism's components into a purposeful whole. In his own words:

> The mechanists have pieced together the sensory and motor organs of animals, like so many parts of a machine, ignoring their real functions of perceiving and acting, and have gone on to mechanize man himself. According to the behaviorists, man's own sensations and will are mere appearance, to be considered, if at all, only as disturbing static. But we who still hold that our sense organs serve our perceptions, and our motor organs our actions, see in animals as well not only the mechanical structure, but also the operator, who is built into their organs as we are into our bodies. We no longer regard animals as mere machines, but as subjects whose essential activity consists of perceiving and acting. We thus unlock the gates that lead to other realms, for all that a subject perceives becomes his perceptual world and all that he does, his effector world. Perceptual and effector worlds together form a closed unit, the *Umwelt*. (von Uexküll, 1957, p. 6)

Von Uexküll (1957) used the example of the tick to illustrate his concept of *Umwelt* and his idea of the organism's embedding in its world through *functional circles* (see Figure 1). It is three such

functional circles in “well-planned succession” which coordinate the interaction of the tick as a subject (and *meaning-utilizer*) and a mammal as its object (and *meaning-carrier*):

(1) The tick typically hangs motionless on bush branches. When a mammal passes by closely its skin glands carry perceptual meaning for the tick: the perceptual signs of butyric acid are transformed into a perceptual cue which triggers effector signs which are sent to the legs and make them let go so the tick drops onto the mammal, which in turn triggers the effector cue of shock.

(2) The tactile cue of hitting the mammal’s hair makes the tick move around (to find the host’s skin).

(3) The sensation of the skin’s heat triggers the tick’s boring response (to drink the host’s blood).

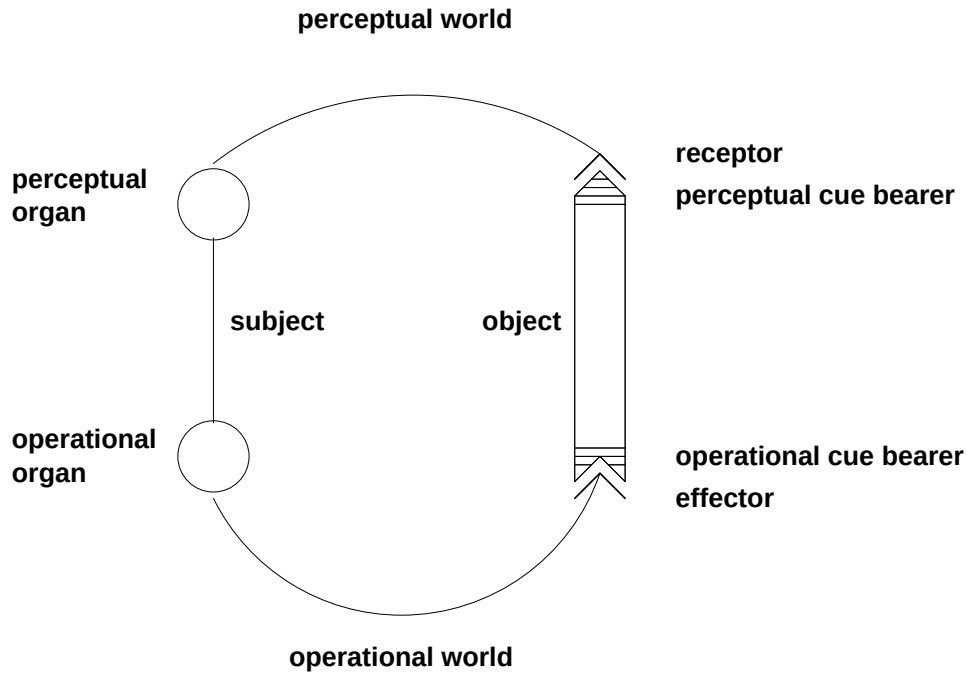

Figure 1: The functional circle according to Jakob von Uexküll. Adapted from von Uexküll (1957).

Von Uexküll did not deny the physical/chemical nature of the organism's components and processes. That means, his view should not, as sometimes done (e.g., Richards, 1987), be considered vitalistic (cf. Emmeche 1990, in press; T. von Uexküll, 1992; Langthaler, 1992). He 'admitted' that the tick exhibits "three successive reflexes" each of which is "elicited by objectively demonstrable physical or chemical stimuli". But he pointed out that the organism's components are forged together to form a coherent whole, i.e. a *subject*, that acts as a behavioral entity which, through functional embedding, forms a "systematic whole" with its Umwelt.

> We are not concerned with the chemical stimulus of butyric acid, any more than with the mechanical stimulus (released by the hairs), or the temperature stimulus of the skin. We are solely concerned with the fact that, out of the hundreds of stimuli radiating from the qualities of the mammal's body, only three become the bearers of receptor cues for the tick. ...What we are dealing with is not an exchange of forces between two objects, but the relations between a living subject and its object. (von Uexküll 1957, p. 11f.)

Closely related to von Uexküll's functional circle is Piaget's concept of '*action schemes*', which are based on the notion of reflexes, but not limited to mere stimulus-response mechanisms. Instead action schemes, e.g. the infant's rooting reflex, contain three elements (cf. von Glasersfeld, 1995, p. 65):

(1) the recognition of a certain situation, e.g. the infant's cheek being touched;

(2) a specific activity associated with this situation, e.g. the infant's turning its head towards the touched side in search for something to suck on;

(3) the expectation that the activity produces a certain (beneficial) result, e.g. finding the mother's breast and milk.

Hence, Piaget's concept of the action scheme is largely compatible with von Uexküll's concept of the functional circle. In both cases knowledge is viewed as tied to action, or as Piaget (1967) formulated it, "to know an object implies its incorporation in action schemes". Furthermore, in both theoretical frameworks the interaction of agent and environment is not conceived as mere stimulus-response, but as meaningfully organized through multiple behavior-guiding structures (functional circles and action schemes, respectively) which tie together an active, meaning-utilizing subject and its meaning-

carrying objects. Von Uexküll sometimes referred to the sign processes in the nervous system as a "mirrored world" (Uexküll, 1985; cf. also T. von Uexküll *et al.*, 1993), but also pointed out that by that he meant a "*counterworld*", i.e. an autonomously constructed *'Vorstellung'* in Kant's sense rather than a 1:1 reflection (*'Darstellung'*) of the external environment in the realist sense of a representation. Thus he wanted to emphasize that

> … in the nervous system the stimulus itself does not really appear but its place is taken by an entirely different process which has nothing at all to do with events in the outside world. This process can only serve as a *sign* which indicates that in the environment there is a stimulus which has hit the receptor but it does not give any evidence of the quality of the stimulus. (von Uexküll, 1909, p. 192)[2]

T. von Uexküll *et al.* (1993) also pointed out that von Uexküll's notion of 'counterworld' should not be equated with a 'mirror' in the narrow sense of a reflection of the environment. They further elaborated that

> … in this phenomenal universe [of the counterworld], the objects of the environment are represented by schemata which are not, as in a mirror, products of the environment, but rather 'tools of the brain' ready to come into operation if the appropriate stimuli are present in the outside world. In these schemata, sensory and motor processes are combined … to form complex programs controlling the meaning-utilizing … behavioural responses. They are retrieved when the sense organs have to attribute semiotic meanings to stimuli. (T. von Uexküll *et al.* 1993, p. 34)

In a similar vein Merleau-Ponty (1962, 1963) argued that organisms do not interact with the objective world in-itself, but with their subjective perception of it (cf. Loren and Dietrich, 1997). In his *Phenomenology of Perception* he characterized the subjective and situation-dependent nature of behavior as follows:

[2] We here use the translation given by T. von Uexküll *et al.* (1993), who translate the original German term "Zeichen" as "sign", rather than "token" as in the earlier translation provided in von Uexküll (1985).

> In fact the reflexes themselves are never blind processes: they adjust themselves to a 'direction' of the situation, and express our orientation towards a 'behavioural setting' … It is this global presence of the situation which gives meaning to the partial stimuli and causes them to acquire importance, value or existence. The reflex does not arise from stimuli but moves back towards them, and invests them with a meaning which they do not possess taken singly as psychological agents, but only when taken as a situation. It causes them to exist as a situation, it stands in a 'cognitive' relation to them, which means that it shows them up as that which it is destined to confront. (Merleau-Ponty, 1962, p. 79)

Since this paper is primarily concerned with AI and the situatedness (or lack thereof) of artifacts (robots, computers, etc.) in particular, it is interesting to note that von Uexküll (1928) considered the *autonomy* of the living as the key difference between mechanisms and organisms. Following the work of Müller (1840), he pointed out that "each living tissue differs from all machines in that it possesses a 'specific' life-energy in addition to physical energy" (von Uexküll, 1982, p. 34). This allows it to react to different stimuli with a 'self-specific' activity according to its own "ego-quality" (*Ich-Ton*), e.g., a muscle with contraction or the optic nerve with sensation of light. Hence, each living cell *perceives* and *acts*, according to its specific perceptual or receptor signs and impulses or effector signs, and thus the organism's behaviors "are not mechanically regulated, but meaningfully organized" (von Uexküll, 1982, p. 26). The operation of a machine, on the other hand, is purely mechanical and follows only the physical and chemical laws of cause and effect. Furthermore, von Uexküll (1928, p. 180)[3] referred to Driesch who pointed out that all action is a mapping between individual stimuli and effects, depending on a historically created basis of reaction (*Reaktionsbasis*), i.e. a context-dependent behavioral disposition (cf. Driesch, 1931). Mechanisms, on the other hand, do not have such a historical basis of reaction, which, according to von Uexküll, can only be grown - and there is no growth in machines. Von Uexküll (1928, p. 217) further elaborated that the rules machines follow are not capable of adaptation. This is due to the fact that machines are fixed structures, and the rules that guide their operation, are not their 'own' but human rules, which have been built into the machine, and therefore also can be changed only by humans. Hence, mechanisms are *heteronomous* (cf. also T. von Uexküll, 1992). Machines can therefore, when they get damaged,

[3] Unless noted otherwise, all translations from German sources have been carried out by the author.

not repair or regenerate themselves. Living organisms, on the other hand, can, because they contain their functional rule (*Funktionsregel*) themselves, and they have the protoplasmic material, which the functional rule can use to fix the damage autonomously. This can be summarized by saying that *machines act according to plans* (their human designers'), whereas *living organisms are acting plans* (von Uexküll 1928, p. 301).

This notion of autonomy is also closely related to what von Uexküll (1982) described as the "principal difference between the construction of a mechanism and a living organism", namely that "the organs of living beings have an innate meaning-quality, in contrast to the parts of machine; therefore they can only develop centrifugally":

> Every machine, a pocket watch for example, is always constructed centripetally. In other words, the individual parts of the watch, such as its hands, springs, wheels, and cogs, must always be produced first, so that they may be added to a common centerpiece.
>
> In contrast, the construction of an animal, for example, a triton, always starts centrifugally from a single cell, which first develops into a gastrula, and then into more and more new organ buds.
>
> In both cases, the transformation underlies a plan: the 'watch-plan' proceeds centripetally and the 'triton-plan' centrifugally. Two completely opposite principles govern the joining of the parts of the two objects. (von Uexküll, 1982, p. 40)

The concept of (autonomous) adaptation in interaction with an environment was also central to Piaget's theory which viewed "cognition as an instrument of adaptation, as a tool for fitting ourselves into the world of our experiences" (von Glasersfeld, 1995, p. 14). This is achieved through (a) the *assimilation* of new experiences into existing structures, and (b) the *accommodation* of these structures, i.e. adaptation of existing ones and/or the creation of new ones. The latter, learning through accomodation, occurs for the purpose of 'conceptual equilibration', i.e. the elimination of perturbations through mismatches between the agent's conceptual structures and expectations on the one hand, and its experience on the other hand. Piaget thus "relinquished the notion of cognition as the producer of representations of an ontological reality, and replaced it with cognition as an instrument of adaptation the purpose of which is the construction of viable conceptual structures"

(von Glasersfeld, 1995, p. 59). Accordingly, in the constructivist framework "the concept of viability in the domain of experience, takes the place of the traditional philosopher's concept of Truth, that was to indicate a 'correct' representation of reality" (von Glasersfeld, 1995, p. 14).

Hereafter we will mostly use the term 'cognition' in Piaget's and von Glasersfeld's sense of organizing an agent's sensorimotor experience and interaction with its environment, thus serving its adaptation tending toward 'viability', i.e. fit with environmental constraints. This view will be referred to as *'interactive cognition'* to distinguish it from the traditional cognitive science notion of cognition as agent-internal processing of explicit representations (cf. Section 3.1). This should, however, not be misunderstood as saying that constructivist theories only cover 'low-level', sensorimotor cognition. As pointed out in detail by Stewart (1996), this is not at all the case. We will see later that the interactive notion of cognition is largely compatible with modern CS and AI notions of situated and embodied intelligence (cf. Section 4) as well as modern theories of the biology of cognition (Maturana and Varela, 1980, 1987; cf. Section 5.3).

3. Computationalist AI

3.1 Cognitivism and Symbolic AI

During the 1940s and 1950s a growing number of researchers, like von Uexküll and Piaget, discontent with behaviorism and mechanistic theories as the predominant paradigm in the study of mind and behavior, became interested in the mind's internal processes and representations, whose study behaviorists had rejected as being unscientific. Craik, in his 1943 book, *The Nature of Explanation*, was perhaps the first to suggest that organisms make use of explicit knowledge or world models, i.e. internal representations of the external world:

> If the organism carries a "small-scale model" of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it. (Craik, 1943)

That means, in Craik's view the organism is not just physically situated in its environment, but it also has its own internal model of it, which allows it to deal with that external reality in a more effective manner. Craik had little to say about the exact form of the internal representations or the processes manipulating them (cf. Johnson-Laird, 1989). However, he elaborated that by a 'model' he meant "any physical or chemical system which has a similar relation-structure" and "works in the same way as the processes it parallels" (Craik, 1943). Hence, his notion of such an internal model was much closer to a 'mirror' of external reality than von Uexküll's notion of a 'counterworld'.

As a result of the increasing use and understanding of computer technology during the 1940s and 50s, researchers began to realize the information processing capabilities of computers and liken them to those of humans. Taken to extremes, this analogy echoes the *computer metaphor for mind*, one of the central tenets of *cognitivism* and traditional AI, which considers cognition to be much like a computer program that could be run on any machine capable of running it. Pfeifer and Scheier summarize the functionalist view (Putnam, 1975) as follows:

> Functionalism … means that thinking and other intelligent functions need not be carried out by means of the same machinery in order to reflect the same kinds of processes; in fact, the machinery could be made of Emmental cheese, so long as it can perform the functions required. In other words, intelligence or cognition can be studied at the level of algorithms or computational processes without having to consider the underlying structure of the device on which the algorithm is performed. From the functionalist position it follows that there is a distinction between hardware and software: What we are interested in is the software or the program. (Pfeifer and Scheier, 1999, p. 43)

Neisser (1967), in his book *Cognitive Psychology*, which defined the field, also stressed that the cognitive psychologist "wants to understand the program, not the hardware". Thus earlier theories, including those of von Uexküll and Piaget, on the interaction between organisms and their

environments were divorced from the dominant themes in the mind sciences. Combining Craik's idea of the organism carrying a "small-scale model" "within its head" with the functionalist view that the essence of cognition and intelligent behavior was to be sought in body-independent computation, traditional AI from then on basically completely neglected both organism (body) and reality. Accordingly, research in CS and AI focused on what von Uexküll (1957) referred to as the "inner world of the subject". The cognitivist view, however, is that this 'inner world' consists of an internal model of a pre-given 'external reality', i.e. representations (in particular symbols) corresponding or referring to external objects ('knowledge'), and the computational, i.e. formally defined and implementation-independent, processes operating on these representations ('thought'). That means, like von Uexküll's theory, cognitivism was strictly opposed to behaviorism and emphasized the importance of the subject's 'inner world', but completely unlike von Uexküll and Piaget it de-emphasized, and in fact most of the time completely ignored, the environmental embedding through functional circles or action schemes. That means, issues like situatedness, agent-environment interaction and the autonomous construction of representations were for a long time simply largely ignored.

The most prominent example of this cognitivist view is the so-called *Physical Symbol System Hypothesis* (PSSH) (Newell and Simon, 1976) which characterizes the approach of traditional AI as dedicated to the view of intelligence as symbol manipulation. The PSSH states that symbol systems, realized in *some* physical medium, have the necessary and sufficient means for intelligent action. Pfeifer and Scheier (1999) further elaborate this view as follows:

> Computational processes operate on representations, the *symbol structures*. A (symbolic) "representation" [see Figure 3] in the sense that Newell and Simon mean it refers to a situation in the outside world and obeys the "law of representation," namely
>
> $$\mathrm{decode}[\mathrm{encode}(T)(\mathrm{encode}(X_1))] = T(X_1)$$
>
> where X_1 is the original situation and T is the external transformation (Newell 1990, p. 59). There is an encoding as well as a decoding function for establishing a mapping between the outside world and the internal representation. (Pfeifer and Scheier, 1999, p. 44)

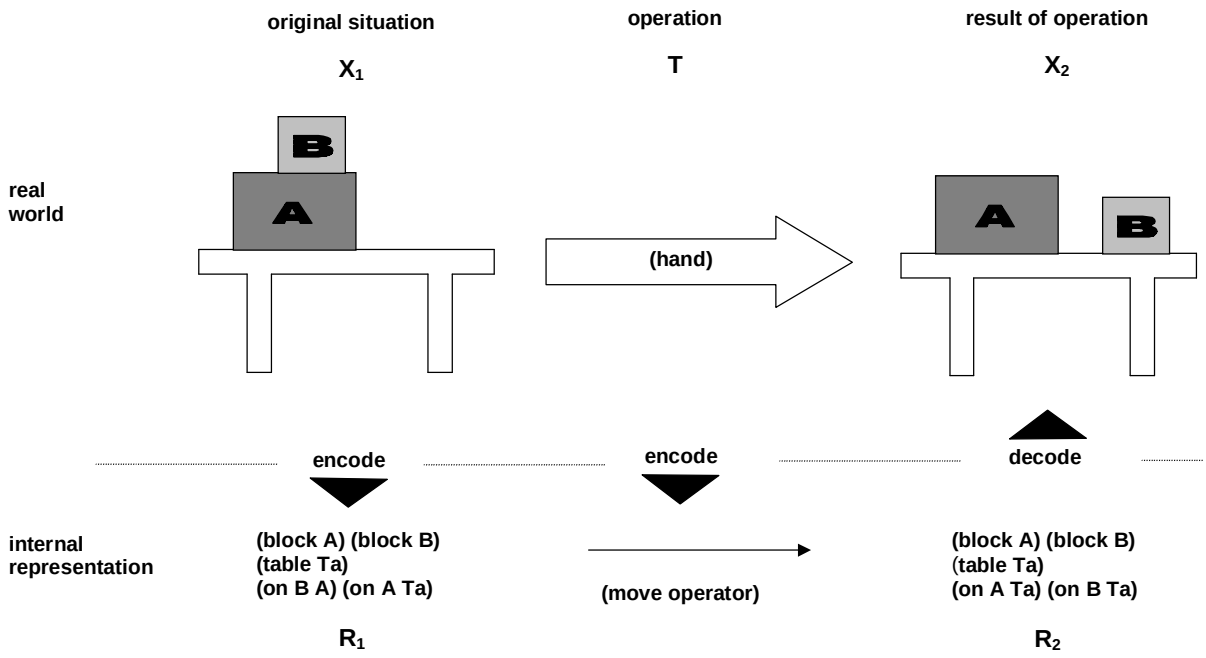

Figure 2: The law of representation (adapted from Pfeifer and Scheier, 1999, p. 45). X_1, the original situation in the real world is mapped onto internal representation R_1. *T*, the operator that moves *B* on the table is mapped onto an internal representation as well – the 'move operator'. When the move has been carried out in the real world and on the internal representation, the resulting real world situation X_2 and the decoding of the resulting internal representation R_2 should be identical.

A number of symbolic knowledge representation schemes were developed, in particular during the 1970s, including frames (Minsky, 1975), semantic networks (e.g., Woods, 1975) and scripts (Schank and Abelson, 1977). We will here only look at the latter in some detail and use it as an example, which will help to illustrate and analyse the problems with symbolic knowledge representation in general in Section 3.2.

Schank and Abelson's (1977) research aimed to enable computers to *understand* natural language stories. The approach was to provide the computer with *scripts*, i.e. symbolic representations of the essence of stereotypical social activities, such as 'going to a restaurant'. The computer should then be able to use its scripts to understand simple stories. A very simple example of a story might be "John went to a fancy restaurant. He ordered a steak. Later he paid and went home.". The computer's

understanding could then be tested by asking it questions such as “Did John sit down?” or “Did John eat?”. In both cases the answer should, of course, be “Yes”, because it is common-sense knowledge that people sit down in restaurants (at least fancy ones), and that they only pay for food they actually have received and eaten.

Scripts were written in an event description language called ‘Conceptual Dependency’ (Schank, 1972) consisting about a dozen primitive, context-free acts, including:

- PTRANS – a physical transfer of an object
- ATRANS – an abstract transfer of, for example, possession, control, etc.
- MBUILD – ‘building’ something mentally; e.g., making a decision

Hence part of a script for going to a restaurant, for example, could look like this (Schank, 1975b):

> Script: restaurant.
> Roles: customer; waitress; chef; cashier.
> Reason: to get food so as to go down in hunger and up in pleasure.
> Scene 1, entering:
> PTRANS – go into restaurant
> MBUILD – find table
> PTRANS – go to table
> MOVE – sit down
> Scene 2, ordering:
> ATRANS – receive menu
> … (Schank, 1975, p. 131)

3.2 Critiques of Computationalism and Symbolic AI

Towards the end of the 1970s traditional AI came under heavy attack from several directions. The most prominent critics were Dreyfus (1979) and Searle (1980), who criticized symbolic AI, and computationalism in general, from different angles. Both their attacks, however, address the issues of

situatedness and embodiment, and the lack thereof in traditional AI, which is why we discuss both of them in some detail in the following two subsections.

3.2.1 Dreyfus' Critique of Explicit Representation

Hubert Dreyfus, in his 1979 book (second edition), *What Computers Can't Do: A Critique of Artificial Reason*, strongly questioned traditional AI's use of explicit, symbolic representations and its focus on limited, isolated domains of human knowledge, such as 'restaurant going', which he referred to as '*micro-worlds'*. He argued that the resulting AI programs represented descriptions of isolated bits of human knowledge "from the outside", but that the programs themselves could never be *"situated"* in any of these descriptions. Basing much of his argument on Heidegger (1962) and his notion of '*equipment*', Dreyfus argued that even simple everyday objects such as chairs could not be defined explicitly (and thus could not be represented to a computer in symbolic form). His argument is thus largely compatible with von Uexküll and Piaget's view of knowledge as tied to action and the constructivist distinction between the physical object in-itself and the meaning attributed to it by a subject (which makes it a semiotic object, i.e. part of an agent's phenomenal *Umwelt*). In Dreyfus' words:

> No piece of equipment makes sense by itself, the physical object which is a chair can be defined in isolation as a collection of atoms, or of wood or metal components, but such a description will not enable us to pick out chairs. What makes an object a *chair* is its function, and what makes possible its role as equipment for sitting is its place in a total practical context. This presupposes certain facts about human beings (fatigue, the way the body bends), and a network of other culturally determined equipment (tables, floors, lamps) and skills (eating, writing, …). Chairs would not be equipment for sitting if our knees bent backwards like those of flamingos, of if we had no tables, as in traditional Japan or the Australian bush. (Dreyfus, 1979)

Commenting on Minsky's (1975) argument that chairs could be identified using certain context-free features (which, however, he left unspecified), Dreyfus pointed out:

> There is no argument why we should expect to find elementary context-free *features* characterizing a chair *type*, nor any suggestion as to what these features might be. They certainly cannot be legs, back, seat, and so on, since these are not context-free characteristics defined apart from chairs which then "cluster" in a chair representation; rather, legs, back, and the rest, come in all shapes and variety and can only be recognized as *aspects* of already recognized chairs. (Dreyfus, 1979)

According to Dreyfus, the "totally unjustified" belief that micro-worlds (such as knowledge about chairs or restaurant-going) could be studied in relative isolation from the rest of human knowledge is based on a "naive transfer" to AI of methods from the natural sciences. An example of such a transfer is Winograd's (1976) characterization of AI research on knowledge representation:

> We are concerned with developing a formalism, or "representation", with which to describe … knowledge. We seek the "atoms" and "particles" of which it is built, and the "forces" that act on it. (Winograd, 1976, p. 9)

In the natural sciences such an approach is valid, Dreyfus argued, due to the fact that many phenomena are indeed the result of "lawlike relations of a set of primitive elements". However, the "sub-worlds" that humans are involved in in their everyday life, such as the 'worlds' of business, of restaurant-going, or of chairs, are not context-free "structural primitives". Hence, they do not compose like building-blocks, but each of them is a "mode [or variation] of our shared everyday world". That means, different domains of human knowledge "are not related like isolable physical systems to larger systems they *compose*; they are local elaborations of a whole which they *presuppose*" (Dreyfus, 1979). The reader should notice the resemblance of this argument concerning the parts and wholes of (human) knowledge to von Uexküll's argument concerning the centrifugal 'construction' of living organisms (cf. Section 2). In both cases the 'parts' presuppose the 'whole', rather than the other way round as in most man-made artifacts.

More specifically, with respect to Schank's above restaurant script, Dreyfus argued that the program, even if it can answer the questions "Did John sit down?" and "Did John eat?", can only do so because what normally happens in a restaurant has been "preselected and formalised by the programmer as default assignments". The situation's background, however, has been left out, such that "a program

using a script cannot be said to understand going to a restaurant at all". This lack of 'true understanding' is revealed as soon as we ask the program non-standard questions such as whether or not the waitress wore clothes, or whether she walked forward or backward – questions it could not possibly answer based on the script alone. Cognitivists could of course rightly argue that scripts were never meant to encode the *whole* background, including common-sense knowledge about clothes, walking, gravity, etc., anyway. Dreyfus' argument is, however, not to be understood as a critique of scripts as such, but as an argument against the explicit style of representation used in symbolic knowledge representation schemes such as scripts. An attempt to prove Dreyfus and other AI critics wrong in this point is the CYC project, started in 1984 (Lenat and Feigenbaum, 1991). This project's ambitious goal is to explicitly formalize human common-sense knowledge, i.e. "the millions of abstractions, models, facts, rules of thumb, representations, etc., that we all possess and that we assume everyone else does" (Lenat and Feigenbaum, 1991, p. 216). Although the project was initially intended as a 10-year project 'only', it has so far failed to convince AI critics that its goal could ever be achieved (e.g., Clark, 1997). However, back to Dreyfus' (1979) original argument; his radical conclusion was that "since intelligence must be situated it cannot be separated from the rest of human life". That 'rest', however, includes bodily skills and cultural practices, which, according to Dreyfus, could not possibly represented explicitly and thus could not be formalized in a computer program.

Dreyfus explanation of human situatedness and why a traditional AI system, i.e. a formally defined computer program, could not possibly have it, is worth quoting at length, because (a) it is to our knowledge the first detailed discussion of situatedness (under that name) in an AI context, and (b) it still today is highly relevant to recent work in situated and embodied AI, which in a sense aims to situate artificial *intelligence* by grounding it in artificial *life* (cf. Section 4).

> Humans … are, as Heidegger puts it, *always already* in a situation, which they constantly revise. If we look at it genetically, this is no mystery. We can see that human beings are gradually trained into their cultural situation on the basis of their embodied pre-cultural situation … But for this very reason a program … is *not* always-already-in-a-situation. Even if it represents all human knowledge in its stereotypes, including all possible types of human situations, it represents them from the outside … It isn't situated *in* any one of them, and it may be impossible to program it to behave as if it were.

> … Is the know-how that enables human beings to constantly sense what specific situation they are in the sort of know-how that can be represented as a kind of knowledge in *any* knowledge-representation language no matter how ingenious and complex? It seems that our sense of our situation is determined by our changing moods, by our current concerns and projects, by our long-range self-interpretation and probably also by our sensory-motor skills for coping with objects and people – skills we develop by practice without ever having to represent to ourselves our body as an object, our culture as a set of beliefs, or our propensities as situation-action rules. All these uniquely human capacities provide a "richness" or a "thickness" to our way of being-in-the-world and thus seem to play an essential role in situatedness, which in turn underlies all intelligent behavior. (Dreyfus, 1979)

3.2.2 Searle's Critique of Computationalism

John Searle (1980) also used Schank's work on script as an example in order to question in general the computationalist nature of traditional AI. Moreover, he suggested to distinguish between the position of *weak* or *cautious AI* which sees the computer as a powerful tool in the study of mind (a position he agreed with), and that of *strong AI* which would hold that "the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states".

Searle went on to present his now famous *Chinese Room Argument* (CRA) which has haunted strong AI researchers ever since – a thought experiment intended to answer the question to what extent, if any, a computer running Schank's script could be said to *understand* a natural language story. The CRA goes approximately like this: Suppose you, knowing no Chinese at all, are locked in a room. Under the door you are passed a first and a second batch of Chinese writing. With you in the room you have a set of rules (in English) for relating the first to the second set of symbols. You further receive a third batch of Chinese writing together with English instructions, which allow you to correlate elements of the batches, and instruct you how to give back Chinese symbols in response to the third batch. Now, the crucial question is, do you *understand* Chinese in the sense that you actually know what any of the symbols mean? The obvious answer, Searle argued, is that you do not. Suppose the people outside the room, call the first batch a 'script', the second one a 'story', the rules

‘a program’, the third batch ‘questions’, and your responses ‘answers’. Further, suppose the ‘program’ is very good; then your ‘answers’ might be indistinguishable from those of a native speaker of Chinese. That means, the point here is that, although from outside the room you might be considered to understand, obviously everybody who knows what goes on inside the room realizes that you are just “manipulating uninterpreted formal symbols”. Furthermore, Searle concluded, since you, inside the room, are “simply an instantiation of the computer program”, any computer using the same script, or any other purely formally defined system for that matter, would have to be said to understand as much of what it processes as you understand Chinese; namely nothing at all.

The reason for this lack of understanding in the computer’s case, Searle elaborated, is that, due to the fact that there are no causal connections between the internal symbols and the external world they are supposed to represent, purely computational AI systems lack *intentionality*[4]. In other words, traditional AI systems do not have the capacity to relate their internal processes and representations to the external world. In semiotic terms, what AI researchers intended was that the AI system, just like humans or other organisms, would be the interpreter in a triadic structure of sign (internal representation/symbol), external object and interpreter. What they missed out on, however, was that the interpreter could not possibly be the AI system itself. This is due to the fact that, in von Uexküll’s terms, the “inner world of the subject” was completely cut off from the external world by traditional AI’s complete disregard for any environmental embedding through receptors and effectors. Hence, as illustrated in Figure 4, the connection or mapping between internal representational domain and the external represented world is really just in the eye (or better: the mind) of the designer or other observers.

[4] This is of course not undisputed; for a symbolist account of intentionality see (Fodor, 1987).

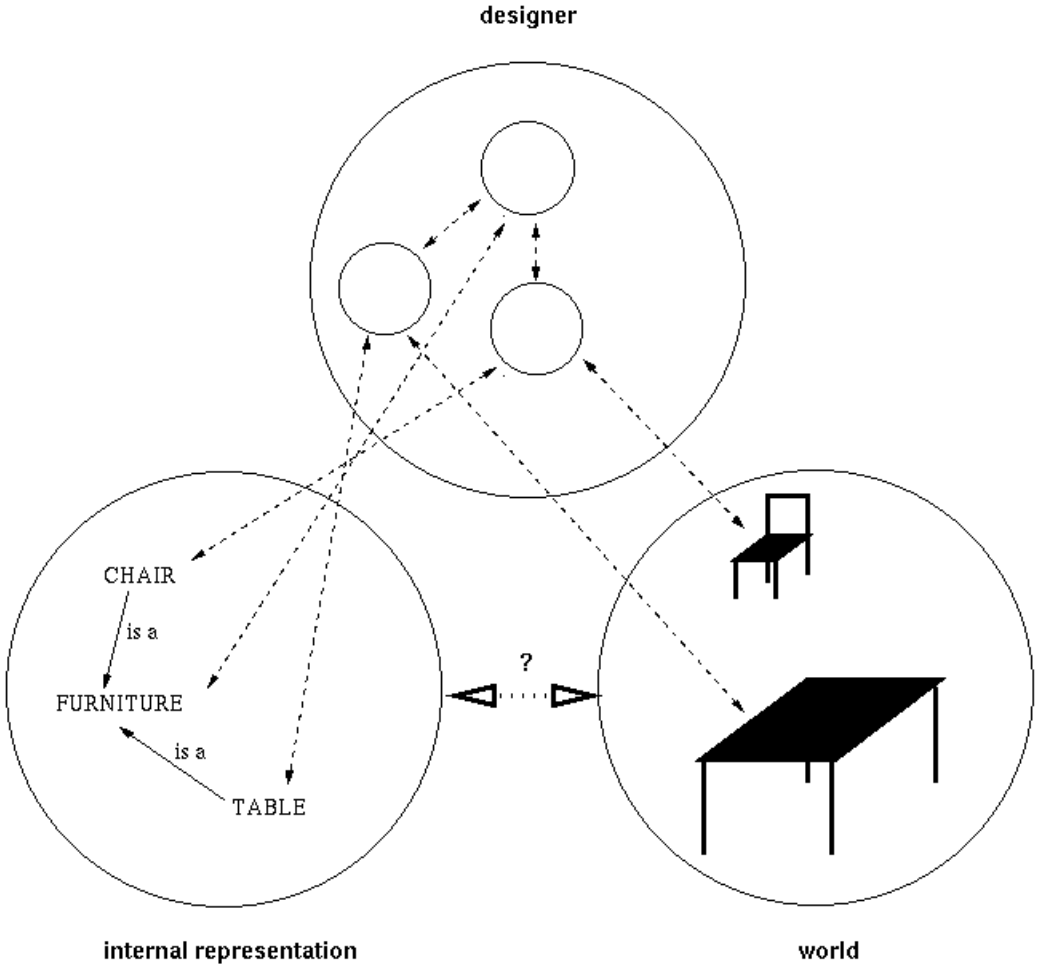

Figure 3: "What 'really' happens in traditional AI representation" (Dorffner, 1997). There are direct mappings between objects in the world and the designer's own internal concepts, and between the designer's concepts and their counterparts in the AI system's representational domain. There is, however, no direct, designer-independent, connection between the AI system and the world it is supposed to represent. Hence, the AI system lacks 'first hand semantics'. Adapted from Dorffner (1997).

It should be noted that Searle himself, contrary to common misinterpretations of his argument, did not suggest that the idea of intelligent machines would have to be abandoned. In fact he argued that humans are such machines and that the main reason for the failure of strong (traditional) AI was that it is concerned with *computer programs*, but "has nothing to tell us about *machines*" (Searle 1980), i.e. physical systems situated in and causally connected to their environments. That means, instead of accusing AI to be materialistic/mechanistic (for its belief that (man-made) machines, could be

intelligent), Searle actually accused AI of dualism, for its belief that disembodied, i.e. body-less and body-independent, computer programs could be intelligent. Hence, his conclusion was that AI research, instead of focusing on purely formally defined computer programs, should be working with physical machines equipped with (some of) the causal powers of living brains/organisms, including perception, action and learning, i.e. the capacity for autonomous construction of an own view of the world. In fact, as we will see in Section 4, that is approximately what modern AI, in particular work in adaptive robotics, does - it focuses on robots, i.e. physical systems, which 'perceive', 'act' and 'learn' (by artificial means) in interaction with the environment they are situated in.

Long before this type of research got started, Searle (1980) himself in fact formulated a possible "robot reply", which argued that putting traditional AI systems into robots would provide them with (some of) the causal powers he had claimed missing in purely computational, disembodied computer programs of traditional AI. He did, however, reject that reply, arguing that it could not possibly make any difference to the person in the Chinese room, if, unknown to that person, some of the incoming symbols came from a robot's sensors and some of the outgoing symbols controlled its motors. We will get back to this argument in Section 5.4 and evaluate to what extent it applies, twenty years later, to contemporary AI research.

3.3 Connectionism

3.3.1 Basics

A standard connectionist network, or (artificial) neural network (ANN), is a network of a (possibly large) number of simple computational units, typically organized in layers (cf. Figure 5). Each unit (or artificial neuron) usually receives a number of numerical inputs from other units it is connected to, calculates from the weighted sum of the input values its own numerical output value according to some activation function, and passes that value on as input to other neurons. The feature of ANNs that allows them to learn functional mappings from examples is the fact that each connection between two units carries a weight, a numerical value itself, that modulates the signal/value sent from

one neuron to the other. By weakening or strengthening of the connection weight, the signal flow between individual neurons can be adapted, and through coordination of the individual weight changes the network's overall mapping from input to output can be learned from examples.

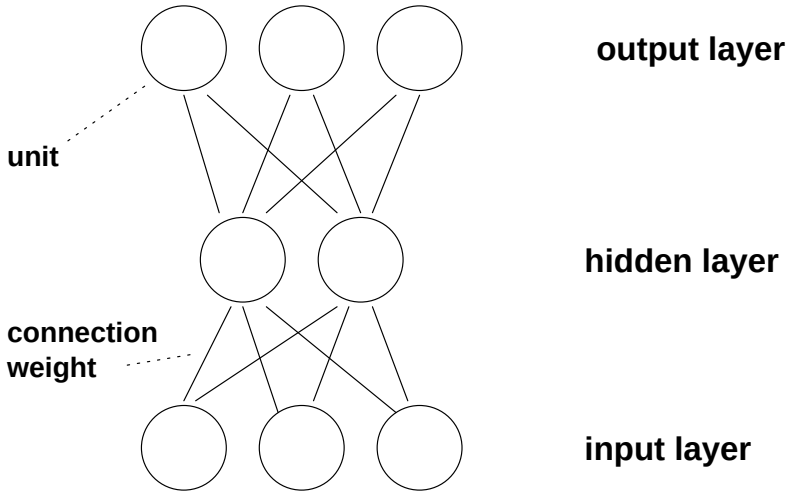

Figure 4: A typical feed-forward artificial neural network (ANN). Each circle represents a unit (or artificial neuron), and each solid line represents a connection weight between two units. Activation in this type of ANN is fed forward only, i.e. from input layer via a hidden layer to the output layer.

A number of learning techniques and algorithms have been applied to training ANNs, which vary in the degree of feedback they provide and the degree of self-organization that they require from the network. During *supervised learning* ANNs are provided with input values and correct target outputs in every time step. That means, the network is instructed on which inputs to use and which output signals to produce, but how to coordinate the signal flow in between input and output is up to the network's self-organization. Hence, internal representations (weights and/or hidden unit activations, cf. Sharkey, 1991) could be considered to be signs (or their modulators) private to the network and often opaque to outside observers. Thus, unlike traditional AI, connectionists do not promote symbolic representations that mirror a pre-given external reality. Rather, they stress self-organization of an adaptive flow of signals between simple processing units in interaction with an environment, which is compatible with an interactivist view of representation (Bickhard and Terveen, 1995;

Dorffner, 1997). Connectionism thus offers an alternative approach to the study of cognitive representation and sign use. In particular the parallel and distributed nature of weight and unit representations in ANNs, and the fact that these representations can be constructed from 'experience', i.e. in interaction with an environment, make connectionism largely compatible with the constructivist view of adaptation driven by the need to fit environmental constraints.

However, in most traditional connectionist work the 'environment' is still reduced to input and output values provided/interpreted by human designers/observers (cf. Lakoff, 1988; Manteuffel, 1992 Clark, 1997; Dorffner, 1997). That means, networks are not, like real nervous systems, embedded in the context of an (artificial) organism and its environment. Thus, although in a technically different fashion, most connectionists are, like cognitivists, mainly concerned with explaining cognitive phenomena as separated from organism-world interaction. Hence, much connectionist research focuses on the modeling of isolated cognitive capacities, such as the transformation of English verbs from the present to the past tense (Rumelhart and McClelland, 1986) or the prediction of letters or words in sequences (Elman, 1990), i.e. 'micro-worlds' in Dreyfus' (1979) sense (cf. Section 3.2.1). In von Uexküll's terms: Most connectionist research is only concerned with the self-organization of the subject-internal part of the functional circle (where input units might be roughly likened to receptors and output units to effectors). Or in Piaget's terms: Knowledge is not at all tied to action. Making the connection between inputs, outputs and internal representations and the objects they are supposed to represent, is again left to the mind of the observer, similar to the situation illustrated in Figure 3.

3.3.2 Recurrent Connectionist Networks

As long as we are using a feed-forward network, i.e. a network in which activation is only passed in one direction, the mapping from input to output will always be the same (given that the network has already learned and does not modify its connection weights anymore). Hence, the network will be a 'trivial machine', i.e. independent of past or input history same inputs will always be mapped to same outputs. However, if we add internal feedback through recurrent connections to the network it

becomes a ‘non-trivial’ machine. We can roughly distinguish between first-order feedback, i.e. the re-use of previous neural activation values as extra-inputs (e.g., in Elman’s (1990) Simple Recurrent Network), and higher-order feedback, i.e. the dynamical adaptation/modulation of the connection weights embodying the input-output mapping (e.g., in Pollack’s (1991) Sequential Cascaded Network). In both cases the mapping from input to output will vary with the network’s internal state, and thus the machine, depending on its past, can effectively be a ‘different’ machine in each time step. For the network itself this means that it no longer merely reacts to ‘external’ stimuli, but it ‘interprets’ inputs according to its own internal state. Or in von Uexküll’s (1982) terms, the network dynamically constructs a ‘historical basin of reaction’, which allows it to imprint its ‘ego-quality’ on incoming stimuli. That means, here the functional circle(s) realized by the recurrent network, and thus the ‘meaning’ it attributes to inputs, do actually vary with time, not completely unlike the varying level of hunger effects the meaning a piece of food has for an animal.

Recurrent connectionist networks play an important role in the study and modelling of cognitive representation and their construction. This is due to the fact that they account for both the (long-term) representation of learning experience in connection weights as well as the (short-term) representation of the controlled agent’s current context or immediate past in the form of internal feedback. Peschl (1997) has pointed out that RNNs, like real nervous systems, are “structure determined” (cf. also Maturana and Varela, 1980), which means that their reaction to environmental stimuli always depends on the system’s current state (or structure), and thus is never determined by the input alone. Peschl referred to this as the “*autonomy* of a representational system”. He further argued that in such recurrent networks traditional concepts of knowledge representation (as a ‘mirror’ of external reality) are not applicable due to the fact that there is “no stable representational relationship of reference”. Hence, the “goal of representation” in such systems, he argued, could not be to achieve an accurate mapping of an external environment to internal referential representations. Instead, recurrent neural systems should be viewed as “physical dynamical devices embodying the (transformation) knowledge for sensorimotor [input-output] integration and generating adequate behavior enabling the organism’s survival”. Thus, Peschl’s view is largely compatible with the earlier discussed

constructivist view of knowledge as tied to action, i.e. knowledge as mechanisms of adequate sensorimotor transformation. This is particularly clear in his characterisation of knowledge as "representation without representations":

> The internal structures do not map the environmental structures; they are rather responsible for generating functionally fitting behaviour which is *triggered* and *modulated* by the environment and *determined* by the internal structure (… of the synaptic weights). It is the result of *adaptive* phylo- and ontogenetic processes which have changed the architecture over generations and/or via learning in an individual organism in such a way that its physical structure embodies the dynamics for maintaining a state of equilibrium/homeostasis. (Peschl, 1997)

Aware of the limitations of disembodied neural networks, Peschl further suggested a *system relative* concept of representation as "determined not only by the environment", but also highly dependent on "the organization, structure, and constraints of the representation system as well as the sensory/motor systems which are embedded in a particular body structure".

3.3.3 Searle and Dreyfus on Connectionist Networks

Dreyfus' and Searle's original criticisms of AI (cf. Section 3.2) were formulated before the re-emergence of connectionism in the mid-1980s. Thus, at the time they were mostly concerned with symbolic AI, such as the work of Schank and others on symbolic knowledge representation. Searle (1990), whose original argument was mostly concerned with the purely computational nature of traditional AI systems, has pointed out that his argument "applies to any computational system", including connectionist networks. He illustrated this with a variation of the CRA, this time replacing the Chinese room with a "Chinese gym", i.e. "a hall of many monolingual, English-speaking men. These men would carry out the same operations as the nodes and synapses in a connectionist architecture … and the outcome would be the same as having one man manipulate symbols according to a rule book". Still, although the people in the gym operate differently from the person in the room in the original CRA, according to Searle, obviously "[n]o one in the gym speaks a word of Chinese, and there is no way for the system as a whole to learn the meaning of any Chinese words."

Hence, to him the use of connectionist networks as such, without embedding in body and environment, does not at all solve the problems of computationalism. In his own words: “You can’t get semantically loaded thought contents from formal computations alone, whether they are done in serial or in parallel; that is why the CRA refutes strong AI in any form.” (Searle, 1990). In a slightly later paper, however, Searle (1991, p. 594) acknowledges as one of the “merits” of connectionism that “at least some connectionist models show how a system might convert a meaningful input into a meaningful output without any rules, principles, inferences, or other sorts of meaningful phenomena in between”.

Similarly, Dreyfus (1996), whose original criticism had been concerned mostly with the explicit nature of traditional AI representation, pointed out that connectionist networks must be considered a powerful alternative to symbolic knowledge representation. This is due to the fact that they “provide a model of how the past can affect present perception and action without needing to store specific memories at all”. Of particular interest, he argued, similar to Peschl (cf. Section 3.3.2), are recurrent connectionist networks, which he referred to as “the most sophisticated neural networks”. The reader might recall that Dreyfus (1979) argued that the problem with traditional AI’s explicit representations was that they were not situated, whereas humans did not have that problem since they are “always already in a situation, which they constantly revise”. Similarly, Dreyfus (1996) describes the working of recurrent connectionist networks (apparently with an SRN-like network in mind): “The hidden nodes of the most sophisticated networks are always already in a particular state of activation when input stimuli are received, and the output that the network produces depends on this initial activation.” Hence, recurrent neural mechanisms provide an agent with the means to actively (re-) construct its own current situation in the sense that it dynamically adapts its behavioral disposition and thus its way of attributing meaning to incoming stimuli (cf. Ziemke, 1999a; Ziemke and Sharkey, in press). In Dreyfus’ words:

> If the input corresponds to the experience of the current situation, the particular prior activation of the hidden nodes which is determined by inputs leading up to the current situation might be said to correspond to the expectations and perspective the expert [an agent] brings to the situation, in terms of which the situation solicits a specific response. (Dreyfus, 1996)

Dreyfus (1996) did, however, also point out that "there are many important ways in which neural nets differ from embodied brains". The main difference that he points out is, as in his 1979 AI criticism, the lack of an embedding in a body and an environment. According to Dreyfus (referring to ANN models of human cognition), "this puts disembodied neural-networks at a serious disadvantage when it comes to learning to cope in the human world, Nothing is more alien to our life-form than a network with no up/down, front/back orientation, no interior/exterior distinction, … The odds against such a net being able to generalize as we do, … are overwhelming". Hence, his conclusion is that ANNs would have to be "put into robots" which would allow them to construct their own view of the world. As discussed in Section 4, roughly speaking, this is what much of modern AI and adaptive robotics does.

3.3.4 Radical Connectionism

As an alternative to non-situated connectionist models, Dorffner (1997) suggested a neural bottom-up approach to the study of AI and CS which he termed *radical connectionism*. The key aspects of Dorffner's formulation of this approach, which he referred to as "one possible implementation" of constructivism, are as follows:

- Self-organisation, i.e. automatic adaptation in interaction with the environment, rather than explicit design, should be the major method of developing internal representations and behavioral structures.
- Systems should interact with their environment via sensorimotor interfaces. That means inputs and outputs should not be "pre-digested representations", but the former should come from sensors and the latter should control motors.

- Systems should exploit rich connectionist state spaces, rather than discrete and arbitrary tokens typical for symbolic representations.
- Any high-level aspects of cognition, such as the formation or use of concepts, should be "embedded and grounded in sensorimotor loops and experiences".
- RC research should focus on interactive and situated models, i.e. "models (or systems) that do not passively receive and process input but are inextricably embedded in their environment and in a constant sensori-motor loop with it via the system's own actions in the environment" (Dorffner, 1997, p. 97).

Furthermore, Dorffner (1997, p. 98-100) identified three notions of representation in cognitive science and AI.

- Type 1: "an explicit encoding of structures in the outside world", i.e. the notion of representation "that is employed for most traditional AI models, where knowledge is seen as some kind of 'mirror' of the external world (or a subset thereof …)". This is the notion of representation as a mapping/correspondence between agent-internal and agent-external structures (i.e. a *'Darstellung'*), as illustrated above in Figures 2 and 3.
- Type 2: "internal structures on which an agent operates to guide its behaviour". Following Bickhard and Terveen (1995), Dorffner refers to this notion as *interactivist* or *interactive representation*, and points out: "The meaning of this type of representation is defined only with respect to the agent itself, its drives and needs, and its behaviour". Hence, this notion is compatible with the constructivist notion of representation as a constructed, subjective view of the world, i.e. a *'Vorstellung'* in Kant's sense or a 'presentation' in von Glasersfeld's terms (cf. Section 2). Dorffner points out that although "no encoding of anything must be presumed … those representations can have *aboutness*, but only for the agent itself".

- Type 3: the very broad notion of representations as causal relationships or correspondences, such as between, e.g., a light stimulus and the corresponding neural firings in the retina.

We will see in Section 4 that framework of radical connectionism is to a high degree compatible with much of modern robotic AI. In particular adaptive neuro-robotics, i.e. the combination of neural robot control mechanisms and computational learning techniques, can be seen as a form of radical connectionism, as will be discussed in detail in Section 4.4.

4. New AI: Situated and Embodied Autonomous Agents

Having reviewed computationalist AI and its problems in Section 3, this one will take a close look at the ins and outs of the alternative bottom-up approach, programmatically titled 'New AI', which has been developed since the mid-1980s. For this purpose we will first overview some of the key ideas and terminology of the new approach in Subsection 4.1, introducing New AI notions of 'situatedness', 'embodiment', and 'autonomous agent'. Subsection 4.2 then discusses *artificial life* models, focusing on one of the earliest examples of a situated artificial autonomous agent, Wilson's (1985) Animat, in order to illustrate some of the key issues in New AI in some more detail. Subsection 4.3 discusses Brooks' *behavior-based robotics* approach and his subsumption architecture. Subsection 4.4, finally, examines in detail *adaptive neuro-robotics*, i.e. the use of artificial neural systems and adaptive techniques for the control of autonomous agents, and illustrates the discussion with examples of experimental work on the construction of interactive representations in robot-environment interaction.

4.1 Key Ideas

Around the mid-1980s a number of researchers began to question not only the techniques used by traditional AI, but also its top-down approach and focus on agent-internal reasoning in general. They suggested a bottom-up approach, often referred to *New AI* or *Nouvelle AI*, as an alternative to the

framework of cognitivism and traditional AI (e.g. Wilson, 1985, 1991; Brooks, 1986a, 1990). In particular, it was agreed that AI, instead of focusing on isolated 'high-level' cognitive capacities ('micro-worlds', in Dreyfus' terms), should be approached first and foremost in a bottom-up fashion, i.e. through the study of the interaction between simple, but 'complete' *autonomous agents* and their environments by means of perception and action (e.g., Brooks, 1991a). Beer (1995) characterises the term 'autonomous agent' as follows:

> By *autonomous agent*, I mean any embodied system designed to satisfy internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated. The class of autonomous agents is thus a fairly broad one, encompassing at the very least all animals and autonomous robots. (Beer, 1995)

This broad notion can be considered a good first approximation which, probably, the majority of researchers in New AI would agree to. However, we will see in the following sections that there is some disagreement as to what exactly is meant by "embodied", "situated" or "its *own* actions" in the above definition. Brooks (1991b), the most influential proponent of the new approach, referred to situatedness and embodiment as "the two cornerstones of the new approach to Artificial Intelligence" and characterized them as follows:

> **[Situatedness]** The robots are situated in the world - they do not deal with abstract descriptions, but with the here and now of the world directly influencing the behavior of the system.
>
> **[Embodiment]** The robots have bodies and experience the world directly - their actions are part of a dynamic with the world and have immediate feedback on their own sensations. (Brooks, 1991b, p. 571, original emphases).

A *word of warning*: It may seem that much of the above and the following discussion presupposes that artificial autonomous agents can have first hand semantics and experience or that they have genuine autonomy, subjectivity, qualia, experience and perception, or that the type of learning and evolution we discuss is the same as in living organisms. That is an incorrect impression, as will be discussed in further detail in Section 5.4 (cf. also Sharkey and Ziemke, 1998; Ziemke and Sharkey, in press). However, instead of marking each term with quotes or qualifications such as "it has been

argued that", we have put in this disclaimer so that we can simplify and improve the flow of the discussion.

4.2 Artificial Life Models

One of the earliest autonomous agents to see the light of the (simulated) day was Wilson's (1985) *Animat*, which also came to coin the term *animat approach* for research on this type of agent. The Animat was a simple 'artificial animal', situated in a simulated grid-world environment which also contained trees and food items, as illustrated in Figure 5. The agent's only task was to find (and thereby automatically consume) food items.

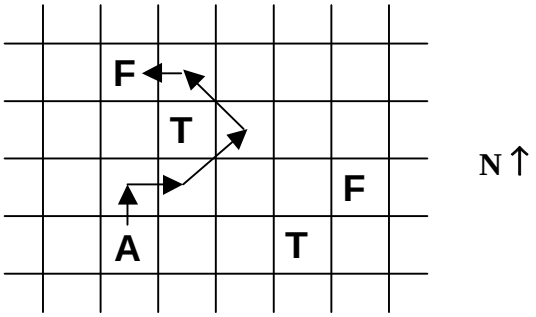

Figure 5: Wilson's (1985) Animat (A), an artificial life agent in part of its simulated grid-world environment. Food (F), in this particular world, is always placed close to trees (T) – a regularity that the Animat learns to exploit. The arrows indicate a sequence of five movements during which the Animat steps around a tree and finally finds a piece of food.

The Animat, whose orientation is constant, is equipped with two sensors each for each of the eight neighboring squares, i.e. it cannot see beyond the cells surrounding it. Of these two sensors one detects edible objects whereas the other detects opaque objects, such that for each of the eight cells close to it the Animat can distinguish between food (input 11 for 'edible' and 'opaque'), tree (01) and an empty square (00). The agent is controlled by a number of classifiers encoding 'situation-action rules'. It adapts these rules through learning in interaction with the environment using genetic

algorithms. Most of the details of the control and learning mechanisms can be ignored here. It is, however, worth noting that in each time step the Animat uses one of its rules to determine its current action. Each rule has a precondition that has to match the current sensor input (possibly using wildcards) and is associated with a certain action, i.e. one of eight possible directions (neighboring cells) that the Animat can move to. For example, one possible rule could be (in abstract from): 'If there is food to the north of me, then I move north'. Thus the Animat's action at time step t is determined based solely on the sensory input at time step t. That means, the agent is *purely reactive*; apart from its 'situation-action rules', it has no means of remembering its past or planning its future.

Nevertheless, Wilson's experiments showed that the Animat learns to carry out what appear to be planned sequences of coordinated actions. For example, as long as it senses neither food nor trees it keeps moving in one direction – an effective 'search strategy' (its environment is shaped like the surface of a doughnut, i.e. there are no walls). When detecting a tree, as illustrated in Figure 5, it 'steps around the tree' until it senses a food item whereupon it moves towards it immediately. Thus, although the Animat has no other memory or 'representation' than its sensor-action rules, it appears to 'know' that food can be found close to trees, despite the fact that it cannot possibly represent this explicitly. Furthermore, as mentioned above, it exhibits coherent sequences of actions or 'strategies', one for 'searching' and one for 'stepping-around-a-tree-in-search-of-food'. The former is of little interest, since it can be explained by a single rule: "if you see neither food nor trees then move to direction X". The latter strategy, however, is more interesting since it requires the agent to deal with a sequence of different situations. In the example illustrated in Figure 5, before the Animat can sense the food, the tree first appears to its north-east, then to its north, etc. Hence, the agent has to 'do the right thing' in each and every time step to carry out a successful sequence, without actually 'knowing' that it is carrying out such a sequence. The reason this works, is of course the tree, which acts as a *scaffold* that 'guides' the agent around itself, and thus towards the food, in a series of purely reactive moves. Similar mechanisms have been shown to be at work in, for example, the mother duck's parenting behavior. Lorenz (1937**)** reported that he first assumed that the mother duck's behavior required some explicit internal representation of a certain duckling being her offspring or

herself being a parent. Later, however, he concluded that each of the different parenting activities was in fact triggered by a different sign stimulus, and that the source of all these stimuli was the duckling 'out there' (cf. also Hendriks-Jansen, 1996). Hence, similar to the Animat's case and in von Uexküll's example of the tick (cf. Section 2), it is not some internal mirror of the external world, but the world 'itself' (as perceived by the agent) that provides continuity and coherence to the agent's behavior. This is exactly what Brooks (1991b) in his above definition of situatedness referred to as the "here and now of the world directly influencing the behavior of the systems". In the New AI this is commonly summarized in slogans such as "The world is its own best model." (Brooks, 1991b).

Although simple simulated AL creatures such as Wilson's Animat certainly have their limitations, it should be pointed out that their value as an approach to the study of intelligent behavior lies in the fact that, unlike traditional AI systems, they interact with their environment directly, i.e. 'on their own'. Franklin (1995) further explains this as follows:

> Symbolic AI systems have been criticized for having their input preprocessed by humans and their output interpreted by humans. Critics maintain that this simplification sloughs off the hardest problems, those of perception and motion. Connectionist systems are often subject to the same criticism. Artificial life creatures, though designed and implemented by humans, sense their environments and act on them directly. Their actions affect the environment and, hence, the creature's future perceptions. All this without further human intervention. The human is taken out of the loop. Thus the semantics of the systems are well grounded and "results are generated by observation rather than by interpretation …. the fruits are 'hard' objective measurements rather than 'soft' subjective ones" (Cliff, 1991, p. 29). (Franklin, 1995, p. 187)

Although, when it comes to situatedness, this type of agent is clearly a step forward from traditional AI systems, it can of course be questioned to what degree its semantics really are "well grounded" and independent of "human intervention" or interpretation. For example, is it really the Animat interpreting the 'F' or '11' as 'food', and even if so, what exactly could that mean to a simulated creature without body and metabolism? Riegler (1997) refers to this as the *PacMan syndrome* in many artificial life models, which do not at all take into account the question of how an agent itself could develop categorical perception or something like a 'food sensor' in actual sensorimotor

interaction with their environment. Riegler (1994, 1997) himself developed an artificial life model directly based on radically constructivist ideas. In this model the environment is not 'pre-digested' and conveniently labelled as in the Animat's case, but built from a number of quasi-physical/-chemical basic elements, thus limiting human intervention at the cognitive/conceptual level to a minimum. Hence, unlike Wilson's Animat, Riegler's agents are largely 'on their own' when it comes to the construction of meaning. Like the Animat, however, they consist of software only, i.e. they are not embodied in the physical sense. From the constructivist point of view, the limitation to computational mechanisms can be argued (as done by Riegler, 1997) to be a benefit, since it avoids the commitment to any assumptions about the existence of a (particular) physical reality. Most AI/robotics researchers, however, do not at all doubt the existence of some physical reality. Hence, work with physical robots rather than simulated agents is commonly argued to be the only way to build and validate models/theories of real-world intelligence (e.g., Brooks, 1991a).

4.3 Brooks' Behavior-Based Robotics and Subsumption Architecture

Rodney Brooks, a roboticist, in the mid-1980s began to argue that the methods of traditional AI, and in particular explicit world models, were simply not suited for use in robots. Typical for traditional AI approaches to robotics is a functional decomposition of the control task following what Brooks (1991b) calls the *sense-model-plan-act (SMPA) framework*. Following the strict perception-cognition-action distinction typical for the cognitivist paradigm, here control is broken down into a series of rather isolated functional units (cf. Figure 6). Input systems handle perception of sensory input, and deliver their results to a modelling module, which integrates the new information into a central world model of the type discussed in Section 3.1. A planner, based on this internal representation of the world alone, then decides on which actions to take. These actions, finally, are executed by the appropriate modules handling, for example, motor control.

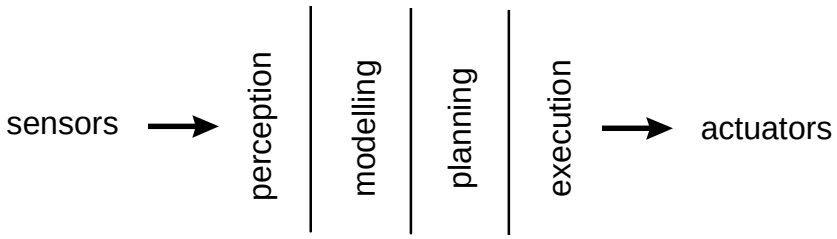

Figure 6: Traditional decomposition of robot control. Adapted from Brooks (1986).

Brooks (1986b, 1991a, 1991b) pointed out a number of flaws in the SMPA framework and instead suggested a decomposition of the control task into *behavior-producing modules*, examples of which might be ‘wandering’, ‘obstacle avoidance’ and ‘light seeking’ (cf. Figure 7). Thus his approach to the study of AI was through the construction of physical robots, which were embedded in and interacting with their environment by means of a number of behavioral modules working in parallel, each of which resembles an Uexküllian functional circle (cf. Section 2). Each of these behavioral modules is connected to certain receptors from which it receives sensory input and, after some internal processing, controls some of the robot’s effectors. Typically all behavioral modules work in parallel, but they are hierarchically organized in a *subsumption architecture* using priorities and subsumption relations for the communication between modules, which allows some of them to override the output of others. Hence the overall behavior of the controlled agent emerges from the interaction of the individual behavioral modules with the environment and among each other. For example, a simple robot with the task to approach light sources while avoiding obstacles, could be controlled by three behavioral modules; one that makes it wander (move forward), a second one that can subsume forward motion and make the robot turn when detecting an obstacle with some kind of distance sensors, and a third one that can subsume the second and make the robot turn towards the light when detecting a light source using some kind of light sensor.

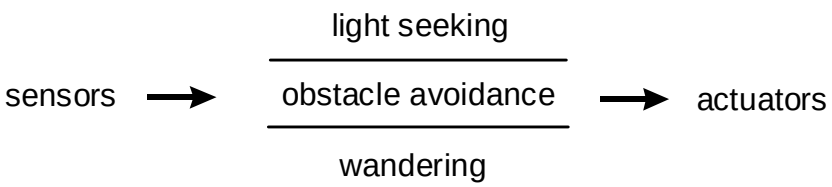

Figure 7: Behavior-based decomposition of robot control. Adapted from Brooks (1986).

In the behavior-based approach robotic agents equipped with sensors and motors are typically considered *physically grounded* as Brooks explains:

> Nouvelle AI is based on the physical grounding hypothesis. This hypothesis states that to build a system that is intelligent it is necessary to have its representations grounded in the physical world. ... To build a system based on the physical grounding hypothesis it is necessary to connect it to the world via a set of sensors and actuators. (Brooks, 1990)

Thus AI has come (or returned) to an Uexküllian view of meaning, in which signs/representations are viewed not as referring to specific external objects, but as embedded in functional circles along which the interaction of agent and environment is organized/structured. Naturally, in the New AI this led to a de-emphasis of representation in the sense of Dorffner's type 1, i.e. an explicit internal world model mirroring external reality (cf. in particular Brooks, 1991a). Brooks (1986a, 1991a) was also, to our knowledge, the first AI researcher to take inspiration directly from von Uexküll's work, and in particular the concept of *Merkwelt* or perceptual world. He pointed out that the internal representations in traditional AI programs really were designer-dependent abstractions. As such, they were based on human introspection, whereas "as von Uexküll and others have pointed out, each animal species, and clearly each robot species with its own distinctly nonhuman sensor suites, will have its own different Merkwelt" (Brooks 1991a). Like Dreyfus (cf. Section 3.2.1), Brooks pointed out that a traditional AI internal representation describing chairs as something one could sit or stand on might be an appropriate representation for a human, but it would probably be entirely meaningless to a computer or a wheeled robot which could not possibly sit down or climb on top of a chair anyway.

A common criticism of Brooks' original architecture is that it does not allow for learning, and thus simply lacks the capacity for autonomous construction of (interactive) representations. Subsumption architectures are typically designed and implemented incrementally, with each step consisting of the implementation and careful testing of one module 'on top' of already tested lower levels. Hence, this type of robot, although operationally autonomous at run-time, remains heteronomous in the sense that the largest parts of its functional circles, namely the processing between receptors and effectors, and thereby the way it interacts with the environment, is still pre-determined by the designer.

4.4 Adaptive Neuro-Robotics

Much research effort during the 1990s has been invested into making robots 'more autonomous' by providing them with the capacity for adaptation and self-organization. Typically these approaches are based on the use of computational learning techniques to allow agents to adapt the internal parameters of their control mechanisms, and thus the functional circles by which they interact with their environment. The robots used in this type of research are often mobile robots (see Figure 8 for a typical example), typically receiving sensory input from, e.g., infrared proximity or simple cameras, and controlling the motion of their wheels by motor outputs. Very often the control mechanism is some form of ANN used as an *'artificial nervous system'* connecting the robot's receptors and effectors. The frequent use of ANNs is mostly due to two reasons. Firstly, there are the advantages of the (radically) connectionist approach from an AI/CS perspective (cf. Section 3.3). Secondly, ANNs have a number of additional advantages from an engineering/robotics point of view, such as their flexibility and robustness to noise. Thus, the use of neurally-controlled robots using learning and/or evolutionary adaptation techniques, hereafter referred to as *adaptive neuro-robotics*, has become a standard methodology in bottom-up AI research.

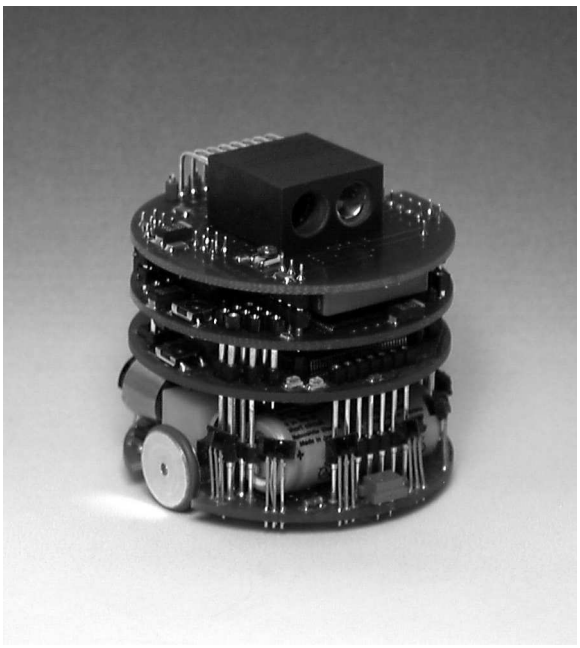

Figure 8: The Khepera, a wheeled miniature mobile robot commonly used in adaptive robotics research (manufactured by K-Team SA; for details see Mondada *et al.*, 1993). The model shown here is equipped with infrared sensors and a simple camera.

The rest of this section is structured as follows: Subsection 4.4.1 discusses the role of adaptive neuro-robotics as a form of radical connectionism from an AI and cognitive science perspective. Different adaptation techniques and their combination with ANNs are then discussed in Subsection 4.4.2, and it is briefly exemplified how such techniques can allow autonomous agents to adapt their interactive representations in order to self-organize their sensorimotor interaction with the environment. Subsection 4.4.3 then discusses in detail an example of experimental work on the construction of interactive representations in adaptive robots.

4.4.1 Adaptive Neuro-Robotics as Radical Connectionism

As Dorffner (1997) himself emphasized, the RC approach has much in common with Brooks' formulation of a Nouvelle AI. Unlike subsumption architectures, however, connectionist networks are typically adaptive and they offer richer representational possibilities. Hence, they allow an artificial agent to construct its *own* view of the world. Thus, "constructivism … finds one possible

implementation in radical connectionism" (Dorffner, 1997) and adaptive neuro-robotics becomes interesting as an approach to the study of interactive cognition. This approach makes a significant difference from an AI/CS perspective. When used as an 'artificial nervous system', the connectionist network can actually, by means of the robot body (sensors and effectors), interact with the physical objects in its environment, independent of an observer's interpretation or mediation. Thus the network becomes an integral part of a robotic agent that is situated and embodied in Brooks' sense. Moreover, its internal representations, now formed in physical interaction with the world they 'represent' or reflect, can be considered physically grounded in Brooks' sense and furthermore constructed through adaptive processes. The robot controller network is in this case part of a complete functional circle (or several circles, as will be discussed below), and can to some degree construct its own *Umwelt*. As an example of this view, imagine a wheeled robot moving about in a room with boxes lying on the floor and pictures hanging on the wall. The robot might be equipped with infrared sensors as receptors sensitive to the perceptual cues of, for example, the reflectance patterns of solid objects in its environment. Thus, the walls and the boxes on the floor would be part of the robot's own perceptual world (*Merkwelt*), cf. Section 4.3. Their 'meaning' to the robot would be that of an 'obstacle', since they limit the robot's motion, assuming the robot has the goal to keep moving while avoiding collisions. The pictures on the wall, on the other hand, would remain 'invisible' to the robot; they are not part of its perceptual world, and they carry no meaning for it. Thus the robot may be considered to be embedded in its own *Umwelt*, consisting of its perceptual world (*Merkwelt*), containing solid objects (or their absence), carrying the meanings 'obstacle' and 'free space' respectively, and its operational world (*Wirkwelt*) of motor-controlled wheeled motion. The "inner world" of the robot would be the ANN's internal sign flow and interactive representations. Hence, unlike in the cases of 'pre-digested' traditional AI programs or Brooks' subsumption architecture, the inner world would here be a self-organized flow of private signs embedded in agent-environment interaction. Thus, adaptation in ANN robot controllers can be viewed as construction of interactive representations through the creation, adaptation and/or optimization of functional circles in interaction with the environment. Although the above example

illustrated only one such circle, we can of course easily imagine several functional circles combined/implemented in a single ANN. For example, if we additionally equipped the robot with a light sensor and added light sources to the environment, we might have three functional circles; one that makes the robot move forward when encountering 'free space', one that makes it turn/avoid when encountering 'obstacles', and one that makes it approach when detecting the light. We will see a number of more concrete examples of the construction of interactive representations in ANN-controlled robots in the following subsections.

4.4.2 Robot Adaptation

In the earlier discussion of connectionist networks we briefly discussed supervised learning. Typically, however, supervised techniques are not used to train robots on complex tasks. This has two reasons: Firstly, in order to allow for a maximum of robot autonomy, it is often desirable to reduce designer intervention to a minimum of feedback/instruction (cf., e.g., Nolfi, 1998). One reason for this is that the designer is likely to view the control task from his/her own distal perspective, which is not necessarily guaranteed to fit the robot's proximal perspective (cf. Nolfi, 1998). Secondly, as Meeden (1996) has pointed out, robot/agent problems are often defined in terms of abstract goals rather than specific input-output pairs. Thus, moment-to-moment guidance is typically not available for a learning robot since for a given situation there is not necessarily just one right or wrong action, and even if there was, it would typically not be known *a priori* (Meeden, 1996). Roughly, this problem can be likened to that of telling a child learning to ride a bike how exactly to move its legs, arms and body at every point in time. For such tasks, the robot, much like the child, simply has to construct itself a viable 'solution', i.e. a way to organize and adapt its sensorimotor transformation knowledge in interaction with the environment. Hence, robots are often trained using *reinforcement learning* or *evolutionary adaptation* techniques. The latter can be considered a special case of the former (e.g., Meeden, 1996). In both these cases the trained robots have to self-organize their internal structures and parameters to fit certain environmental constraints. Roughly this can be likened to the constructivist view of organisms striving towards a 'conceptual

equilibration' through adaptation of their internal/conceptual structures to fit their experience (cf. Section 2).

During conventional reinforcement learning, an agent is provided only with occasional feedback, typically in terms or positive and negative reinforcement ('good' or 'bad'). From this feedback the agent can adapt its behavior to the environment in such a way as to maximize its positive reinforcement ('reward') and minimize its negative reinforcement ('punishment'). The use of evolutionary techniques is an approach to 'push' the designer even further 'out of the learning loop' and aims to let robots learn from the interaction with their environment with a minimum of human intervention (cf. Nolfi, 1998; Nolfi and Floreano, 2000). Evolutionary methods are abstractly based on the Darwinian theory of natural selection. Thus feedback is not instructive as in supervised learning, but only evaluative. Typically, a population of individuals (i.e. 'artificial genotypes', encoding, e.g., robot control networks as strings of bits or numbers) is evolved over a large number of generations, in each of which certain individuals are selected according to some *fitness function*, and 'reproduced' into the next generation, using recombinations and slight mutations mimicking natural reproduction. Due to the selective pressure the average fitness in the population is likely to increase over generations, although the individuals typically do not learn during their 'lifetime' (for a discussion of the combination of evolutionary and learning techniques see Nolfi and Floreano, 1999, 2000). The very idea of evolving robots was well illustrated by Braitenberg (1984) who likened evolution to the following scenario: There are a number of robots driving about on a tabletop. At approximately the same rate that robots fall off the table, others are picked up randomly from the table, one at a time, and copied. Due to errors in the copying process, the original and the copy might differ slightly. Both are put back onto the table. Since the fittest robots, those which stay on the table longest, are most likely to be selected for 'reproduction' the overall fitness (i.e. the time robots remain on the table) of the robot population is likely to increase in the course of the 'evolutionary' process.

A concrete example of *evolutionary robotics* research is the work of Husbands *et al.* (1998) who evolved recurrent ANN robot controllers for a target discrimination task, which required a mobile robot, equipped with a camera, to approach a white paper triangle mounted on the wall, but to avoid rectangles. In these experiments both the network topology and the visual morphology (or receptive field), i.e. which parts/pixels of the camera image the controller network would use as inputs, were subject to the evolutionary process. The analysis of the experimental runs showed that structurally simple control networks with complex internal feedback dynamics evolved which made use of low bandwidth sensing (often only two pixels of visual input were used) to distinguish between the relevant environmental stimuli. Thus in these experiments both the internal flow of signals and use of feedback, as well as the 'external' sign use, i.e. which environmental stimuli to interpret as signs of what, are constructed in an artificial evolutionary process. The evolved sign processes are often difficult to analyze and understand in detail, due to the fact that they are *private* to the robot and in many cases radically differ from the solutions the human experimenters would have designed based on their own distal perspective. Husbands *et al.* point out that this "is a reminder of the fact that evolutionary processes often find ways of satisfying the fitness criteria that go against our intuitions as to how the problem should be 'solved'" (Husbands *et al.* 1998, p. 206).

The influence of the human designer can be reduced even further using *co-evolutionary methods*. Nolfi and Floreano (1998), for example, co-evolved two robots controlled by recurrent connectionist networks to exhibit predator- and prey-behavior. The 'predator', a Khepera robot equipped with a simple camera (cf. Figure 8), which allowed it to observe the prey from a distance, had to catch (make physical contact with) the 'prey'. The latter is another Khepera robot, equipped only with short-range infrared sensors but also with the potential to move faster than the predator. By simply evolving the two 'species' with time-to-contact as an *implicit* fitness and selection criterion, quite elaborate pursuit- and escape-strategies evolved in the respective robots. The predator species, for example, in some cases developed a dynamics that allowed it to observe and interpret the prey's current behavior as a symptom of its current behavioral disposition, and thus of its behavior in the

immediate future. Hence, it would only ‘strike’ when it had a realistic chance to catch the prey ‘off guard’.

These examples illustrate that the adaptive neuro-robotics approach of letting robots construct their own sensorimotor mechanisms and (interactive) representations through self-organization in interaction with their environment is largely compatible with (radically) constructivist ideas and an interactive view of cognition (cf. Section 2). The robots used in this type of work are, of course, still the product of much human design when it comes to physical construction, experimental setups and other pre-determined aspects. However, typically no explicit knowledge is built into these systems (and no categorical perception, as in the Animat’s case, either), but the construction of sensorimotor transformation knowledge is to a high degree left to the robot itself. Hence, this approach is largely compatible with the constructivist view of knowledge as actively built up by cognizing subjects and serving their organization of their sensorimotor interaction with the environment, rather than the discovery of some objective reality. Furthermore, the view of adaptation as tending towards fit or viability is, at least at a first glance, compatible with the use of evaluative feedback/selection in reinforcement and evolutionary robot learning. A more detailed and critical discussion of the differences between adaptive robots and living organisms, as well as the mechanisms of their respective situatedness will be presented in Section 5.

4.4.3 A Detailed Example of Interactive Representations in an Adaptive Robot

Wilson’s Animat is an example of a purely reactive system which despite its limitations can exhibit non-trivial behavior due to the fact that it uses trees as scaffolds to guide its behavior. While this might be comparable to the behavior of lower animals such as snakes (cf. Sjölander, 1999), for most animals the environment is not that benevolent in the sense that sufficiently reliable scaffolds are already built into it. Similarly, robots often have to solve context-dependent or temporally extended tasks like homing or finding a goal location, but are faced with the problem of *perceptual aliasing*. That means, many locations in the environment look the same from the robot’s current point of view, such that they cannot be distinguished without knowledge/memory of where the robot came from.

Hence, in many other cases robots have to exhibit adaptive behavior to deal with requirements changing over time, and they have to do so without reliable external scaffolds.

Meeden, for example, discussed the case of Carbot (cf. Figure 9), a toy-car-like robotic vehicle of about 23 cm length and 15 cm width placed in a rectangular environment of approximately 120 cm by 60 cm (Meeden *et al.*, 1993; Meeden, 1996). Apart from the 'low-level' goal of avoiding bumping into the walls surrounding the environment, Carbot's 'high-level' goal periodically changed between having to approach a light source placed in one corner of the environment, and having to avoid it. This means that it should, depending on the current goal, maximize or minimize the readings of two light sensors directed towards the front of the vehicle (cf. Figure 9). Apart from that, Carbot was equipped only with digital touch sensors at the front and the back of the vehicle which detected collisions when they occurred, but did not give any advance warning. The task was further complicated by the fact that the smallest dimension of the environment was smaller than Carbot's turning radius. The vehicle could therefore only execute a 180-degree turn in a series of backward and forward movements. In all experiments documented by Meeden *et al.* (1993) and Meeden (1996) neural robot controllers were trained in a tailor-made simulation of Carbot and its environment, and trained networks were transferred to and evaluated on the real robot.

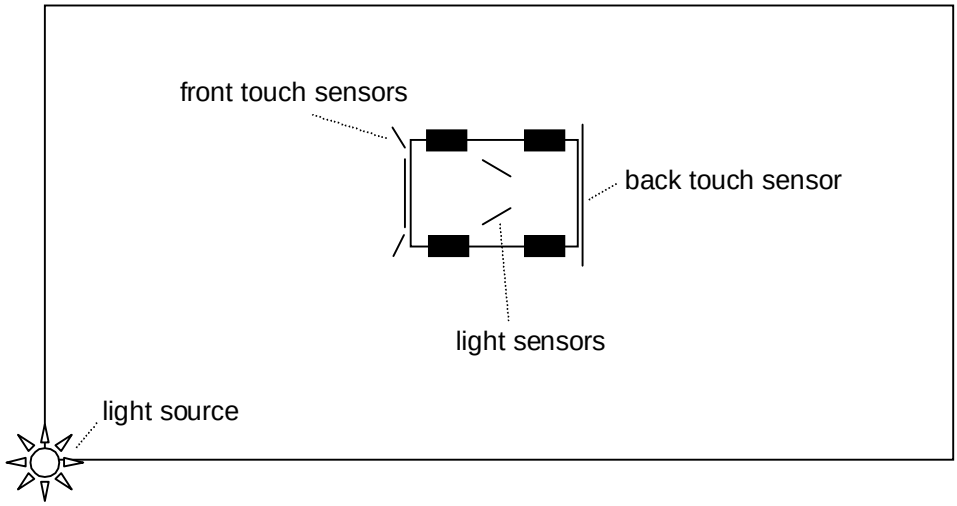

Figure 9: Schematic illustration of Meeden's Carbot and its environment

Meeden *et al.* (1993) carried out extensive experimental comparisons of different feed-forward and recurrent control architectures for varying Carbot tasks. The basic recurrent control architecture (cf. Figure 10) was similar to Elman's (1990) Simple Recurrent Network (cf. Section 3.2). The network's inputs came from light and touch sensors (plus in some experiments an extra input explicitly indicating the current goal), its outputs controlled the motor settings, and the hidden unit activation values were copied back and used as extra inputs in the next time step. Not surprisingly, the experimental results of Meeden *et al.* (1993) showed that recurrent networks consistently outperformed feed-forward networks.

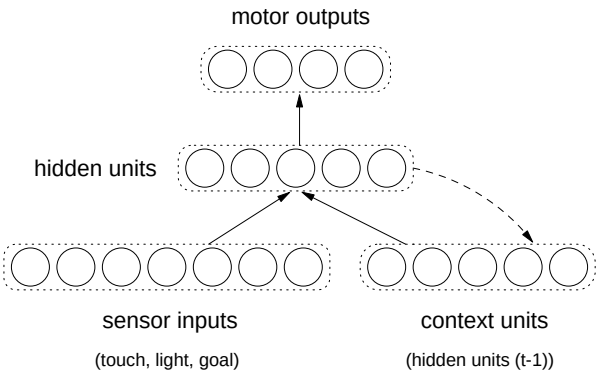

Figure 10: Meeden's recurrent robot control architecture. Solid arrows represent fully connected layer of weights between two layers of units (indicated by surrounding dotted lines). Hidden unit values are fed back via a weightless 1:1 copy connection (dashed arrow) and used as extra inputs in the next time step. Adapted from (Meeden, 1996).

Meeden *et al.* (1993) analysed the recurrent control networks and showed that they utilized their internal state (i.e. the hidden unit activation values which were fed back) to carry out coherent behavioral sequences corresponding to one-, two- and multi-turn strategies instead of merely reacting to the current input. Let us have a closer look at a particular strategy, one that uses multiple turns. In this case, as illustrated in Figure 11, to avoid the light the robot (starting from a position facing the light) first moved backwards towards the centre of the environment during time steps 1-3. It then carried out a series of alternating forward-right and backward-left movements (time steps 3-8) until it faced away from the light in time step 8 and thus satisfied its current high-level goal. This multi-turn strategy effectively overcame the problem that the environment's smallest dimension is smaller than its own turning radius. The robot executed multiple turns in a position where it did not risk colliding with walls. Having fulfilled the light-avoidance goal, Carbot switched itself back into seek mode, and a similar strategy was executed during time steps 9-16 to approach the light source again (cf. Figure 11). This time the robot oriented itself towards the light through a series of backward-right and forward-left movements during time steps 9-14, and then moved forward towards the light during time steps 14-16.

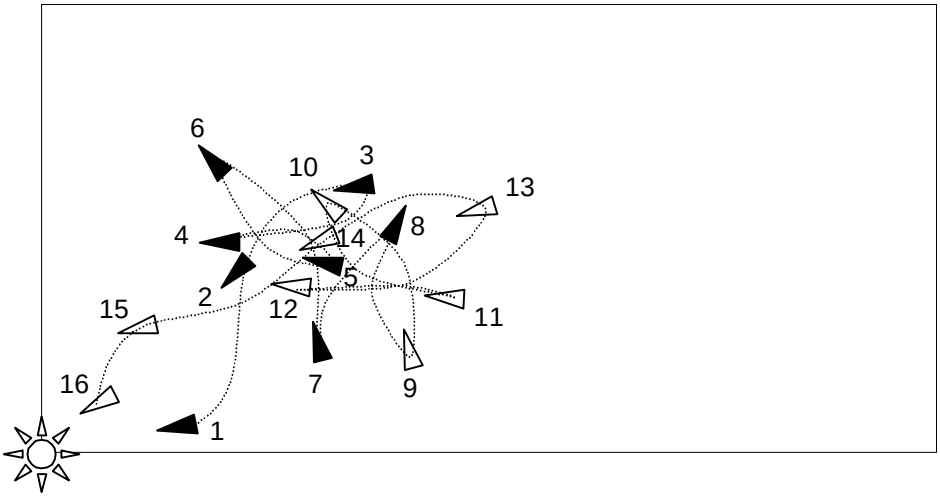

Figure 11: Schematic illustration of a (simulated) multi-turn strategy (adapted from Meeden *et al.*, 1993). The triangles indicate Carbot's position and orientation in every time step. Steps 1-8 occurred during avoid mode (black triangles; goal achieved at step 8), 9-16 during seek mode (white triangles; goal achieved at step 16). See text for a detailed explanation.

Figure 12 illustrates the trajectory in internal state space which corresponds to Carbot's sequence of actions illustrated in Figure 11. During avoid mode backward motion away from the light was achieved through three times steps (1-3) in/near cluster D, and the multi-turn orientation away from the light was implemented through the network's alternation between areas C and D. During seek mode then the multi-turn orientation towards the light was achieved through alternation between areas A and B during time steps 10-14, triggering forward- and backward-movements respectively, followed by approach of the light source in time steps 14-16. Thus Carbot's behavior was guided by an internal dynamic that allowed it not merely to react to its environment. That means, it did not simply follow the light gradient in every time step, but instead executed a coherent series of actions/movements that led to achievement of the goal eventually.

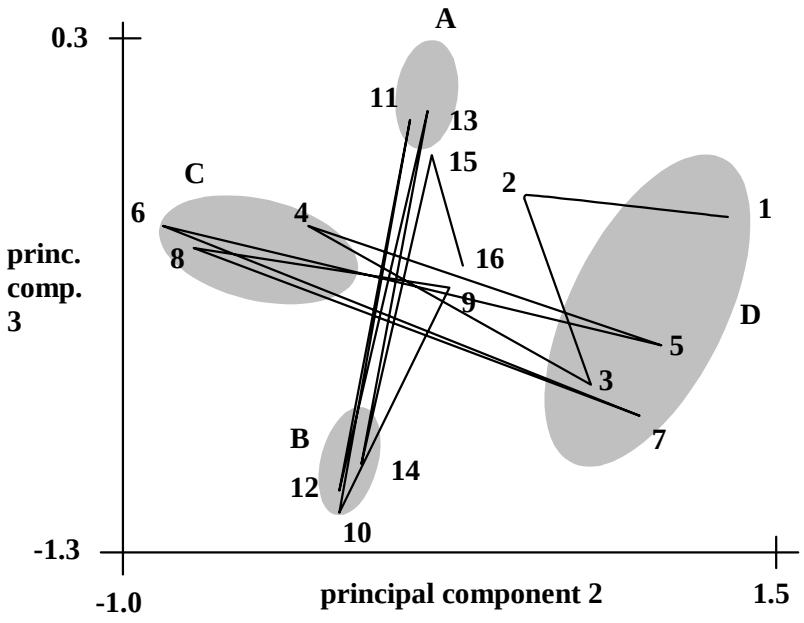

Figure 12: Internal state trajectory corresponding to the action sequence illustrated in Figure 11 (redrawn from Meeden et al., 1993). The 5-dimensional hidden unit space is here reduced to two principal components. Shaded regions indicate the most visited areas. Carbot alternates between A and B while orienting towards the light in seek mode, between C and D while orienting away from light in avoid mode.

Meeden *et al.* (1993) argued that Carbot's behavior was *plan-like* in the sense that (a) it associated abstract goals with *sequences* of primitive actions, (b) the behavior could be described in hierarchical terms (cf. Figure 13), and (c) the robot maintained its overall strategy even when flexibly reacting to the environmental conditions. For example, when encountering a wall while carrying out the above light avoidance strategy, it reacted to the wall first and then returned to its high-level strategy. On the other hand, the behavior was not plan-like in the traditional sense that the robot would explicitly anticipate the future, and the number and complexity of Carbot's strategies were admittedly limited.

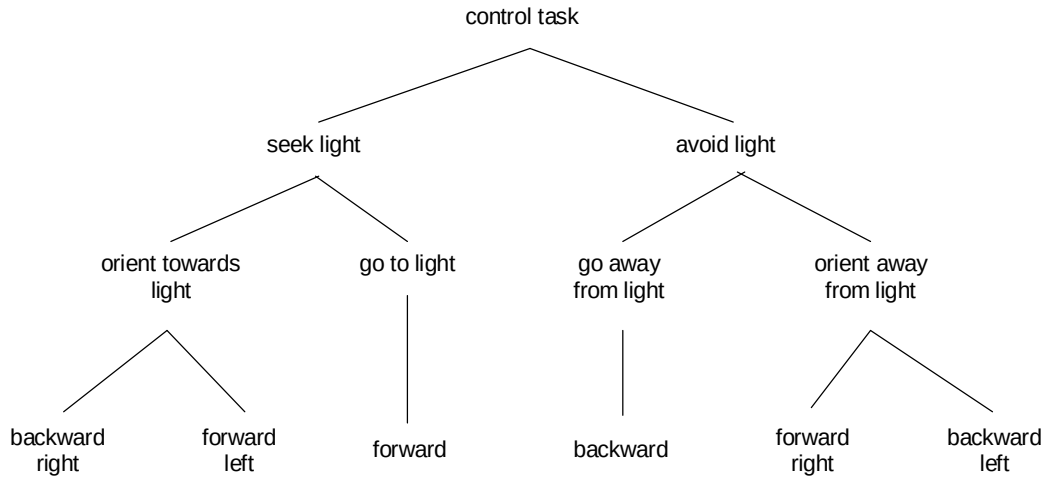

Figure 13: A hierarchical *description* of Carbot's behavior. Adapted from Meeden *et al.* (1993).

It might be worth noting that the internal states and emergent behavioral structures/hierarchies in Carbot's control networks can be considered good examples of interactive (type-2) representations in Dorffner's sense. That means

- they are internal structures on which Carbot operates to guide its behavior;
- they are the result of self-organization in interaction with the environment, rather than explicit design, and thus can be considered grounded representations 'subjective' to Carbot;
- they do not encode anything in the external world in the traditional, referential sense; in fact there are several different strategies and internal representations that solve the same task in the exact same world; i.e. each of these is a viable 'view' of the robot's world in the sense that each of them allows it to solve the task and 'survive'.

5. Discussion and Conclusion

Through Sections 2 to 4 we have incrementally narrowed the discussion down from constructivist theories about the relation between organisms and their environment, over different types of AI, to experiments with recurrent neural robot controllers which realize a form of situatedness in the sense of continual context-dependent self-revision. In this section we intend to incrementally broaden the discussion again and put things in perspective. We start in Subsection 5.1 by summarizing and discussing different types/forms/levels of situatedness from an AI-internal perspective, focusing on the role of constructive processes. Hence, 5.1 is mostly concerned with robot situatedness and the main question is *'What are the differences, with respect to situatedness and constructive processes, between today's situated robots and the traditional AI systems criticized by Dreyfus and Searle?'*. Subsection 5.2 then connects back to von Uexküll's discussion of the differences between organisms and the (man-made) mechanisms of his time (cf. Section 2), in order to see to what extent modern AI's biological inspiration might have contributed to closing that gap. Thus, the main question discussed here is *'What are the differences between situated robots and conventional mechanisms?'*. Subsection 5.3 then takes a closer look at the situatedness and autonomy of living organisms. Here we compare von Uexküll's theory to its modern counterpart, the work of Maturana and Varela on autopoiesis, embodied cognition, and their biological basis. Hence, the guiding question in 5.3 is *'What exactly are the mechanisms of organismic situatedness?'*. Subsection 5.4 then puts the discussions of robotic and organismic situatedness together, and asks *'What are the (remaining) differences, with respect to situatedness and constructive processes, between situated robots and organisms?'*. Subsection 5.5, finally, presents a brief summary and conclusion.

5.1 Robotic Situatedness

5.1.1 Summary: Forms and Mechanisms of Robotic Situatedness

We have seen in this paper that the New AI directly grew out of a number of criticisms accusing its traditional counterpart for its lack of situatedness. The various attacks of Dreyfus, Searle, and

Brooks, came from different perspectives, but they all agreed in the judgement that the traditional approach was 'wrong', and that situatedness was one of the key components missing. Thus, it has become one of the "cornerstones of the new approach to Artificial Intelligence", as Brooks (1991b) formulated it. There is, however, more agreement about the failure of traditional AI than about how to proceed with the new approach. While probably everybody in the New AI agrees to Brooks' formulation of situatedness as the property of dealing "with the here and now of the world directly influencing the behavior of the system", there are very different interpretations of this slogan. In particular there was disagreement about the role and relevance of *embodiment*, the other 'cornerstone'. Some researchers, including Brooks (e.g., 1991a) himself, are strongly dedicated to *physical* embodiment, and thus physical grounding/situatedness in the 'real world', whereas others are 'content' with simulated creatures, such Wilson's Animat, situated in simple man-made environments. Franklin (1997), for example, argued: "Software systems with no body in the usual physical sense can be intelligent. But, they must be embodied in the situated sense of being autonomous agents structurally coupled with their environment."

Similarly there is disagreement about what exactly is meant by 'autonomy'. A number of researchers have criticized Brooks' original approach for the lack of adaptivity, something they consider an essential aspect of autonomy (cf. Ziemke, 1998). Sharkey and Heemskerk (1997), for example, used the metaphor of the *environmental puppeteer* guiding the robot by the 'strings' of predetermined reactive mechanisms and they pointed out: "An important goal of modern "scientific" robotics is to *cut the strings* and give the robot its autonomy." We have elsewhere (Ziemke, 1997) discussed in detail the relation of this metaphor to different aspects of autonomy, including those discussed by von Uexküll (cf. Section 2). At the risk of overusing the metaphor, an agent using feedback can be likened to a puppet, which in addition to external strings pulled by the environmental puppeteer pulls some internal strings of its own. Obviously, this is useful where the "here and now" is not as reliable as in the Animat's case. Hence, this agent is not only situated in its "here and now", but it is furthermore situated in its own history of percepts.

An agent capable of self-organization, on the other hand, would correspond to a puppet with the capacity to adapt its own strings in order to maximize its 'fit' with environmental constraints (as, in the robot's case, expressed in reinforcement or fitness functions). As discussed in Section 4 there is extensive research activity in this area. In particular the combination of neural control mechanisms and adaptive algorithms has been argued to be a promising route, due to the fact that ANNs are flexible and open to adaptive mechanisms while requiring a minimum of designer intervention (e.g Meeden, 1996). The adaptive neuro-robotics approach, which can be considered a form of radical connectionism, thus allows agents to construct/form representations in a grounded fashion, i.e. in interaction with the environment. As a result agents are no longer merely situated in the physical sense, but they are also situated in their own representation or conceptualization of the world. Thus, they do not just have their own *Merkwelt* (perceptual world) in the Brooksian sense that their perception depends on their specific sensors, but they have also developed an 'own' interpretation of their sensory input, i.e. 'own' categories or behavioral mechanisms.

An agent that can selectively use different sensorimotor mappings, finally, could be likened to a puppet that has different sets of strings to choose between. One way of achieving this is the use of synchronically structured, modular architectures, such as Brooks' subsumption architecture, that break down a control mechanisms into a number of sensorimotor mappings and some way of integrating or mediating between them. This can be contrasted with diachronically structured control mechanisms realized in recurrent neural robot controllers, in which different sensorimotor mechanisms 'emerge' over time as a result of constant, context-dependent self-adaptation (cf. Figures 11-13). Using experiments by Meeden, we illustrated how a robot can dynamically bias its own behavior in the immediate future depending on its immediate past. Or, in terms of the puppet metaphor, it keeps pulling internal strings in order to complement or override the effect of the external strings pulled by the environment. Other experiments by ourselves (e.g., Ziemke, 1999a), not discussed in detail here, used higher-order networks (cf. Section 3.3.2) in which the connection weights embodying the sensorimotor mapping can be dynamically adapted from moment to moment. This can be likened to a puppet that, depending on its own current 'needs', adapts the strings to be

pulled by the environment. A similar metaphor is also used by Merleau-Ponty (1963) who rejects the notion of an organism as a keyboard played on by environmental stimuli as follows:

> The organism cannot properly be compared to a keyboard on which the external stimuli would play and in which their proper form would be delineated for the simple reason that the organism contributes to the constitution of that form. … [I]t is the organism itself - according the proper nature of its receptors, the thresholds of its nerve centers and the movements of the organs - which chooses the stimuli in the physical world to which it will be sensitive. "The environment (*Umwelt*) emerges from the world through the actualization or the being of the organism – [granted that] an organism can exist only if it succeeds in finding the world an adequate environment." [(Goldstein, 1934)] This would be a keyboard which moves itself in such a way as to offer - and according to variable rhythms - such or such of its keys to the in itself monotonous action of an external hammer. (Merleau-Ponty, 1963, p. 13)

Thus, recurrent neural control mechanisms allows robots to exhibit adaptive, context-dependent responses to otherwise identical stimuli, e.g. light-avoidance and –seeking in identical external conditions in Meeden's experiments. In the sense described by Merleau-Ponty above, this mechanism of (varying) meaning attribution brings forth a phenomenal world (*Umwelt*) in the interaction of robot and environment. In this interaction it is in fact the environment pulling the strings, but it is the agent which, as an active subject, (re-) constructs the strings and thus uses the environment as a resource. Moreover, the robot is "always already in a situation" in the sense that it always encounters the world with some behavioral disposition, i.e. a way of attributing meaning to stimuli, which provides it, to a limited degree, with what Dreyfus referred to as a 'sense' of situatedness. Furthermore, as in Dreyfus account of (human) situatedness, the context in which perceptually-guided action occurs is not some background of explicit representations, but it is a constantly revised situation or *Umwelt* realized through a fluctuating behavioral disposition.

5.1.2 Beyond Software Situatedness?

The reader might have noticed already that, although discussing robots, we have hardly mentioned hardware at all. The discussion of robotic situatedness has so far been limited to situatedness in the

sense of robots having self-organized representations to various degrees. We spoke of '*physical* grounding', but so far the role of the physical has in fact been limited to sensors that provide inputs from the world and motors that transform outputs into action in the world. The mechanisms in between sensors and motors have in all cases been limited to *computational* mechanisms, i.e. software. Accordingly, the forms of situatedness discussed in the previous subsection, could all be characterized as various degrees of *software situatedness*.

Recently, however, researchers have begun to apply evolutionary methods also to the construction of physical structures and robot morphologies (in simulation) (e.g., Funes and Pollack 1997; Lund *et al.* 1997), in some cases in co-evolution with robot controllers (Cliff and Miller 1996, Lund and Miglino 1998. Cliff and Miller (1996), for example, simulated the co-evolution of 'eyes' (number and position of visual sensors) and 'brains' (ANN controllers mapping visual input to motor output) of simple robotic agents which pursued and evaded each other in a two-dimensional plane. Their results show that predator and prey species evolve significantly different morphologies suited to solve their respective tasks. Experiments with evolving body structures actually realized in hardware, constructed from LEGO™ bricks, were carried out by Funes and Pollack (1997) as well as Lund and Miglino (1998). In the first case structures, such as cranes or bridges, were evolved in simulation to fit certain demands (e.g., the weight a crane should be able to carry), and then built and successfully tested in hardware. In the second case LEGO™ robot bodies were co-evolved with artificial nervous systems for the control of these bodies. Again, the evolutionary process was here carried out in simulation, but later evolved robots/controllers were successfully tested in real hardware. The process of artificial 'brain-body co-evolution' has been further automated by Lipson and Pollack (2000) who co-evolved thermoplastic robot bodies and neural controllers. The evolutionary process is carried out in simulation, as in the above examples, but here the evolved body does not have to be assembled by a human according to an evolved body plan, but the body (except for the motors) can be built automatically using 3D solid printing technology.

Hence, in these cases the evolved agents are not just situated in their own 'software' (control mechanism), but they are also situated in the sense that they possess a self-organized body that fits their world and task requirements. This aspect will be discussed in further detail in Section 5.4; for a detailed discussion of the epistemological implications of robotic devices which evolve/construct their own hardware see also Cariani (1992).

5.2 How Situated Robots Differ from Conventional Mechanisms

We have now seen a number of examples of autonomous agents and their self-organization. Together these examples illustrate that situated robots, although certainly mechanisms in the technical sense, in a number of points radically differ from the type of mechanism that von Uexküll discussed, and in fact exhibit some of the constructive processes that he ascribed to organisms alone (cf. Section 2).

Firstly, the use of 'artificial nervous systems' in combination with computational learning techniques allows autonomous agents to adapt to their environment. Furthermore, when using internal feedback the behavioral disposition of autonomous agents can vary over time. Thus, although they do not "grow" in the physical sense (except for the above cases of robot body evolution), they do adapt to their environment, such that they do in fact have what Dreisch and von Uexküll referred to as a "historical basis of reaction" (*Reaktionsbasis*) (cf. Section 2). Self-organized situated robots thus have a 'subjective' quality, in the sense that the way they react is not determined by built-in rules (alone), but is specific to them and their history of 'experience' and self-organization.

Secondly, and closely related to the previous point, artificial autonomous agents are clearly involved in sign processes, and they 'make use' of signs 'themselves', unlike the mechanisms von Uexküll discussed. Furthermore, unlike typical computer programs, the sign processes of situated robots are usually (a) not (fully) determined by their human designers, (b) independent of interpretation through external observers (at least at the operational level), and (c) in many cases not even interpretable to humans at a close look at the internal processes (despite the fact that these are much easier to observe than in the case of a living organism). Much of the sign usage of such systems is therefore, due to

their self-organization, indeed *private* and specific to them. Autonomous agents therefore have been argued to have certain degree of *epistemic autonomy* (Prem, 1997; cf. also Bickhard, 1998), i.e. like living organisms they are "on their own" in their interaction with their environment.

Thirdly, the use of self-organization, especially evolutionary techniques, *does* nowadays (to some degree) allow the construction of robot controllers, and to some degree even robot bodies, following *centrifugal principles*. As illustrated with Meeden's experiments (cf. Section 4.4.3), in recurrent neural robot control architectures the control of a robot is distributed over a number of functional circles in a process of dynamic adaptation and differentiation. In these cases the control mechanism is not constructed along centripetal principles, i.e. not broken down into sub-tasks or -competences by a designer to be integrated later, but instead constructed making use of what might be called *centrifugal task decomposition*. That means, a single control mechanism breaks itself down into a number of sub-mechanisms in a process of adaptation and differentiation. Similar principles have even been applied to the co-evolution of physical structures and robot morphologies with controllers (cf. Section 5.1.2). Here robot body and controller are no longer treated as isolated elements to be constructed separately, but instead they are co-evolved in an integrated fashion.

5.3 Organismic Situatedness

Having illustrated the principles of situated robots and their self-organization and having outlined the differences between such systems and conventional mechanisms in the previous section, we will now turn to the differences between situated robots and living organisms. Subsection 5.3.1 presents a brief comparison between von Uexküll's theory and its modern counterpart, the work of Maturana and Varela on autopoiesis and the biology of cognition. Here we focus on what the mechanisms of organismic autonomy and situatedness are according to these two theoretical frameworks. Subsection 5.3.2 then further relates this to the earlier discussions of the (recurrent neural) mechanisms of robotic situatedness. The implications of the lack of a living body for the possibilities and limitations of robotic situatedness will then be considered in detail in Subsection 5.4.

5.3.1 Von Uexküll vs. Maturana and Varela

We have discussed elsewhere (Ziemke and Sharkey, in press) in detail the similarities between the theories of von Uexküll and their modern counterpart, the work of Maturana and Varela's work on the biology of cognition and autopoiesis (e.g., Maturana and Varela, 1980, 1987). Of particular interest in the discussion of the differences between robotic and organismic situatedness and the underlying constructive processes is Maturana and Varela view of living systems as characterized by their autopoietic organization. The *organization* of a system, similar to von Uexküll's notion of a building-plan (*Bauplan*), denotes "those relations that must exist among the components of a system for it to be a member of a specific class" (Maturana and Varela 1987, p. 47). An *autopoietic* system is a special type of homeostatic machine for which the fundamental variable to be maintained constant is its own organization. This is unlike regular homeostatic machines, which typically maintain single variables, such as temperature or pressure. A system's *structure*, on the other hand, denotes "the components and relations that actually constitute a particular unity, and make its organisation real" (Maturana and Varela 1987, p. 47). Thus the structure of an autopoietic system is the concrete realization of the actual components (all of their properties) and the actual relations between them. Its organization is constituted by the relations between the components that define it as a unity of a particular kind. These relations are a network of processes of production that, through transformation and destruction, produce the components themselves. It is the interactions and transformations of the components that continuously regenerate and realize the network of processes that produced them.

Hence, according to Maturana and Varela (1980), living systems are not at all the same as machines made by humans. The latter, including cars and robots, are *allopoietic*. Unlike an autopoietic machine, the organization of an allopoietic machine is given in terms of a concatenation of processes (note the similarity to von Uexküll's notion of centripetal construction, cf. Section 2). These processes are not the processes of production of the components that specify the machine as a unity. Instead, its components are produced by other processes that are independent of the organization of the machine. Thus the changes that an allopoietic machine goes through without losing its defining

organization are necessarily subordinated to the production of something different from itself. In other words, it is not truly autonomous, but heteronomous. In contrast, a living system is truly autonomous in the sense that it is an autopoietic machine whose function it is to create and maintain the unity that distinguishes it from the medium in which it exists. Hence, despite differences in terminology, Maturana and Varela's distinction between autopoietic and allopoietic machines, is very similar to von Uexküll's (1928) earlier discussed distinction between human-made mechanisms, which are constructed centripetally by a designer and act according to his/her plan, and organisms, which as 'living plans' 'construct' themselves in a centrifugal fashion.

The two-way fit between organism and environment is what Maturana and Varela refer to as *structural congruence* between them, which is the result of their *structural coupling*:

> Ontogeny is the history of structural change in a unity without loss of organisation in that unity. This ongoing structural change occurs in the unity from moment to moment, either as a change triggered by interactions coming from the environment in which it exists or as a result of its internal dynamics. As regards its continuous interactions with the environment, the .. unity classifies them and sees them in accordance with its structure at every instant. That structure, in turn, continuously changes because of its internal dynamics. ...
>
> In these interactions, the structure of the environment only *triggers* structural changes in the autopoietic unities (it does not specify or direct them), and vice versa for the environment. The result will be a history of mutual congruent structural changes as long as the autopoietic unity and its containing environment do not disintegrate: there will be a *structural coupling*. (Maturana and Varela 1987, p. 74)

5.3.2 Structural Coupling and Situatedness

There is a close match between the views of agent-environment interaction expressed in the biologically based theories of von Uexküll and Maturana and Varela, and the notion of situatedness as discussed by Dreyfus. Maturana and Varela's above characterization of an organism as a "unity" that "classifies" its interactions with the environment "in accordance with its structure at every instant" is exactly what von Uexküll described as the subject imprinting its ego-quality on stimuli,

i.e. its objects, depending on its current behavioral disposition. Furthermore this is also what happens in recurrent neural robot controllers; in every time step incoming stimuli are interpreted based on the current biases and sensorimotor mapping. Hence, in Dreyfus and Heidegger's terms, the agent is "always already in a situation". Furthermore, that situation is constantly revised through the "structural coupling" between agent and environment, i.e. the interaction of internal and external dynamics resulting in what Maturana and Varela refer to as "a history of mutual congruent structural changes". This is the case in both von Uexküll's example of the tick and Meeden's Carbot. In both cases the main mechanism of being "always already in a situation" is that of structural coupling, which ensures a "history of mutual congruent structural changes" by means of a (diachronically structured) sensorimotor mechanism that unfolds itself into a causal sequence/chain of different 'internal' sensorimotor mechanisms that is finely 'tuned' to match the temporal sequence of varying situations and requirements[5] (for more detailed discussions of this aspect see Ziemke, 2000a, 2000b).

5.4 How Situated Robots Differ from Organisms

Having discussed the mechanisms of robotic situatedness in Section 5.1, the differences between situated robots and conventional mechanisms in Section 5.2, and the mechanisms of organismic situatedness in Section 5.3, this section will examine what exactly the (remaining) differences between situated robots and living organisms are.

As discussed in Section 4, modern AI research on the interaction between situated robots and their environments has, unlike the still pre-dominant computer metaphor, certainly taken a lot of inspiration from biology. Nevertheless, the robot's situatedness is very different from the living organism's. A robot might have adapted its control system, possibly even its physical structure to some degree, in interaction with its environment, and thus have acquired a certain degree of "epistemic autonomy" (Prem, 1997; cf. also Cariani, 1992). This (partial) self-organization, however,

[5] As pointed out by Manteuffel (1992), the continual context-dependent adaptation of behavioral dispositions allows neurally controlled robots to avoid the *frame problem* (Pylyshyn, 1987) that comes with the explicit style of representation (type 1) of traditional AI.

typically starts and ends with a bunch of physical parts and a computer program. Furthermore, the process is determined, started and evaluated by a human designer, i.e. the drive to self-organize does not lie in the robot's components themselves and success or failure of the process is not 'judged' by them either. For example, in the evolution of robot bodies (Section 5.1.2), there is no growth or adaptation of the individual robot body. Instead body plans are first evolved in the computer, i.e. 'outside' the robot, and then implemented in a robot body. In von Uexküll's terms (cf. Section 2), the evolution of the body plan might have followed centrifugal principles, the resulting robot bodies are, however, still built in a centripetal fashion and from then on can no longer self-organize. Hence, these bodies are not at all 'living plans' in von Uexküll's sense, which construct themselves, but they still are constructed according to an extrinsic plan. The components might be better integrated after having self-organized; they might even be considered 'more autonomous' for that reason, but they certainly do not become alive in that process, i.e. they remain allopoietic rather than autopoietic.

Any living organism, on the other hand, starts its self-organizing process from a single autonomous cellular unity (*Zellautonom*). The drive to self-organize is part of its 'building plan' (*Bauplan*), and it is equipped, in itself, with the resources to 'carry out that plan'. From the very beginning the organism is a viable unity, and it will remain that throughout the self-organizing process (until it dies). T. von Uexküll (1997) has pointed out that living organisms are autopoietic systems, which selectively assimilate parts of their environment and get rid of parts they do not need anymore. According to T. von Uexküll, selection and assimilation of the required elements can be described as sign processes, whose interpretants correspond to the living systems' biological needs. The criterion for the correctness of the interpretation described by the sign process is the successful assimilation. Today's robots, however, do not assimilate anything from their environment, and, as mentioned above, they have no intrinsic needs that the self-organizing process would have to fulfill to remain 'viable'. Following the arguments of von Uexküll, Maturana and Varela as well as Piaget, every organism can be considered a living, self-constructing and self-modifying 'hypothesis' in von Uexküll's sense of an 'acting plan', maintaining its viability through adaptation under environmental constraints. Hence, in the organism's case viability in the biological sense of survival and viability in

the sense of fit between behavioral/conceptual mechanisms and experience are closely connected. A robot, on the other hand, always embodies a human hypothesis. It 'lacks' the *intrinsic* requirement of biological viability. Hence, the viability of its behavioral/conceptual mechanisms can ultimately always only be evaluated from the outside (with respect to fitness function, reinforcement, error measures, etc.). Thus, for the robot the only criterion of success or failure is still the designer's and/or observer's evaluation or interpretation, i.e. this criterion is entirely *extrinsic* to the robot.

A key problem with New AI research on adaptive robots artificial life, we believe, is that, despite claims to the contrary and despite the emphasis of 'embodiment', many researchers are still devoted to the *computationalist/functionalist* view of *medium independence*, i.e. the idea that the "characteristics of life and mind are independent of their respective material substances" (Emmeche 1992, p. 471; cf. also Section 3.1). Much research effort is spent on control mechanisms, or 'artificial nervous systems', and how to achieve certain behaviours in robots through self-organization of these control mechanisms. However, to compare a robot's 'artificial nervous system' to an animal's nervous system, because they exhibit 'the same behavior', implies that the relation between behavior and (artificial) nervous system is actually independent of the controlled body. In other terms, it implies that the operation of the nervous system is computational and largely independent of the body it is carried out in. That means, the body is reduced to the computational control system's sensorimotor interface to the environment.

Hence, several researchers have argued that robotic agents are, at least in theory, capable of possessing 'first hand semantics' or 'intrinsic meaning' (e.g., Harnad, 1990; Brooks, 1991b; Franklin, 1997; Bickhard, 1998). In particular, it is held, their epistemological interest arises from the fact, that unlike conventional machines, their use of signs and representations is often self-organized, and thus, as for living systems, largely private and typically only meaningful to themselves. Many researchers therefore no longer draw a strict line between animals and robots. An example of this view is Beer's (1995) above characterization of animals and robots as 'autonomous agents' (cf. Section 4.1). Another example is Prem (1998) who refers to both these categories as 'embodied autonomous

systems', and does not at all distinguish between living and non-living in his discussion of semiosis in such systems. An even more extreme view, although not at all uncommon in New AI circles, is that of Franklin (1997) who argued that "[s]oftware systems with no body in the usual physical sense can be intelligent" if only they are "embodied in the situated sense of being autonomous agents structurally coupled with their environment".

Maturana and Varela, however, have argued (again, similar to von Uexküll (1928); cf. also Hoffmeyer, 1996), that in living organisms body and nervous system are not at all separate parts:

> ... the nervous system contains millions of cells, but all are integrated as components of the organism. Losing sight of the organic roots of the nervous system is one of the major sources of confusion when we try to understand its effective operation. (Maturana and Varela 1987, p. 34)

Similarly, T. von Uexküll *et al.* (1993), in their discussion of *endosemiosis* (sign processes inside the organism), point out that the living body, which we experience to be the centre of our subjective reality (*Wirklichkeit*), is the correlate of a neural *counterbody* (*Gegenkörper*) which is formed and updated in our brain as a result of the continual information flow of proprioceptive signs from the muscles, joints and other parts of our limbs. This neural counterbody is the center of the earlier discussed neural counterworld (cf. von Uexküll, 1909, 1985; cf. Section 2), created and adapted by the brain from the continual stream of signs from the sensory organs. According to T. von Uexküll *et al.*, counterbody and counterworld form an undividable unity, due to the fact that all processes/events we perceive in the world really are 'countereffects' to real or potential effects of our motor-system, and together with these they form the spatial structure within which we orient ourselves. A robot, on the other hand, has no endosemiosis whatsoever in the body (its physical components) as such. Thus, there is no integration, communication or mutual influence of any kind between parts of the body, except for their purely mechanical interaction. Further, there is no meaningful integration of the 'artificial nervous system' and the physical body, beyond the fact that some parts of the body provide the control system with sensory input, which in turn triggers the motion of some other parts of the body (e.g., wheels) (cf. also Sharkey and Ziemke, 1998; Ziemke and Sharkey, in press). Thus, New

AI, although acknowledging the role of the physical, is still largely ‘stuck’ in the old distinction between hardware and software, which was central to computationalism and traditional AI (cf. Section 3.1).

In summary, it can be said that, despite all biological inspiration, today’s situated robots are still radically different from living organisms. In particular, despite their capacity for a certain degree of self-organization, today’s so-called ‘autonomous’ agents are actually far from possessing the autonomy of living organisms. Mostly, this is due to the fact that today’s robots typically are composed of mechanical parts (hardware) and control programs (software). The autonomy and subjectivity of living systems, on the other hand, emerges from the interaction of their components, i.e. autonomous cellular unities (*Zellautonome*). Meaningful interaction between these first-order unities, and between the resulting second-order unity and its environment, is a result of their structural congruence, as pointed out by von Uexküll as well as Maturana and Varela (cf. previous subsection). Thus, autonomy is a property of a living organism’s organization right from its beginning as an autonomous cellular unity, and initial structural congruence with its environment results from the specific circumstances of reproduction. Its ontogeny maintains these properties throughout its lifetime through structural coupling with its environment. Providing artifacts with the capacity for self-organization and constructive processes can be seen as the attempt to provide them with an artificial ontogeny. However, the attempt to provide them with autonomy this way seems to be doomed to fail, since it follows from the above argument that autonomy cannot from the outside be ‘put’ into a system, that does not already ‘contain’ it. Ontogeny preserves the autonomy of an organization, it does not ‘construct’ it. The attempt to bring the artifact into some form of structural congruence with its environment, on the other hand, can ‘succeed’, but only in the sense that the criterion for congruence cannot lie in the heteronomous artefact itself, but must be in the eye of the observer. This is exactly what happens when a robot is trained to adapt its structure in order to solve a task defined by its designer.

A major problem with New AI and adaptive robotics research, we believe, is that, despite its strong biological inspiration, it has focused on establishing itself as a new paradigm *within* AI and cognitive science. Relatively little effort has been made to make the connection to other theories and disciplines, where issues like autonomy, representation, situatedness and embodiment have been discussed for a long time, although not necessarily under those names. In particular the New AI distinguishes itself from its traditional counterpart in its view of knowledge as sensorimotor transformation knowledge. Thus, although many researchers do not recognize that this is an 'old' idea, recent work in adaptive robotics is largely compatible with the radically constructivist view of the construction of such knowledge through sensorimotor interaction with the environment with the goal of achieving some 'fit' or 'equilibrium' between internal behavioral/conceptual structures and experience of the environment. However, the organic roots of these processes which were emphasized by von Uexküll and Piaget (as well as others such as Merleau-Ponty, Searle, Maturana and Varela), are often ignored in the New AI. Instead its view of the body is largely compatible with mechanistic theories, whereas the view of control mechanisms is in fact still largely compatible with computationalism. That means the robot body is typically viewed as some kind of input- and output device (one with sensors and motors instead of the computer's keyboard and monitor) that provides 'physical grounding' to the internal computational mechanisms. Thus the New AI has in practice become a theoretical hybrid, or in fact a 'tribrid', combining (1) a mechanistic view of the body's role with (2) the constructivist notion of sensorimotor transformation knowledge, and with (3) the functionalist/computationalist hardware-software distinction and its notion of the activity of the nervous system as computation, i.e. largely medium-independent information or signal processing.

Hence, Searle's (1980) general critique of exactly that computationalist hardware-software distinction, as well his rejection of the 'robot reply' in particular (cf. Section 3.2.2), still must be said to hold for most adaptive robotics research. As in the Chinese Room Argument, the robot body is typically reduced to an input-output device with no further influence on the software core of the control mechanism, i.e. the activity 'inside the room', which consists of hardware-independent computational processes in the functionalist sense (cf. Section 3.1.2). Thus, from the perspective of

AI and cognitive science, adaptive robotics research seems to be stuck in what might be called the 'software trap', due to its acceptance of the functionalist/computationalist hardware-software distinction. It is therefore highly questionable that the New AI could ever 'escape' the Chinese Room as long as sticks to that distinction. It might be argued that work on evolvable hardware blurs that distinction. However, from the arguments of von Uexküll as well as those of Maturana and Varela discussed in the previous section, it clearly follows that even today's so-called 'evolvable hardware' is allopoietic and heteronomous, no matter whether it can re-configure 'itself' by means of artificial evolutionary methods or not. In a nutshell, this is why, despite the emphasis on 'situatedness', 'embodiment' and 'autonomy', even New AI and adaptive robotics research, in its current form, cannot be a 'strong AI' either (cf. Section 3.2.2). On the one hand, as long as AI researchers are committed to computationalism and 'software intelligence', their models will always fall foul of Searle's Chinese Room/Gym argument. If AI researchers counted on 'hardware intelligence' instead, on the other hand, they would have to face the fact that they would have to build autopoietic hardware, i.e. *construct* truly *self-constructing* hardware; something that, at least with current (man-made) hardware technology, seems to be impossible.

5.5 Summary and Conclusion

As discussed in Section 2, in the days of von Uexküll (1928) organisms and mechanisms/artifacts clearly could be distinguished by the way they were 'constructed'. The latter were the objects of human design and construction and they 'interacted' with their environment in a relatively passive and merely mechanical manner. Organisms, on the other hand, could be said to autonomously construct themselves according to their own 'building-plan', both in the biological and the psychological/conceptual sense, and to imprint their own subjective qualities on all interaction with their physical environment. Traditional AI, as discussed in Section 3, took a completely different route that involved neither (direct) interaction with any environment nor autonomous construction of any kind. Instead it focused on human-constructed computational, i.e. purely formally defined, systems, which therefore only by an observer could be interpreted to be models of cognition and

intelligent behavior. The 'New AI', on the other hand, as discussed in Section 4, aims to reduce this observer-dependence by (a) physically situating artificial agents in an environment (possibly simulated), such that they can interact with it directly and 'on their own', and (b) in the case of adaptive robots, equipping it with some capacity of self-construction/-organization. The latter is intended to provide the robot with the means to 'autonomously' construct its own 'subjective' view of the world and what might be called conceptual/phenomenal situatedness. We discussed constructive processes at several levels and time scales, from a short-term sense of situatedness through self-revision of behavioral dispositions in recurrent neural robot controllers, to the construction of complete 'artificial nervous systems' and bodies. It could be noted that much of adaptive robotics research is in fact largely compatible with radical constructivist views of sensorimotor transformation knowledge and its autonomous construction driven by the need to viably fit environmental constraints. However, we argued in Section 5.4, while the use of computational learning techniques has provided adaptive robots with some aspects of the situatedness and constructive processes that von Uexküll considered limited to organisms, ultimately these robots remain heteronomous, i.e. their 'viability' is still always in the eye of the observer. Hence, they might very well turn out be useful tools for the study of constructive processes, i.e. a 'weak constructivist AI' or 'synthetic radical constructivism', but they will always lack the deeply rooted, constructed nature of the biological embodiment and situatedness of living organisms.

Acknowledgements

The author would like to thank Claus Emmeche, Brian Goodwin, Gerhard Manteuffel, John Stewart and the anonymous reviewers for very useful comments on an earlier version of this paper. Moreover, the author would like to thank Noel Sharkey, who supervised the author's doctoral dissertation (Ziemke, 2000a) of which the work presented here is part. Financially the author has been supported by a grant (1507/97) from the Knowledge Foundation, Stockholm, Sweden.

References

Andersen, Peter B., Hasle, Per, and Brandt, Per A. (1997). Machine semiosis. In Posner, R., Robering, K., and Sebeok, T. A., editors, *Semiotik / Semiotics - Ein Handbuch zu den zeichentheoretischen Grundlagen von Natur und Kultur / A Handbook on the Sign-Theoretic Foundations of Nature and Culture*, pages 548-571. Berlin / New York: Walter de Gruyter.

Beer, Randy D. (1995). A dynamical systems perspective on autonomous agents. *Artificial Intelligence*, 72:173-215.

Bickhard, Mark H. (1998). Robots and representations. In *From animals to animats 5 - Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, pages 58-63. Cambridge, MA: MIT Press.

Bickhard, Mark H. and Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science - Impasse and Solution*. New York, NY: Elsevier.

Braitenberg, Valentino (1984). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Brooks, Rodney A. (1986a). Achieving artificial intelligence through building robots. Technical Report Memo 899, MIT AI Lab.

Brooks, Rodney A. (1986b). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2:14-23.

Brooks, Rodney A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1-2):1-16.

Brooks, Rodney A. (1991a). Intelligence Without Representation. *Artificial Intelligence*, 47:139-159.

Brooks, Rodney A. (1991b). Intelligence Without Reason. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 569-595. San Mateo, CA: Morgan Kaufmann.

Cariani, Peter (1992). Some epistemological implications of devices which construct their own sensors and effectors. In Varela, F. J. and Bourgine, P., editors, *Toward a practice of autonomous systems - Proceedings of the First European Conference on Artificial Life*, pages 484-493. Cambridge, MA: MIT Press.

Clancey, William J. (1997). *Situated Cognition: On Human Knowledge and Computer Representations*. New York: Cambridge University Press.

Clark, Andy (1997). *Being There - Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Cliff, Dave T. (1991). Computational Neuroethology: A Provisional Manifesto. In Meyer, J.-A. and Wilson, S. W., editors, *From Animals to Animats*, pages 29-39. Cambridge, MA: MIT Press.

Cliff, Dave T. and Miller, G. F. (1996). Co-evolution of pursuit and evasion II: Simulation methods and results. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J. B., and Wilson, S. W., editors, *From animals to animats 4 - Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 506-515. Cambridge, MA: MIT Press.

Craik, Kenneth J. W. (1943). *The Nature of Explanation*. Cambridge, UK: Cambridge University Press.

Dorffner, Georg (1997). Radical connectionism - a neural bottom-up approach to AI. In Dorffner, G., editor, *Neural Networks and a New Artificial Intelligence*, pages 93-132. London, UK: International Thomson Computer Press.

Dreyfus, Hubert L. (1979). *What Computers Can't Do – A Critique of Artificial Reason* (revised edition). New York: Harper & Row.

Dreyfus, Hubert L. (1996). The Current Relevance of Merleau-Ponty's Phenomenology of Embodiment. *The Electronic Journal of Analytic Philosophy*, 4. Originally appeared in Haber, H. and Weiss, G., editors, *Perspectives on Embodiment*, New York: Routledge.

Driesch, H. (1931). *Das Wesen des Organismus*. Leipzig, Germany.

Elman, Jeffrey (1990). Finding Structure in Time. *Cognitive Science*, 14:179-211.

Emmeche, Claus (1990). Kognition og omverden – om Jakob von Uexküll og hans bidrag til kognitionsforskningen. *Almen Semiotik*, 2:52-67.

Emmeche, Claus (1992). Life as an abstract phenomenon: Is artificial life possible?. In Varela, F. J. and Bourgine, P., editors, *Toward a practice of autonomous systems - Proceedings of the First European Conference on Artificial Life*, pages 466-474. Cambridge, MA: MIT Press.

Emmeche, Claus (in press). Does a robot have an Umwelt? Reflections on the qualitative biosemiotics of Jakob von Uexküll. *Semiotica*, special issue on the work of Jakob von Uexküll, to appear in late 2000.

Fodor, Jerry A. (1987). Psychosemantics. Cambridge, MA: MIT Press.

Franklin, Stan (1995). *Artificial Minds*. Cambridge, MA: MIT Press.

Franklin, Stan (1997). Autonomous agents as embodied AI. *Cybernetics and Systems*, 28(6):499-520.

Funes, Pablo and Pollack, Jordan B. (1997). Computer evolution of buildable objects. In *Proceedings of the Fourth European Conference on Artificial Life*, pages 358-367. Cambridge, MA: MIT Press.

Hallam, John C. and Malcolm, Chris A. (1994). Behaviour: Perception, Action and Intelligence – The View from Situated Robotics. *Proc. Royal Society Land A,* 349:29-42.

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

Heidegger (1962). *Being and Time*. New York: Harper & Row. Originally appeared as Heidegger, M. (1927). *Sein und Zeit*. Tübingen, Germany.

Hendriks-Jansen, Horst (1996). *Catching Ourselves in the Act – Situated Activity, Interactive Emergence, Evolution, and Human Thought*. Cambridge, MA: MIT Press.

Hoffmeyer, Jesper (1996). *Signs of Meaning in the Universe*. Bloomington: Indiana University Press.

Husbands, Phil; Harvey, Inman and Cliff, Dave (1993). An Evolutionary Approach to Situated AI. In Sloman, A.; Hogg, D.; Humphreys, G.; Ramsay, A. and Partridge, D., editors, *Prospects for Artificial Intelligence*, pages 61-70. Amsterdam: IOS Press.

Husbands, Phil; Smith, Tom; Jakobi, Nick and O'Shea, Michael (1998). Better Living Through Chemistry: Evolving GasNets for Robot Control. *Connection Science*, 10(3-4):185-210.

Johnson-Laird, Philip N. (1989). *Mental Models*. In Posner, M. I., editor, *Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Kant, Immanuel (1781/7). Kritik der reinen Vernunft. In *Kants Werke*, Akademieausgabe, Vol. IV, Berlin.

Lakoff, George (1988). Smolensky, semantics, and the sensorimotor system. *Behavioral and Brain Sciences*, 11:39-40.

Langthaler, Rudolf (1992). *Organismus und Umwelt – Die biologische Umweltlehre im Spiegel traditioneller Naturphilosophie*. Hildesheim, Germany: Georg Olms Verlag.

Lenat, Doug and Feigenbaum, Edward P. (1991). On the Thresholds of Knowledge. *Artificial Intelligence*, 47(1-3):199.

Loren, Lewis A. and Dietrich, Eric (1997). Merleau-Ponty, Embodied Cognition and the Problem of Intentionality. *Cybernetics and Systems*, 28:345-358.

Lorenz, Konrad (1957). The Nature of Instinct: The Conception of Instinctive Behavior. In Schiller, C. H., editor, *Instinctive Behaviour - The Development of a Modern Concept*, pages 129-175. New York: International Universities Press. Originally appeared as Lorenz, K. (1937) Über die Bildung des Instinktbegriffes, *Die Naturwissenschaften*, 25:289-300,307-318,324-331.

Lipson, Hod and Pollack, Jordan B. (2000). Evolution of machines. In *Proceedings of the International Conference on Artificial Intelligence in Design*. Worchester, MA.

Lund, Henrik H., Hallam, John, and Lee, W. (1997). Evolving robot morphology. In *Proceedings of the IEEE Fourth International Conference on Evolutionary Computation*. IEEE Press.

Lund, Henrik H. and Miglino, Orazio (1998). Evolving and breeding robots. In *Proceedings of the First European Workshop on Evolutionary Robotics*. Berlin/Heidelberg, Germany: Springer Verlag.

Manteuffel, Gerhard (1992). Konstruktivistische künstliche Intelligenz. In Schmidt, S. J., editor, *Kognition und Gesellschaft – Der Diskurs des Radikalen Konstruktivismus 2*. Frankfurt a. M., Germany: Suhrkamp Verlag.

Maturana, Humberto R. and Varela, Francisco J. (1980). *Autopoiesis and Cognition - The Realization of the Living*. Dordrecht, The Netherlands: D. Reidel Publishing.

Maturana, Humberto R. and Varela, Francisco J. (1987). *The Tree of Knowledge - The Biological Roots of Human Understanding*. Shambhala, Boston, MA. NB: All page numbers refer to the revised edition of 1992.

Meeden, Lisa A. (1996). An incremental approach to developing intelligent neural network controllers for robots. *IEEE Transactions on Systems, Man, and Cybernetics*, 26.

Meeden, Lisa A., McGraw, Gary, and Blank, Douglas (1993). Emergence of control and planning in an autonomous vehicle. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*, pages 735-740. Hillsdale, NJ: Lawrence Erlbaum.

Merleau-Ponty, Maurice (1962). *Phenomenology of Perception*. London: Routledge & Kegan Paul. Originally appeared as Merleau-Ponty (1945) *Phenomenologie de la Perception*, Paris: Gallimard.

Merleau-Ponty, Maurice (1963). *The Structure of Behavior*. Boston, MA: Beacon Press. Originally appeared as Merleau-Ponty (1942) *La Structure du Comportment*, Presses Universites de France.

Minsky, Marvin (1975). A Framework for Representing Knowledge. In Winston, P., editor, *The Psychology of Computer Vision*, pages 211-277. McGraw-Hill.

Mondada, Francesco, Franzi, E., and Ienne, P. (1993). Mobile robot miniaturisation: A tool for investigating in control algorithms. In *Third International Symposium on Experimental Robotics*, Kyoto, Japan.

Müller, Johannes (1840). *Handbuch der Physiologie des Menschen,* Band 2. Koblenz, Germany.

Neisser, Ulric (1967). *Cognitive Psychology*. New York: Appelton.

Newell, Alan (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, Alan and Simon, Herbert A. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19:113-126.

Nolfi, Stefano (1998). Evolutionary Robotics: Exploiting the full power of self-organisation. *Connection Science*, 10(3-4):167-184.

Nolfi, Stefano and Floreano, Dario (1998). Co-evolving predator and prey robots: Do 'arm races' arise in artificial evolution? *Artificial Life*, 4(4).

Nolfi, Stefano and Floreano, Dario (1999). Learning and Evolution. *Autonomous Robots*, 7(1).

Nolfi, Stefano and Floreano, Dario (2000). *Evolutionary Robotics*. Cambridge, MA: MIT Press.

Peschl, Markus (1997). The representational relation between environmental structures and neural systems: Autonomy and environmental dependency in neural knowledge representation. *Nonlinear Dynamics, Psychology and Life Sciences*, 1(2):99-121.

Peschl, Markus and Riegler, Alexander (1999). Does Representation Need Reality. In Riegler, A.; Peschl, M. and von Stein, A., editors, *Understanding Representation in the Cognitive Sciences*, pages 9-18. New York: Plenum Press.

Pfeifer, Rolf and Scheier, Christian (1999). *Understanding Intelligence*. Cambridge, MA: MIT Press.

Piaget, Jean (1954). *The Construction of Reality in the Child*. New York: Basic Books. Originally appeared as Piaget (1937) *La construction du réel chez l'enfant*. Neuchâtel, Switzerland: Delachaux et Niestlé.

Piaget, Jean (1967). *Six Psychological Studies*. New York: Vintage.

Pollack, Jordan B. (1991). The induction of dynamical recognizers. *Machine Learning*, 7:227-252.

Prem, Erich (1997). Epistemic autonomy in models of living systems. In *Proceedings of the Fourth European Conference on Artificial Life*, pages 2-9. Cambridge, MA: MIT Press.

Prem, Erich (1998). Semiosis in embodied autonomous systems. *In Proceedings of the IEEE International Symposium on Intelligent Control*, pages 724-729. Piscataway, NJ: IEEE.

Pylyshyn, Zenon, editor (1987). *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Norwood: Ablex Publishing.

Richards, Robert J. (1987). *Darwin and the emergence of evolutionary theories of mind and behavior*. Chicago: The University of Chicago Press.

Riegler, Alexander (1994). Constructivist Artificial Life: The constructivist-anticipatory principle and functional coupling. In Hopf, J., editor, *Genetic Algorithms with the Framework of Evolutionary Computation*, pages 73-83. Max-Planck-Institute für Informatik, MPI-I-94-241, Saarbrücken, Germany.

Riegler, Alexander (1997). Ein kybernetisch-konstruktistisches Modell der Kognition. In Müller, A; Müller, K. H. and Stadler, F., editors, *Konstruktivismus und Kognitionswissenschaft. Kulturelle Wurzeln und Ergebnisse*, pages 75-88. Vienna, New York: Springer.

Risku, Hanna (2000). Situated Translation und Situated Cognition – ungleiche Schwestern. In Kadric, M.; Kaindl, K. and Pöchhacker, F., editors, *Translations wissenschaft. Festschrift für Mary Snell-Hornby*, pages 81-91. Tübingen: Stauffenburg.

Rumelhart, David E. and McClelland, Jay L. (1986). On learning the past tense of English verbs. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2. Psychological and Biological Models*, pages 216-271. Cambridge, MA: MIT Press.

Schank, Roger C. (1972). Conceptual Dependency: A Theory of Natural Language Understanding. *Cognitive Psychology*, 3:552-631.

Schank, Roger C. (1975). Using Knowledge to Understand. *Theoretical Issues in Natural Language Processing*. Cambridge, MA.

Schank, Roger C. & Abelson, Robert P. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Searle, John (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3:417-457.

Searle, John (1990). Is the Brain's Mind a Computer Program? *Scientific American*, January 1990:20-25.

Searle, John (1991). Consciousness, Explanatory Inversion and Cognitive Science. *Behavioral and Brain Sciences*, 13:585-642.

Sejnowski, Terrence and Rosenberg, C. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145-168.

Sharkey, Noel E. (1991). Connectionist representation techniques. *Artificial Intelligence Review*, 5:143-167.

Sharkey, Noel E. and Ziemke, Tom (1998). A consideration of the biological and psychological foundations of autonomous robotics. *Connection Science*, 10(3- 4):361-391.

Sjölander, Sverre (1999). How Animals Handle Reality – The Adaptive Aspect of Representation. In Riegler, A.; Peschl, M. and von Stein, A., editors, *Understanding Representation in the Cognitive Sciences*, pages 277-282. New York: Plenum Press.

Stewart, John (1996). Cognition = Life: Implications for higher-level cognition. *Behavioural Processes*, 35:311-326.

Suchman, Lucy A. (1987). *Plans and Situated Action: The Problem of Human-Machine Communication*. New York: Cambridge University Press.

Varela, Francisco J.; Thompson, Evan and Rosch, Eleanor (1991). *The Embodied Mind Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

von Glasersfeld, Ernst (1995). *Radical Constructivism – A Way of Knowing and Learning.* London: Falmer Press.

von Uexküll, Jakob (1909). *Umwelt und Innenwelt der Tiere*. Berlin: Springer Verlag.

von Uexküll, Jakob (1928). *Theoretische Biologie*. Berlin: Springer Verlag. NB: All page numbers refer to the first paperback edition, 1973, Frankfurt/Main, Germany: Suhrkamp.

von Uexküll, Jakob (1957). A stroll through the worlds of animals and men - a picture book of invisible worlds. In Schiller, C. H., editor, *Instinctive Behaviour - The Development of a Modern Concept*, pages 5-80. New York: International Universities Press. Appeared also in *Semiotica*, 89(4):319-391. Originally appeared as von Uexküll (1934) *Streifzüge durch die Umwelten von Tieren und Menschen*, Berlin: Springer.

von Uexküll, Jakob (1982). The Theory of Meaning. Semiotica, 42(1):25-82.

von Uexküll, Jakob (1985). Environment [Umwelt] and Inner World of Animals. In Burghardt, G. M., editor, *Foundations of Comparative Ethology*. New York: Van Nostrand Reinhold. Partial translation of von Uexküll (1909) *Umwelt und Innenwelt der Tiere*, Berlin: Springer.

von Uexküll, Thure (1992). Introduction: The sign theory of Jakob von Uexküll. *Semiotica*, 89(4):279-315. Originally appeared as von Uexküll, T (1987) The sign theory of Jakob von Uexküll, in Krampen, M. *et al.*, editors, *Classics of Semiotics*, pages 147-179, New York: Plenum.

von Uexküll, Thure (1997). Biosemiose. In Posner, R., Robering, K., and Sebeok, T. A., editors, *Semiotik / Semiotics - Ein Handbuch zu den zeichentheoretischen Grundlagen von Natur und Kultur / A Handbook on the Sign-Theoretic Foundations of Nature and Culture*, pages 447-457. Berlin / New York: Walter de Gruyter.

von Uexküll, Thure; Geigges, Werner, and Herrmann, Jörg M. (1993). Endosemiosis. *Semiotica*, 96(1/2):5-51.

Wilson, Stewart W. (1985). Knowledge growth in an artificial animal. In Grefenstette, J., editor, *Proceedings of an International Conference on Genetic Algorithms and Their Applications*, pages 16-23. Hillsdale, NJ: Lawrence Erlbaum.

Wilson, Stewart W. (1991). The animat path to AI. In Meyer, J.-A. and Wilson, Stewart, editors, *From Animals to Animats: Proceedings of The First International Conference on Simulation of Adaptive Behavior*, pages 15-21. Cambridge, MA: MIT Press.

Woods, W. A. (1975). What's in a Link: Foundations for Semantic Networks. In Bobrow, D. G. and Collins, A. M., editors, *Representation and Understanding: Studies in Cognitive Science*, pages 35-82. Academic Press.

Ziemke, Tom (1997). The 'Environmental Puppeteer' Revisited: A Connectionist Perspective on 'Autonomy'. In *Proceedings of the 6th European Workshop on Learning Robots (EWLR-6)*, pages 100-110, Brighton, UK.

Ziemke, Tom (1998). Adaptive Behavior in Autonomous Agents. *Presence*, 7(6):564-587.

Ziemke, Tom (1999a). Remembering how to behave: Recurrent neural networks for adaptive robot behavior. In Medsker, L. and Jain, L. C., editors, *Recurrent Neural Networks: Design and Applications*. New York: CRC Press.

Ziemke, Tom (1999b). Rethinking Grounding. In Riegler, A.; Peschl, M. and von Stein, A., editors, *Understanding Representation in the Cognitive Sciences*, pages 177-190. New York: Plenum Press.

Ziemke, Tom (2000a). *Situated Neuro-Robotics and Interactive Cognition*. Doctoral Dissertation, Department of Computer Science, University of Sheffield, UK.

Ziemke, Tom (2000b). On 'Parts' and 'Wholes' of Adaptive Behavior: Functional Modularity and Diachronic Structure in Recurrent Neural Robot Controllers. In *From animals to animats 6 - Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press.

Ziemke, Tom and Sharkey, Noel. E. (in press). A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life. *Semiotica*, special issue on the work of Jakob von Uexküll, to appear in late 2000.