

The Adaptive Behavior Approach to Psychology

Bram Bakker

Unit of Experimental and Theoretical Psychology

Leiden University

P.O. Box 9555; 2300 RB, Leiden

The Netherlands

`bbakker@fsw.leidenuniv.nl`

Abstract

This paper discusses a rapidly evolving field with great relevance for cognitive psychology, the field of adaptive behavior. Its goal is to learn about biological behavior by constructing agents exhibiting that behavior, but it is very different from traditional artificial intelligence. It is argued that rather than providing mere implementations of the abstract theories of cognitive psychology, the adaptive behavior approach yields many completely new insights, as well as indications that some of mainstream cognitive psychology's assumptions and theoretical ideas may be in need of revision. Specific examples of adaptive behavior research are presented to support these claims.

1 Introduction

Cognitive psychology is the subdiscipline of psychology that is concerned with the basic mechanisms underlying human behavior. Its ultimate goal is to understand how language is produced and perceived, how memory works, how perception comes about, how muscles are controlled, how emotions work, etc.; in short, how all the faculties function that make up the mind. Derived from that goal are the goals of understanding how those faculties may fail and how they behave in different circumstances, and applying the insights of cognitive psychology to clinical, educational, or industrial settings.

In recent years, forceful arguments have been put forward to the effect that cognitive psychology will and must become more intimately related to the neural sciences (Crick, 1988; Churchland, 1986). Among other developments, technological advances have made it possible to “look” inside the active brain, using techniques with acronym names such as PET, fMRI, ERP, and MEG. The new insights into brain mechanisms that are likely to emerge from those developments will have to be taken into account by cognitive psychology.

This paper, however, discusses another rapidly evolving field with great relevance for cognitive psychology. It is the field of *adaptive behavior*, alternatively referred to as the field of autonomous agents, animats, behavior-based cognitive science, bottom-up cognitive science, or synthetic psychology. This field takes an artificial intelligence approach to psychology (although it is very different from traditional artificial intelligence), in that systems are constructed that are

capable of intelligent behavior. The idea is that by constructing an artificial system one can learn something about the biological system. After all, in the process of construction one may encounter the same problems that the biological system encounters, and hypotheses about how the biological system overcomes these problems can be put to the test.

One might expect that this approach amounts to straightforward “implementation” of the abstract theories of mainstream cognitive psychology into concrete models. At best, this would lead to arbitration between competing abstract theories, and the filling in of some of the details left open by those abstract theories. In contrast, I will argue that the adaptive behavior approach yields insights that are far more revolutionary. The systems that are constructed and that work best are, in many cases, very different from any of the systems described in the mainstream cognitive psychology literature. In those cases where attempts were made to implement existing abstract theories directly, often insurmountable problems have appeared—suggesting falsification of those theories rather than a mere filling in of the details.

To back up this argument it is necessary to describe the current standard view of human and animal information processing in cognitive psychology, and to present results from the adaptive behavior approach that are different from or even incompatible with that view. Section 2 contains a description of current mainstream cognitive psychology, together with some representative examples. Section 3 describes the adaptive behavior approach by contrasting it with mainstream cognitive psychology. Section 4 presents a number of examples of artificial systems or “agents” that are very different in architecture and functioning from what was or would be expected given the standard view. In that way they serve to back up the argument that some of the assumptions and ideas in mainstream cognitive psychology may be in need of revision, and that a closer relationship between cognitive psychology and adaptive behavior research is needed. Section 5, finally, contains a general discussion of the ideas presented in this paper.

2 Mainstream cognitive psychology

2.1 Theoretical foundations

After the behaviorist era with its exclusive focus on behavior, the cognitive revolution halfway through the 20th century brought renewed attention for the internal mechanisms underlying behavior. Theoretical ideas regarding those mechanisms were heavily influenced by another development of that time. The theory of computation showed that a very large class of mathematical functions could be performed by machines that mechanistically follow stored procedures, computers. In fact, it was made plausible that no physical machine could reliably perform functions beyond this class. Importantly, all computable functions can be performed by one and the same machine, the “hardware”, only by changing the stored procedures, the “software”. Actual computer applications displayed capabilities such as arithmetic, pattern recognition, and game playing, which were previously thought to require human intelligence, and which in many cases outperformed human capabilities.

Ever since then, the internal mechanisms of the mind are theorized to amount

to a kind of computer. Combining several ideas from computation theory, Newell and Simon (1976) stated the *symbol system hypothesis*: for a system to be capable of human-level intelligence, it is both necessary and sufficient for it to constitute a symbol system, a computer. Just like a computer, the brain is viewed as a single piece of hardware capable of a wide variety of tasks. One idea of computation theory has been particularly important in that respect. That idea is that a computer can in principle be realized (implemented) in many different types of hardware, whether they are mechanical switches, electronic circuits, or even beer cans. Thus, an implementation in neural tissue is also possible. The end result of the computation is not dependent on the specific implementation, in the same sense in which Microsoft Word works equally well (or equally badly) on Windows computers and Macs. In that sense, the hardware is irrelevant. This has led to a widespread doctrine called *functionalism*, which states that for this reason cognitive psychology needs not be concerned with the physical implementation, the brain, but only needs to focus on the functional level, the “software running on” the brain (e.g. Pylyshyn, 1984).

2.2 Functions and functional modules

The theoretical foundations described in the previous paragraph should be considered in combination with a number of practical principles of scientific methodology to understand current cognitive psychology. These practical principles amount to ideas about how research on the mechanisms underlying behavior should proceed. It is clear that the entire human brain or mind is too complex and multi-faceted to oversee and assess for single researchers, or to describe all aspects and details of its functioning in terms of a single simple model. For this reason, and influenced by the computer analogy, the system is decomposed into different *functions*: perception (itself sorted by sensory modality), memory, language, attention, motor control, reasoning, etc. In practice, each function has its own, more or less separate community of researchers and a distinct body of literature. Figure 1 presents a fairly characteristic general overview of the human information processing system as it appears in many textbooks.

Within each function, a similar decomposition into constituent functional components or modules is assumed. Each component is dedicated to some subfunction. Information is exchanged between components through communication lines although, in close analogy with computer systems, these communication lines are often assumed to have a “limited bandwidth”, i.e. have a limited capacity. Exchanged information consists of the results computed by one module which are the prerequisite for another module. Many models assume strictly serial processing; an array of modules activated one after the other, each one using the final results of the previous module as input. Other models, recent ones in particular, allow parallel processing. Multiple modules may be active at the same time and possibly interact, and the results are integrated into a final, response module. Components are usually represented visually as boxes and communication lines as arrows, yielding what are known as box-arrow models. Figures 2 and 3 show two well-known models representing this approach, one model of language production (Levitt, Roelofs, & Meyer, 1999) and one model of working memory (Baddeley, 1990).

To be sure, the idea of decomposition into functions and functional modules is not based on practical considerations and the computer analogy alone.

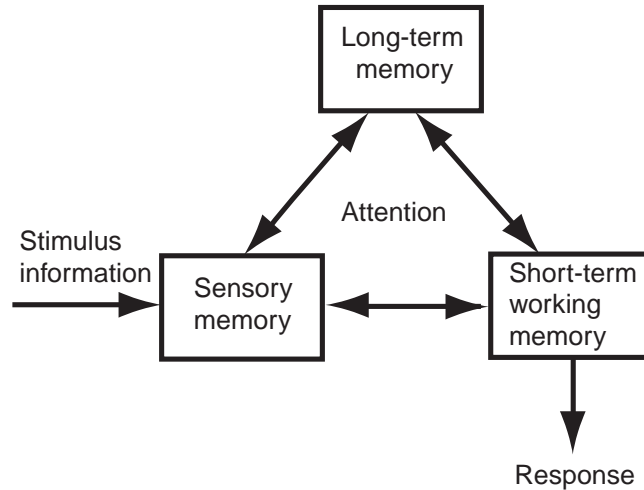


Figure 1: An overview of the human information processing system. Adapted from Ashcraft (1998). Similar overviews appear in many textbooks.

Even though functionalism dictates that the hardware of the brain is, in an important sense, irrelevant for the functioning of the mind, cognitive psychologists have been quick to embrace findings from the neural sciences that seem to support the idea of decomposition into functions. It is known from studies where animal brains were purposely lesioned as well as from brain imaging and cell recording studies that different anatomical regions in brains are involved in different aspects of animal and human functioning. Furthermore, neuropsychological studies show that people with accidental brain lesions, e.g. caused by car accidents, strokes, or bullet wounds, often display fairly specific defects rather than an overall deterioration of functioning. For instance, a person may lose the ability to produce language but still comprehend language, or a person may be able to visually recognize all objects except faces.

Many models in mainstream cognitive psychology assume as one of the functional modules some kind of central workspace or central processor where information from different sources comes together, where selections are being made, where information is temporarily stored and manipulated, and where decisions are made regarding actions. In figure 1, for instance, there is a short-term working memory, which, guided by attention, receives information from long-term memory and from sensory memory. In working memory this information is processed, and the appropriate response is given as output. Figure 3 depicts a central executive, which has, as slave systems, temporary stores for auditory information and visuo-spatial information, and which uses that information for its reasoning and decision processes. Figure 2 has a central executive in disguise, in the form of the self-monitoring process. Central processors and workspaces are based on the way computers function. There are hardly any data from the neural sciences which indicate that such central processors or workspaces actually exist in animal and human brains (Dennett, 1994; 1991), but in this case such objections are brushed aside by referring to functionalism.

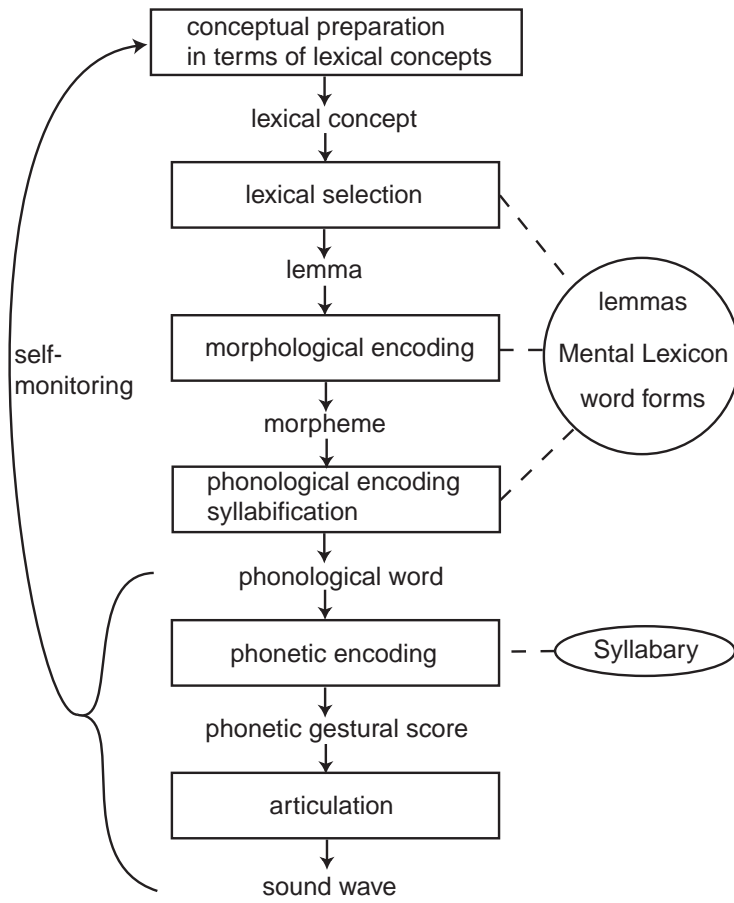


Figure 2: A model of language production. Adapted from Levelt, Roelofs, & Meyer (1999).

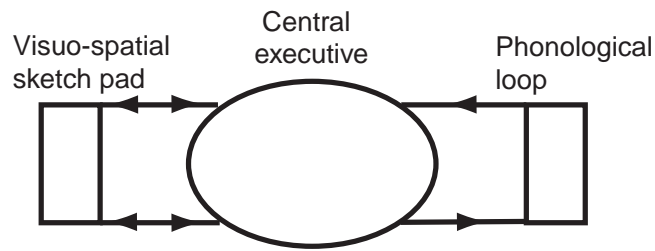


Figure 3: A model of working memory. Adapted from Baddeley (1990).

2.3 Experiments and effects

Other practical considerations that have been important in establishing current cognitive psychology concern not the theoretical models themselves but the experiments that are being done to develop and support the models. Experimental data are the foundation for cognitive psychology's theories, but they have the problem of providing only *indirect* information on the architecture and functioning of internal mechanisms. This is true even for brain imaging data, because these may, for instance, give rough indications as to which brain area receives more blood than others, but not say in what way that brain area is involved exactly. But the problem is even more acute for the kind of data on outward behavior that mainstream cognitive psychology typically works with: reaction time, strength, frequency, and accuracy of responses. Interestingly, these measures have in large part remained the same since the dawn of experimental psychology in the late 19th century. To cope with the indirectness of those data, experiments are usually set up such that a manipulation of a single controlled variable results in a changed response from the subjects, a so-called *effect*. The type and direction of the effect may then be used to derive conclusions about the underlying mechanisms that are involved and affected by the experimental manipulation. The subtractive procedure or the method of additive factors, developed by Donders (1862) and still widely used, is a particular instance of this idea. An experimental manipulation results in a change of responses, in this case longer reaction times, and this is interpreted as indicating an additional stage of processing or an additional involved functional component.

However, drawing straightforward conclusions from such effects can be difficult. Even if we have a large, clear-cut effect, its implications for the underlying mechanism are often unclear. Consider a case from language production research based on picture-word interference experiments (e.g. Glaser & Dünghoff, 1984; Levelt et al., 1999). These experiments are based on the Stroop task, an experimental paradigm that is more than 60 years old (Stroop, 1935). In picture-word interference experiments, the subject must name a picture (or classify it or respond in another way). Within the boundaries of the picture, a distracting word appears before, during, or after the appearance of the picture itself. A reliable finding is that if the so-called Stimulus Onset Asynchrony (SOA), the time between the presentation of the picture and the word, is roughly between -100 milliseconds and +100 milliseconds, and the distractor word is semantically related to the picture (such as the word "dog" is related to a picture of a cat), naming the picture takes significantly longer than in other conditions. This finding is displayed in figure 4.

What does this *semantic inhibition* effect mean for models of language production? In the interpretation of additive factors, it could mean that an additional stage of processing or functional module is involved. Alternatively, it could mean that semantically related words are stored closely together in memory, and "activation" of words "spreads out" to their neighborhood in a kind of blurring process, making it harder to distinguish semantically related words and pick out the correct one. Yet another interpretation is that the distractor word is processed faster than the picture and prepares the response component for saying the wrong word, which subsequently needs to be suppressed before the right word can be said.

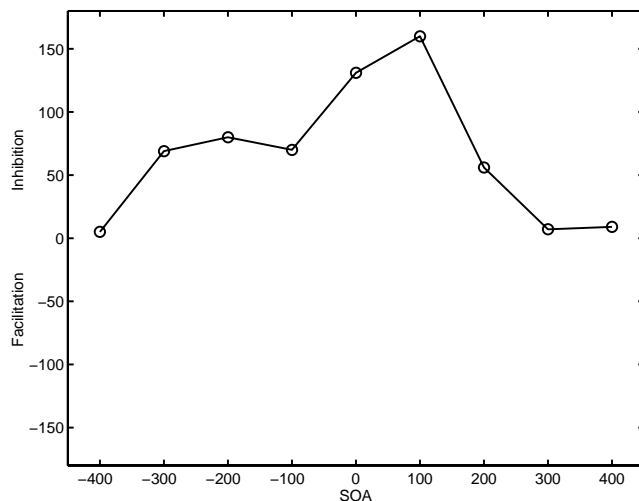


Figure 4: Semantic inhibition in the picture-word interference task where subjects have to name the picture, as a function of Stimulus Onset Asynchrony (SOA). Adapted from Glaser & Dungelhoff (1984).

The standard way to decide between the alternatives is to run new experiments. However, effects may disappear or even be reversed by slight changes to the experimental setup or by alternative manipulations. In the case of picture-word interference tasks, the effect is indeed reversed and becomes a *facilitation* effect when the pictures must be *categorized* rather than named (Levelt et al., 1999). In many cases research then tends to “zoom in” on details regarding those changes and manipulations, mapping out exactly when effects appear and how large they may become. In doing that, researchers are following the seemingly reasonable principle that one must first come to grips with the subtleties of the effect before one can move on. In addition, in new experiments a completely new effect may appear, much to the delight of its discoverers. The details of this new effect are then explored in more detail, etcetera.

In effect (no pun intended), the end result is that much of cognitive psychology is more concerned with effects by themselves than with the implications of effects for theoretical models. In a way, measurements and the effects derived from them take on a life of their own, with models being postponed to a later time “when we have more data”. The argument that “there is not yet enough data” is similarly used to justify the usual state of affairs that if a model is proposed, it is not specified beyond the very abstract level of a small number of boxes and arrows, as described above and illustrated by figures 1–3.

2.4 Mathematical and computational models

In those rare cases where a cognitive psychology model is further specified and a mathematical or computational model is implemented, an observation can be made that illustrates the strong focus of mainstream cognitive psychology on raw experimental data and effects. Such models typically replicate the raw experimental data and effects very precisely, but they do not aim to replicate

the general behavior that the research initially set out to investigate.

For example, the computational model of language production proposed by Levelt et al. (1999), which is an implementation of the box-arrow model depicted in figure 2, replicates the semantic inhibition effect very well, along with a number of other experimental effects, but it says little about language production in general: how the picture is “transformed” into processes dedicated to language, what those processes look like and how they are operated upon, how the complex structure of language and the large number of words are dealt with, and how processes concerned with language are in turn transformed into muscular movements corresponding to saying a word. In other words, one cannot present an actual picture of a cat to the computational model and obtain an auditory response “cat”—let alone have the language production model produce language as it is normally produced, in the form of comprehensible sentences. Reaction times and errors are produced, rather than actual language. Again this is justified by referring to the intuitively reasonable arguments that there is not enough data and that one must stay close to the data that one has got, before one can go on and “generalize to the unknown”. However, this leaves the peculiar situation that data that were initially gathered to say something about the mechanisms underlying a general capability of behavior, become the primary focus of the model at the expense of those general mechanisms. Such models explain the errors in and speed of certain behavior, without explaining the behavior itself.

The (implicit) idea behind this is that successive models of raw data and effects will encompass more and more data and, in the long run, converge to models which will indeed be able to produce actual behavior. After all, if we keep doing experiments, more and more data will become available, and if all conceivable data are eventually taken into account, this should include data on actual behavior. However, we have seen that more data often amounts to more details regarding the existence, disappearance and reversal of an effect under different experimental variations, and not necessarily to more data on actual behavior in natural circumstances. In practice, after many years of careful experimentation and corresponding theorizing, models have rarely, if ever, converged to models aiming more and more to replicate actual behavior in natural circumstances.

3 The adaptive behavior approach

Adaptive behavior research takes a very different approach to the study of the mechanisms underlying behavior. This approach is conveniently and effectively described by contrasting it with mainstream cognitive psychology on a number of issues.

1. General capabilities, as opposed to experimental effects

The previous section ended with a discussion of the kind of data that current models in cognitive psychology focus on, and ideas about how successive models will encompass more data. In general, this concerns the question of deciding where to start: which kind of data should be modeled first? Mainstream cognitive psychology opts for the kind of exact and quantitative but, as we saw,

very indirect data that can be and have been obtained in carefully controlled experiments.

The adaptive behavior approach can be understood as choosing another kind of “data” to focus on first. It attempts to replicate general capabilities of behavior that humans or animals exhibit. A system, or *agent*, is constructed that is capable of, for instance, locomotion behavior, or navigation behavior, or cooperation behavior, etc. These capabilities are not data in the regular sense of that which is gathered in controlled experiments, but they are data in the sense that they are non-trivial phenomena exhibited by humans and animals and they should be accounted for by a model. The idea is that a good model of the mechanisms underlying behavior should first and foremost account for what those mechanisms are *for*, what makes them interesting in the first place: the generation of successful behavior.

Viewed from this perspective, it is clear that the adaptive behavior approach is not merely a branch of engineering. It is a branch of cognitive science, in that it attempts to explain known facts about the behavior of organisms, facts concerning the existence, complexity, and limitations of that behavior. The difference between mainstream cognitive psychology and the adaptive behavior approach is not that the first explains actual data on organisms and the latter does not; the difference is in the type of data that is being explained.

Obviously, the type of data described as “general capabilities” cannot be measured as rigorously and quantitatively as the typical data used by cognitive psychology. In some cases, we may only be able to say that there is a qualitative fit between the constructed agent’s behavior and an organism’s behavior: the agent is capable of some behavior, but the degree to which it matches the capability of a particular organism is not entirely clear. In other cases, it may be possible to determine this fit more quantitatively, by taking measures of effectiveness and efficiency. However, in many cases an agent’s behavior is not even intended to model the behavior of a specific species, but rather a type of behavior exhibited by many species.

Mainstream cognitive psychology stays close to quantitative data gathered in neatly controlled experiments at the expense of the actual behavior itself. The adaptive behavior approach does the opposite: sacrificing some quantitative measures in favor of “qualitative data” that are claimed to have at least as high a priority.

2. Basic behavior, as opposed to high-level behavior

Building a system capable of behavior is difficult. At the moment, complex behavior is out of reach. For this reason, the adaptive behavior approach simplifies by focusing on very basic behavior first. This can be contrasted with mainstream cognitive psychology which can be characterized as simplifying by taking simple behavioral measures of mostly high-level cognition such as natural language, decision making, complex perception and attention, and intentional processes. It can similarly be contrasted to “traditional” artificial intelligence, which has focused on mimicking those higher-level cognitive processes. Traditional artificial intelligence has met with some success, but only when the domain to which the high-level cognition applies is very limited, such as chess or medical diagnosis. The idea to go back to basic behavior first (Keijzer, in press), which adaptive behavior research adopts, is based in part on the brittle-

ness of traditional artificial intelligence systems in real-world domains.

The adaptive behavior approach has the disadvantage that, initially, many of the most interesting types of behavior, especially the ones exhibited uniquely by humans and not by other animals, are not dealt with. On the other hand, this may not be such a bad idea given the history of research in biology, for instance. The very successful field of genetics has worked its way up from the relatively simple genomes of the fruit fly up to more and more complex genomes, with the human genome coming into the picture only recently. Thus, adaptive behavior research starts with the very basic behaviors exhibited by virtually all animals and slowly works its way up: locomotion, orientation, approach to desirable stimuli and movement away from danger, collision detection and collision avoidance, simple navigation, prey following and fleeing, cooperation, communication, etc.

This initial focus on basic behavior, as opposed to a focus on high-level cognition, does not imply a return to behaviorism or a denial of the existence and importance of high-level cognition. It is mainly a pragmatic choice, based on perceived limitations of our understanding of the mechanisms underlying even very basic behavior. In sharp contrast with behaviorism, adaptive behavior research is very much concerned with the internal mechanisms behind behavior, and it does not view behavior as simple stimulus-response relationships. In the long run, the adaptive behavior approach hopes to tackle more high-level cognitive behavior. In fact, there are already some efforts in this direction, especially with regard to planning (Nolfi & Floreano, 1998; Werner, 1994), communication using higher-level concepts (de Jong, 1999), and language (Steels, 1997).

3. Learning by constructing, as opposed to learning by measuring

By constructing an agent capable of a certain type of behavior, one can learn something about how biological systems, organisms, accomplish that behavior (Braitenberg, 1984). At the very least, one will learn about the *problems* that are involved in accomplishing that behavior, which are the same problems that the biological system must overcome. We can think of the evolved mechanisms underlying behavior as a *solution* to those problems. It seems reasonable that understanding that solution requires a sufficient understanding of the problems.

Interestingly, during construction one's intuitions about what will be and what will not be severe problems are often falsified. The history of the field of artificial intelligence is very illuminating in this regard. On the one hand, problems may turn out to be much more severe than was envisioned. In research on navigating robots it was initially thought that the process of transducing sensory information into a world model, as well as the process of transducing planned actions into motor commands was fairly trivial, and that all the hard work is done in sorting out a plan given the world model (Nilsson, 1984; Moravec, 1982). This turned out to be a gross underestimation of the difficulties of those "transduction" processes, and research on navigating robots stalled for many years as a result (Brooks, 1991).

On the other hand, problems that seem to be very serious may turn out to be pretty easy. It was previously thought that transforming English verbs from the present tense into the past tense requires a complicated system of rules, containing the rules and how they are applied, as well as the exceptions to which they do not apply. However, it was shown that this task can be performed relatively

successfully and even learned fairly easily as a straightforward stimulus-response association, using a neural network (a simplified model of nervous systems) of very basic architecture and only a few dozen neurons (Rumelhart & McClelland, 1986).

In general, the history of artificial intelligence shows that construction of an intelligent system is by no means a trivial enterprise. This suggests that it is not true that for any abstract theory on how to achieve some behavior, it will be easy or even possible to construct an implementation of that theory that works. The conclusion that the adaptive behavior approach draws from this observation is that the test of implementation is a much more serious one than is usually assumed by mainstream cognitive psychology. Implementation reveals genuine problems associated with accomplishing a particular type of behavior; and it reveals weaknesses (if any) of abstract theories in overcoming those problems.

Conversely, properties of a successful artificial system may suggest ideas, and even specific hypotheses, about how the biological system does the job. The engineered solution can be compared and contrasted with the biological solution. Having something to compare with may help in making sense of the data on biological mechanisms that are available from psychology and the neurosciences, data that are there in large quantities but that are often hard to interpret.

4. Implementationism, as opposed to functionalism

The degree to which constructed solutions, agents, tell us anything about nature's solutions, organisms, is a matter of some debate. After all, functionalism, which is widely adhered to in mainstream cognitive psychology, tells us that the same function can be accomplished in many ways. The way the engineered agent performs the function may be completely different from the way the organism performs it, and what we are interested in, in the end, is only the latter. How much does an airplane really teach us about how birds fly? Perhaps surprisingly, the answer to that question is: quite a bit. One example is that even though airplanes do not flap their wings and the wings are (fairly) rigid, the principle by which their wings accomplish their task is very similar to birds. Airplane wings and bird wings have similar, though not exactly the same, curvatures that cause lower air pressure above the wing than below during horizontal movement, creating lift. This understanding of bird wings almost completely depends on the development of airplanes and the corresponding understanding of aerodynamics.

In general, there are not as many possible ways to accomplish a particular capability effectively as seems to be implied by functionalism. The original idea from computation theory on which functionalism is based only says that the end result of a computation can be accomplished in multiple ways. Firstly, it says nothing about how long it will take. The archetypical computer, the Turing Machine, would be impractically slow for nearly all functions if it were implemented.

Secondly, it does not claim that *just any* idea about how to achieve a capability will work. In the enormous space of possibilities of different systems, or "design space", only a few regions of that space amount to systems that are actually capable of accomplishing anything interesting. And one can reduce the number of regions even more and zoom in to the region of interest by using self-imposed constraints on the "building blocks" of the artificial system.

Sure enough, a wheeled robot does not tell us very much about legged locomotion. But if the robot has legs and its artificial brain consists of artificial neurons, it becomes a different matter altogether. It then becomes more likely that one lands in more or less the same region of possibilities as the biological system, and the constructed solution has accordingly more similarities with the biological solution.

Thus, functionalism's suggestion that the particular implementation is irrelevant is rejected. The alternative doctrine, which we may call *implementationism*, states that the implementation is neither irrelevant nor trivial; in contrast, in order to understand a system's functioning one must understand the implementation and in principle be able to construct one.

There seems to be a paradox here. If a system's functioning is so closely tied to its implementation, how can we ever expect to learn something about an organism's functioning, with its biological implementation, by studying an artificial implementation? To solve this apparent paradox, it is necessary to specify the functionalism-implementationism dimension further. At one end of the spectrum, there is radical functionalism, which says that a system's functioning has nothing to do with the system's implementation. This would imply that any random stack of bricks could be intelligent. At the other end, there is radical implementationism, which says that a system's functioning is inseparably linked to all details of its implementation. Radical implementationism implies that each and every protein used in the cell is critical in achieving a neuron's function, and an artificial neuron that leaves out one such protein will fail. The argument that was made here for implementationism really argues for *mild* implementationism. Not *every* detail in an implementation is crucial for a system's functioning. It is possible to abstract away from some of the details (such as a particular protein) when constructing an artificial system and still be able to make meaningful comparisons with the biological system. In other words, within a single region of design space, there are still many systems whose details differ to some extent, but whose functioning is not affected significantly by those details. Of course, it is not known *a priori* which details are essential and which are not; this has to be found out.

The adaptive behavior approach follows mild implementationism and attempts to construct a successful implementation. In the process, the constructed agent's capabilities, its limitations, and the problems it overcomes can be understood; and the agent is compared to biological implementations. In most cases, adaptive behavior research takes inspiration from biology and attempts to use in its implementations the same type of building blocks as nature does, in order to facilitate meaningful comparisons. Sometimes it even becomes a matter of testing specific hypotheses about how the biological system works (e.g. Webb, 1994; Beer, 1990). As an extra advantage of the adaptive behavior approach, known biological features can be added to and removed from the agent at will, and the effects on the resultant behavior can be observed. In this way, it may become clear whether that particular feature is essential for the organism's functioning or not. These suggestions can subsequently be put to the test by the neuroscience and psychology communities (see Webb, 1994 and Beer, 1990 for examples). These are important ways in which adaptive behavior research feeds back to neuroscience and psychology and fruitful interactions may take place.

5. Detailed models, as opposed to abstract models

In accordance with implementationism, the adaptive behavior approach proposes implemented agents that actually demonstrate some capability of behavior, and which are therefore specified in large detail. This is in contrast with the very abstract models typically proposed by mainstream cognitive psychology. Those abstract models usually consist of a few boxes and arrows (see figures 1–3) and they contain hardly any details, sometimes to the point where one may wonder how much information they contain at all. A box labeled “short-term memory” contains only the information that humans are capable of remembering information presented to them a little while before; it says nothing about the underlying mechanisms.

Mainstream cognitive psychology often takes the position that details are bad, because the “principle of parsimony” says that models should be as simple as possible. The adaptive behavior approach takes the position that details are good or at least unavoidable, because actual behavior cannot be accomplished without them, neither in artificial systems nor in biological systems. As argued above in the context of “learning by constructing”, it is very hard to predict in advance which details are arbitrary and which are crucial. Only by implementing them will this become clear. In any case, a constructed agent that is successful at its behavior can be seen as an “existence proof” that proposed mechanisms deemed necessary for functioning actually do the job, and this requires the details to be filled in. Such an existence proof can never be given by an abstract box-arrow model.

A related objection is that details make it hard to see the bigger picture: the general principles that govern the mechanisms behind the behavior. If there are too many details, one may not be able to see “the wood for the trees”. First of all, if a detail turns out to be crucial in the implementation phase, it is apparently not part of the trees, but of the wood: without it, the bigger picture would not be complete, and without implementing it, the bigger picture could not have been obtained at all. On the other hand, one should not drastically go the other way toward radical implementationism, reasoning that all details may be important, and concluding that therefore one should always start by modeling individual molecules, or atoms, or elementary particles. One should constrain oneself to details for which it is reasonable to suppose that they may have relevance for the system’s functioning; the details by virtue of which the system may be doing what it is doing. Admittedly, finding this right level of detail is an art in itself. It is as easy to become too detailed as it is to become too abstract.

The agents that are constructed within the adaptive behavior approach are either actual robots or computer simulations. In both cases, the agent is investigated as a complete system of brain and body interacting with an environment, to make the behavioral task as realistic as possible. Both computer simulations and robots have advantages and disadvantages. An important advantage of computer simulations is that they are easy to work with and easy to modify. In addition, certain things are practically impossible to realize in robots but are possible to simulate in the computer, such as evolution, large populations of agents, muscles, a variety of environments, etc. On the other hand, it is very hard to simulate the full complexity of the real world. There is a danger of oversimplifying the problems that the agent is faced with. This is not a problem if one uses real robots. A real robot needs to confront the real world and show

it is capable of successful behavior—one may argue that only then there is a genuine existence proof that proposed mechanisms underlying behavior actually work. But research on robots is very difficult and can get lost in uninteresting technical problems. Usually, the types of behavior exhibited by robots are simpler than the ones in simulation, because it is so hard to achieve even the simple ones.

6. Perception to action loops, as opposed to functional modules

Mainstream cognitive psychology decomposes the animal and human information processing system into separate functions, and each function in turn into functional components. Each function is studied in isolation, with little attention for interactions with other functions or with the environment. The adaptive behavior approach uses another kind of decomposition, one that isolates all mechanisms that are involved in the particular, usually simple behavior of interest. Since the goal is to achieve complete behavior, all aspects from perception to action which are critical in achieving that particular type of behavior must be dealt with. This also includes interaction of the brain with the body and with the environment. Actions typically change the state of the body and the world and what is perceived next in an immediate feedback-like way. This is an important component of the behavioral task, and it has to be taken into account if successful behavior is to be generated. Mainstream cognitive psychology does not usually acknowledge the importance of this fact, and it treats body and environment as “passive”, independent receivers of actions and providers of sensations.

Dealing with all aspects from perception to action sounds like adding up all problems involved in different functions. If an individual function studied in isolation is so complex as to warrant a large, separate field of research, how can we expect to deal with all those functions at once? The trick is to study and solve only those problems of, for instance, perception or memory or motor control that are necessary to accomplish the single behavior of interest. Agreed, full-blown human perception is very complex and cannot be simply built into an agent. That is why the focus is on basic behavior and on the corresponding aspects of perception, memory, and motor control that by themselves are relatively simple when compared to the full complexity of human perception, memory, and motor control. But much can be and has been learned from how these simple aspects of a task constrain each other, how they interact with each other, and how interaction with the environment constrains the whole system (Brooks, 1989, 1991; Mataric, 1991).

In fact, as we shall see in the next section, one of the most important insights resulting from adaptive behavior research is that successful agents often resist a clear-cut decomposition into functions and functional modules. Memory may be distributed across the whole agent rather than located in a separate component (Rumelhart & McClelland, 1986; Dennett, 1994). Perception and action may be intricately linked to the point where they are no longer usefully thought of as two separate components (Brooks, 1989, 1991; Beer, 1990; Beer & Gallagher, 1992). Agents may behave “as if” they pay selective attention to certain sensory information, without having an explicit mechanism or component for attention (Werner, 1994). For these agents, the intuitively natural decomposition into functions and functional components is not the most fruitful one. This in turn

suggests that it is not necessarily the most fruitful one when one thinks about biological systems.

But what about those findings from the neural sciences that seem to support the idea of decomposition into functions and functional modules? It should be noted that those data do not directly implicate such clear-cut functional isolation, but rather show something much weaker: there is some level of specialization, *not all* brain areas are directly involved in a particular task, but only parts of the brain. The same is usually true of constructed agents: there is some level of specialization, some parts of the artificial brain are involved in some tasks, and other parts in other tasks.

Within biological brain regions that are involved, it is unclear how to further assign subtasks to brain areas. But what is suggestive is that there are typically many connections back and forth to brain areas; a finding that argues against strict functional isolation. Furthermore, one should be careful with deriving conclusions from failing systems. If a radio starts to make a howling sound if one particular transistor is broken, this does not mean that this transistor is the “howl inhibitor” (Arbib, 1989). In general, the phenomenon that only a specific capability fails if one of the physical building block is broken, is a characteristic of many systems, and not only systems with clear-cut, isolated functional modules.

7. Decentralized control, as opposed to centralized control

Mainstream cognitive psychology usually proposes as one of the functional modules a central workspace where information is temporarily stored for processing and organization by a central controller. There is a danger of attributing all capabilities that are not yet understood to this central controller, which is not specified in more detail. Sometimes this central controller can rightly be called a “homunculus”, a little man in the head, which takes over all the hard work that the overall system is supposed to do. Such a model does not explain the capabilities of the system, but only “pushes back” the problems deeper into the system, into an unspecified functional component named “central controller” (Dennett, 1991).

The adaptive behavior approach cannot, in principle, resort to this strategy because it forces itself to replicate the capabilities. If a central controller were to be proposed, it would have to be implemented for behavior to be accomplished. In the process, its mechanisms have to be made explicit, and its weaknesses, if any, are revealed.

Almost all of traditional artificial intelligence has used central controllers and workspaces in its systems. This has met with very limited success, and it is therefore an example of a specific cognitive psychology idea about underlying mechanisms that seems to be falsified when put to the test of implementation. Among the problems is the creation of a serious bottleneck if everything has to pass through the central workspace, slowing down performance tremendously. There is also the problem of information access (Dennett, 1994; Lenat & Guha, 1990). One can have a long-term memory storing huge amounts of information, but how does one get the relevant piece of information into working memory in time? Furthermore, for systems that interact with the world, it has proven to be very hard to maintain and update the central model of the world that the central processor is operating upon—this is known as the frame problem.

Using sensory information to decide what has to be changed in the central world model, as well as predicting what will change in the world model if some action is executed, is notoriously difficult (e.g. Brooks, 1991; Moravec, 1982; Nilsson, 1984; Dennett, 1991; Krotkov & Simmons, 1996).

For these reasons, the adaptive behavior approach typically attempts to accomplish behavior without using a central controller or workspace. Control is decentralized, distributed among local controllers operating in parallel. Each local controller performs a simple task, such as moving a single leg when its own sensors say so, or activating another local unit when its sensors detect a specific feature in the world. Through the interaction of these local controllers between themselves and with the environment, the behavior as a whole “emerges”, without a need for a central guiding or monitoring mechanism. This is often called *self-organization*, because there is no central system actively organizing the behavior.

8. Distributed, continuous internal state, as opposed to symbolic representation

Based on the computer analogy, the content of the central workspace (as well as long-term memory) is usually theorized to amount to so-called symbolic representations. These are language-like structures consisting of arrays of symbols, manipulated by logical operations. Symbols are the atoms of knowledge that “stand for” something in the outside or inside world, such as “Mary” or “love”. An important, powerful property of symbolic representations is their combinatorial structure. Symbols can be combined into large symbol structures in many different ways, representing many different meanings. Certain operations on these representations may depend on the combination of the symbols, rather than the meaning of the individual symbols themselves. This is called structure-sensitive processing, and it allows a single type of operation to generalize to many different contexts.

Just as there are no indications that there is a central workspace in the brain, it is not obvious at all where and how the symbolic representations are encoded. That is not to say that there are no symbolic representations in the brain. However, the structure of and processes in the brain suggest other types of encoding that are used next to, or perhaps even instead of symbolic encodings. These other types of encodings are investigated in depth in the field of artificial neural networks or connectionism (see Rumelhart & McClelland, 1986). Artificial neural networks are simplified models of biological neurons and biological neuron interactions. They are inherently based on decentralized control. Furthermore, no clear distinction can be made between the controlling part and the information used by the controller, which is very different from standard computers and models in mainstream cognitive psychology. As it turns out, the type of continuous-valued, distributed internal states encoded by artificial neural networks, and therefore—it is hypothesized—by biological nervous systems, affords a type of processing well suited for many parts of intelligent behavior.

Symbolic representations are best suited for logical, all-or-nothing types of reasoning and applications of strict rules. Much of intelligent behavior is handled better and perceived more fruitfully as recognition and classification of patterns, completion of partial information, association between related pieces

of information, and decisions based on incomplete information and on the satisfaction of multiple, “soft” constraints. Those types of tasks are done more naturally and effectively in neural networks than in symbol systems. This is particularly acute for a system interacting with the environment, which is an important focus of adaptive behavior research. As described above, the transduction process from sensory information to a central world model, which is supposed to consist of symbolic representations, is very hard. The type of continuous, distributed internal states of neural networks lends itself much better to connections and interactions with sensory and motor apparatus than discrete, symbolic representations. As for structure-sensitive processing, it was shown that this is not limited to symbolic representations but can also be done in neural networks (e.g. Chalmers, 1990; Elman, 1990).

For these reasons, a lot of adaptive behavior research uses artificial neural networks, or some other type of system based on decentralized control and distributed internal states (e.g. CMAC or classifier systems). However, in contrast with connectionism (Rumelhart & McClelland, 1986), no absolute commitment is made to distributed internal states. If it turns out that for some types of behavior (e.g. high-level cognition) symbolic representations are necessary or very useful, the adaptive behavior approach, with its focus on making systems work, will use them. There are more differences with connectionism, so it is certainly a mistake to simply equate the two approaches. In a way, connectionism is situated in between mainstream cognitive psychology and adaptive behavior research. Like the adaptive behavior approach, it emphasizes implementations and decentralized control. But unlike the adaptive behavior approach, connectionist models are often intended as straightforward implementations of the functional modules of mainstream psychological theories, and they are often used to directly fit typical experimental psychology data on reaction times and errors. The adaptive behavior approach emphasizes much more than connectionism the importance of building complete agents interacting with realistic environments, exhibiting general capabilities of behavior.

9. Bottom-up engineering principles, as opposed to top-down engineering principles

The decomposition into functions and functional modules that mainstream cognitive psychology assumes reflects—and is probably inspired by—standard engineering principles. In fact, it is basically how traditional artificial intelligence constructs an artificially intelligent system.

First, it is determined what the system is supposed to do, what its overall task is. Next, this task is divided into subtasks which are handled by dedicated components. Each component’s subtask is relatively independent and well-defined, such that the component can be individually built and tested. To avoid the notorious problem of unwanted side effects and complications caused by interactions between components, each component’s functioning is isolated as much as possible from other components. Finally, all components are put together. This can be called top-down engineering, because one starts with the abstract idea of the overall task, working one’s way down to more and more concrete subtasks and finally physical realization of the components and combination into a complete system.

However, this is not the way nature constructs systems. Nature does not

start with an abstract idea about what the final system should do. It does not neatly figure out subtasks and assign these to functionally isolated subsystems. It does not care about whether or not any clear-cut decomposition into sub-functions is possible at all, or whether the end result is easy to comprehend for scientists. And it does not build and test the subsystems individually before they are recombined. In contrast, nature blindly tries out all kinds of systems without any foresight on what the system should do or how it should do it. It builds on systems that were successful before, varies them randomly and selects the lucky ones that happened to work one way or another. It selects a system as a whole and does not develop subsystems in isolation, opportunistically allowing side effects and complex interactions between subsystems as well as multiple functions within a subsystem if they happen to be beneficial.

To contrast this with top-down engineering, it may be called bottom-up engineering (e.g. Dennett, 1994): starting with building blocks from previous generations, those building blocks are varied and more or less randomly combined into a new system, allowing strong interdependencies and interactions between them, without a preconceived and neatly worked out plan on the overall design. But the system is selected on the basis of success in its environment, such that unsuccessful systems (most systems, in fact) are filtered out and we are left with successful ones. Only with hindsight, then, one can say that a successful system is “designed to perform a particular task”.

It seems plausible that top-down engineering often leads to different types of systems than bottom-up engineering. Thinking again in terms of the metaphorical space of possibilities of systems, design space, top-down engineering is constrained to certain regions of design space; in particular, regions where systems are easily decomposable into functional modules (these may be very good systems: airplanes are an example). Bottom-up engineering does not have those constraints. Of course, it has other constraints, such as the constraint that a design is always heavily based on a previous design. Because of these differences in constraints, bottom-up engineering may end up in very different regions of design space. Those regions may yield systems from the easily imaginable to the bizarre—the only criterion used by bottom-up engineering is success of the system.

In order to learn more about the systems designed by nature and to have access to the same regions in design space as nature has, a lot of adaptive behavior research attempts to mimic nature’s bottom-up engineering processes. Agents are developed over many generations using a simulated evolution process, employing so-called evolutionary algorithms; or they are developed using learning, nature’s way of developing an organism during its lifetime. As it turns out, in many cases the artificial systems developed in this way are indeed very different from what was expected given the standard top-down view in traditional artificial intelligence and cognitive psychology (e.g. Beer, 1995; Beer & Gallagher, 1992; Nolfi & Floreano, 1998). This challenges the (usually implicit) assumption in cognitive psychology that the top-down approach is the most logical and fruitful one, when it comes to understanding and reconstructing biological intelligence.

10. A posteriori analysis, as opposed to a priori analysis

The overall emphasis on constructing working systems, the fact that bottom-

up engineering does not work from an abstract theory, and the explicit statement of implementationism, all seem to suggest that the adaptive behavior approach is not very interested in abstract theories. The main concern seems to be with building an agent that is specified in large detail and that is successful at its particular task, and not so much with finding more generalized, high-level theories of behavior and of the mechanisms behind behavior.

First of all, it is true that the adaptive behavior approach has a somewhat different attitude towards abstract theories. Adaptive behavior research looks at organisms from an engineering perspective. This leads to a general view of cognitive science as an enterprise that is more like reverse engineering than like physics. A reverse engineer attempts to understand complex machinery made by someone else, with the goal of being able to build a similar device herself (Dennett, 1994). She does not come up with a single “theory of video cassette recorders” or a single “theory of cars”, in the same sense as there is a theory of elementary particles. What she is looking for is insights in the workings of the machine. This requires descriptions of many parts and their interactions, and a number of general principles; but not a single theory. In the same way, it may be an idle hope to find a relatively simple “theory of behavior” for complex, bottom-up engineered systems such as organisms.

Having said that, the goal is still to find those general principles of the mechanisms underlying behavior, and taken together, the general principles can be said to constitute the abstract theory. It is important to note that constructing detailed implementations does not preclude an adaptive behavior researcher from thinking about the abstract theory, both in the process of constructing and in the analysis of the completed agent. In the end, abstract theories (general principles) are what the whole enterprise is about; not single, detailed agents, built for one specific environment. The adaptive behavior researcher is, in fact, in one of the *best* positions to understand the general principles, because she has constructed a complete agent herself and knows, like no one else, about the problems that have been overcome and how the solution works.

The doctrine called implementationism does not say that one should not be concerned with the abstract theory; it says that one should build and look at implementations if one wants to find the abstract theory. Compared to functionalist psychology, the order of doing things can be said to be reversed. Functionalist psychology starts out, *a priori*, with an abstract theory, based on an analysis of the demands of the task and on preconceived ideas on how to deal with the demands, and expects that an implementation can easily be found which realizes this theory in a physical machine. Adaptive behavior research, in contrast, focuses first and foremost on building a successful machine and only then on analyzing it, *a posteriori*, so as to arrive at the abstract theory. If one employs bottom-up engineering to construct the agent, this is even the only possible methodology.

The idea is that it does not make sense to establish the abstract theory until one has extensively investigated the feasibility of different ideas on how to accomplish a certain capability. In other words, implementation is part of the process of theorizing right from the start. Once there are successful implementations, it becomes possible to say which ideas worked and which failed and, with hindsight, it may become easier to see *why* certain ideas worked and others failed. This then yields the more generalized, abstract theories.

As one example, implementations of multilayer feedforward neural networks

in detailed computer simulations have yielded the general theoretical insight that a lot of complex rule-like behavior can be accomplished with, and understood as a *nonlinear mapping of vectors* (e.g. Rumelhart & McClelland, 1986; Sutton & Barto, 1998). This is the abstract theory behind multilayer feedforward networks, but it was not and probably could not have been conceived a priori by functionalist psychology. The whole concept of “nonlinear mappings of vectors” emerged from and depended on investigations of detailed implementations of neural networks.

An example that is more typical of adaptive behavior research is Beer’s (1995) analysis of an evolved locomotion controller, described in more detail in the next section. The high-level abstract theory describing the evolved controller is stated in terms of dynamical systems, and again it is very different from anything that was or even could have been conceived a priori.

As a final example of adaptive behavior research on finding more generalized theoretical insights, a number of recent studies focus on classification of agents and environments (e.g. Wilson, 1991; Bakker & de Jong, 2000). Such classification methods may afford better judgements of the difficulty of different environments and the complexity of different agents. In this way, they may afford more meaningful comparisons between different studies, which at this time often consist of single examples of agents in specific environments.

4 Examples of adaptive behavior research

In this section a number of examples of adaptive behavior research are presented that are intended to illustrate the issues described in the previous section. At the same time, they are examples of some of the theoretical insights that have been gained using the adaptive behavior approach. A number of these insights directly oppose ideas in mainstream cognitive psychology. This suggests that those classical concepts should not be taken for granted, but may be in need of revision.

4.1 Locomotion

Locomotion, in particular legged locomotion, was one of the first types of behavior investigated using the adaptive behavior approach. In part, this resulted from dissatisfaction about traditional artificial intelligence efforts on locomotion. Those efforts had resulted in large robots using a central controller which carefully planned and executed each single step or change of position before anything else could happen, yielding very slow and unnatural locomotion (e.g. Krotkov & Simmons, 1996).

Taking inspiration from biology, researchers, most notably Brooks (1989, 1991), decided to use simple decentralized controllers in small, insect-like legged robots. Each leg is controlled in a local, reflexive way, aided by individual timers: if the leg is down and forward, swing backward; if it is down and backward, lift the leg and swing forward, etc. Coordinating these six moving legs such that successful locomotion is accomplished seems like a complex problem. In contrast, this can be achieved by simply letting the robot move about in the world and exploiting the sensors on the legs, using simple inhibition between the legs. If a leg is swinging forward, the other legs that stand on the ground are told

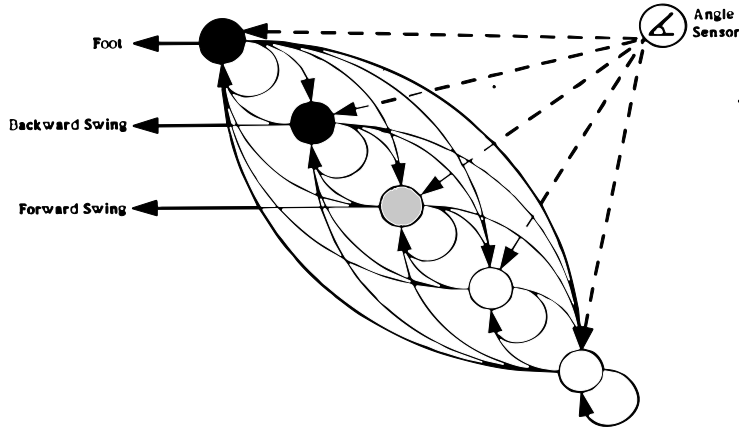


Figure 5: Part of the neural locomotion controller connected to a single leg. From Beer & Gallagher (1992).

to swing backward a little, etc. When put into action, the robot quickly settles into an efficient, lifelike gait of locomotion and is even able to negotiate somewhat rough terrain. It can easily be supplemented with similar reflex-like mechanisms that allow it to detect and deal with collisions. This work was an important “step” because it showed how complex coordinated movements can arise from parallel, distributed control, without a central coordinating mechanism and without using explicit central representations of the task and the environment. Thus, it is one of the first successful examples of exploiting self-organization to achieve complex behavior.

Following up on this work, Beer and co-workers (Beer, 1990; Beer, Chiel, Quinn, Espenschied, & Larsson, 1992) used a neural network as locomotion controller for their simulated insects and real robots. This artificial neural network was a simplified model of the nervous system of the cockroach. Previously, locomotion of this type of animals was thought to be controlled by a central locomotion system carrying out a fixed motor program and sending commands to the legs at precisely timed moments (see Marder & Calabrese, 1996; Simmons & Young, 1999); these models were akin to typical cognitive psychology models. In contrast, the neural network model uses highly distributed control and, again, inhibition between legs, but this time only local inhibition between neighboring legs.

The cockroach is known to have different gaits, which it uses at different speeds. To explain this, classical models have to assume that the central locomotion system contains different motor programs, one for each gait. In the neural network model, however, these different gaits arise spontaneously when the speed is varied—another instance of self-organization. If one is to understand how this works, one cannot fruitfully think in terms of the classical concepts of motor programs, functional components, and the like, but one has to view the neural network in interaction with body and environment as a single dynamical system in which different periodic attractors are stable at different speeds.

Later work by Beer and colleagues (Beer, 1995; Beer & Gallagher, 1992) included the development of neural network locomotion controllers using evolu-

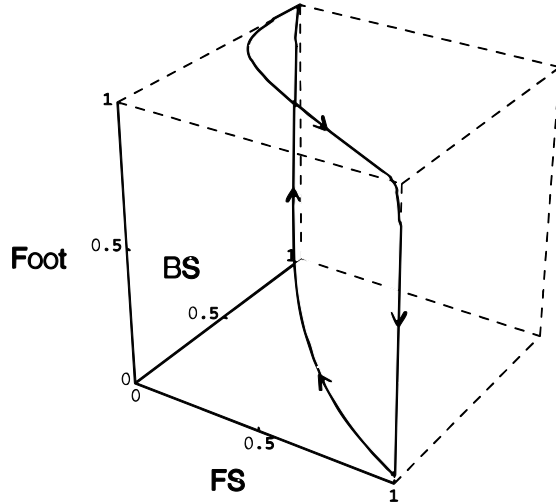


Figure 6: Phase space plot of the limit cycle of the leg controller depicted in figure 5. The output of the Foot, Backward Swing (BS), and Forward Swing (FS) motor neurons are plotted. From Beer (1995).

tionary algorithms, thus mimicking nature’s bottom-up engineering style. This provided additional confirmation for the idea that, at least for organisms exhibiting this type of behavior, dynamical systems notions may provide better explanations than cognitive psychology’s classical notions. A part of the overall neural network that is connected to a single leg is shown in figure 5 (Beer & Gallagher, 1992). It cannot be meaningfully decomposed into different functional modules, nor is it functionally isolated from other parts of the network; and there is no central control. Rather, each neuron continuously affects and is affected by the other neurons as well as the body and the environment, and the overall capability emerges from the interactions. To illustrate the difference in the type of explanations of the mechanisms underlying behavior, figure 6 depicts a so-called phase plot of the limit cycle of this single leg’s local controller interacting with the environment (Beer, 1995). What is important here is that this new, dynamical systems type of abstract theory (and the diagram illustrating the theory) is very different from mainstream cognitive psychology theories, and this insight depended on the implementation of agents.

4.2 Navigation

Locomotion can be used by an organism just to wander around at random, until food is encountered. Locomotion can be employed much more efficiently, however, if the animal knows where it is and where to go to achieve its goals. This is called navigation, and all but the simplest animals use it. The animal exploits cues in the environment available through its sensors, or a memory of what is has done and experienced since it left “the nest”, or a combination of these options, to decide where to go next.

It was navigation which prompted Tolman (1948) to suggest that pure stimulus-response behavior was insufficient to account for certain behavior, in

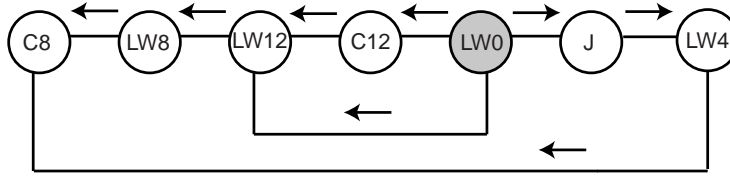


Figure 7: Cognitive map learned by a robot in a cluttered office environment. LW8 means Left Wall heading south, for instance. Arrows denote the spreading of activation from the goal (gray node). Adapted from Mataric (1991).

one of the studies leading up to the cognitive revolution. Tolman argued that rats trained in a maze learn a cognitive map, a map encoded in the brain, representing their environment and their current position. Early efforts to implement such cognitive maps by traditional artificial intelligence were not very successful. A robot developed at Stanford failed dramatically when the angle of the sun changed over time, changing the shadows in an otherwise static environment (Moravec, 1982; Brooks, 1991). Shakey the robot at SRI had some success at navigation, but it stood still for long periods of time to “think” (in the meantime shaking a bit, hence its name), and it operated in a highly simplified, small world of a few rooms and brightly colored boxes (Nilsson, 1984; Dennett, 1991; Brooks, 1991). These examples illustrate the difficulty of using sensory information and planned actions to maintain a central world model, in this case a central cognitive map module, even when the layout of the cognitive map is carefully programmed in beforehand.

More recent attempts by adaptive behavior researchers have taken another route. Rather than using a central functional module containing the cognitive map, separate from sensory and motor apparatus and operated upon by a central controller, they use systems based on distributed control and close interaction with the environment. In addition, they use learning, one of nature’s bottom-up engineering methods, rather than a preprogrammed cognitive map.

One example is Mataric (1991). A robot was constructed capable of navigating successfully in a cluttered office environment. It applies the perception-action loop philosophy described earlier. One perception-action loop (or “layer”) is used to avoid obstacles and follow walls. Building upon the first loop, another loop detects and registers landmarks in combination with its own concurrent movements. A third loop uses that information to develop a kind of cognitive map (see figure 7).

However, this cognitive map is different from traditional conceptions of the cognitive map. It consists of a network of nodes, each representing a registered landmark and corresponding movement. The current location of the robot is represented by one of the nodes being active. Activation spreads to other nodes, thus generating “expectations” about what will be perceived when the robot performs the corresponding movement. A goal location can also become active,

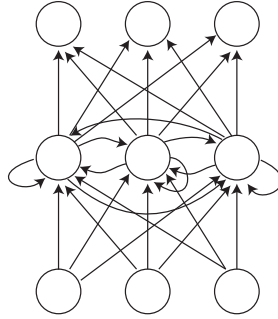


Figure 8: Simple Recurrent Network. The internal feedback connections are in the intermediate, hidden layer.

and activation will similarly spread out to neighboring nodes. This spreading of activation depends on the physical distance between landmarks. Consequently, locally at each landmark suggestions can be made as to which direction to go to reach the goal most rapidly. Overall, this results in the robot choosing the globally shortest path to the goal. Determination of the current position, map building, and action selection are not separated into distinct functional components, but they are all combined in this single map. There is no central planning mechanism figuring out the optimal path to the goal; the navigation behavior emerges as the result of interacting local units.

This cognitive map investigated by Mataric (1991) still looks considerably like a map as we normally construe it. Distinct places in the world are represented by distinct nodes, and a change in position of the robot in the world is represented by moving from one node to a connected node. However, it is possible to achieve similar capabilities without the use of anything remotely similar to such a map. There is a large body of literature on learning to navigate using reinforcement learning (see Sutton & Barto, 1998). The common principle is to let an agent explore an environment in which rewards are present in certain locations. On the basis of these (scarce) rewards, and without further instruction, the agent must learn to find its way to the goal from different positions in the environment. One of the more difficult, as well as realistic, variations of this task is the case where sensory information by itself is not sufficiently informative to allow action selection to be based on that alone. In other words, the sensory information is ambiguous. For instance, one T-junction in a maze looks exactly like another T-junction; but in the first case the best action is to go left, and in the second case the best action is to go right. Such tasks are called non-Markovian tasks, and the agent must use some kind of variable internal state to resolve the ambiguity of the sensory information. One option is to constrain this internal state to an explicit cognitive map. But the option that will be discussed here is to take the bottom-up engineering approach even further and have the system develop its own internal state based on the rewards it obtains.

To this end, a Simple Recurrent Network (Elman, 1990), a neural network with internal feedback connections, can be used (see figure 8). The feedback connections provide the variable internal state, or a kind of short-term memory.

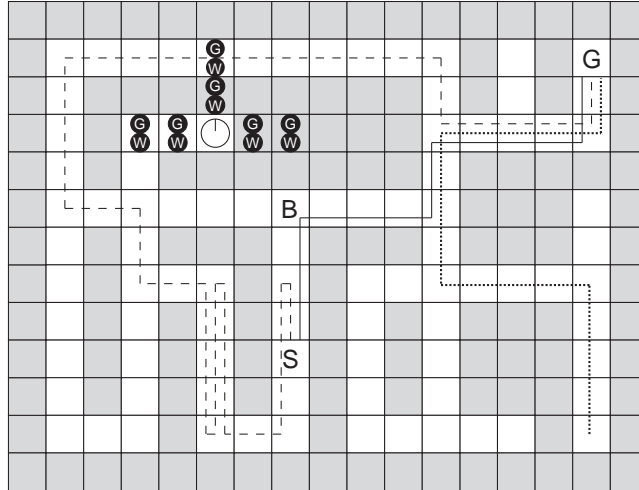


Figure 9: The maze. The agent is the white circle. It is oriented to the north, and is depicted together with its sensors. The solid line indicates the path taken by the agent to the goal (G) from the normal starting point (S). The dotted line indicates an example of a path taken when the agent is moved to another, remote starting position. The dashed line indicates the path taken when the position denoted by B is blocked.

Long-term memory is encoded in the weights of the connections between the neurons. Both are distributed and continuous-valued. It is apparent that long-term memory and short-term memory are not functionally isolated from each other, as is presumed by mainstream cognitive psychology, but are different aspects of the same network.

This type of network was used as controller for a simulated agent learning to navigate in mazes (see also Bakker & van der Voort van der Kleij, 2000). The agent’s perception is limited, such that it literally has the described problem of ambiguous T-junctions (a non-Markovian task). The agent successfully develops internal states that allow it to navigate from the starting position to the goal position using the shortest path (see figure 9).

It is interesting to go back once more to Tolman’s (1948) work on navigating rats. He similarly used mazes with few perceptual cues, combined with reinforcement learning. His strongest arguments for the existence of a cognitive map come from experiments where the rats were moved to another starting position, and from experiments where the optimal path which the rat had learned was suddenly blocked. In both cases, almost immediately the rats tended to choose the path that was best given the new situation. They had not been explicitly rewarded to do that, thus showing that they had not simply memorized a sequence of actions but could generalize over the maze in a “smart” way.

The same experiments were applied to the simulated agent in the maze. Figure 9 shows that here too the agent chooses the optimal or near-optimal path given the new situation. According to Tolman’s criteria, the agent must contain a cognitive map. But this capability is accomplished using an architecture in which nothing similar to a “map” can be discerned. One useful way to

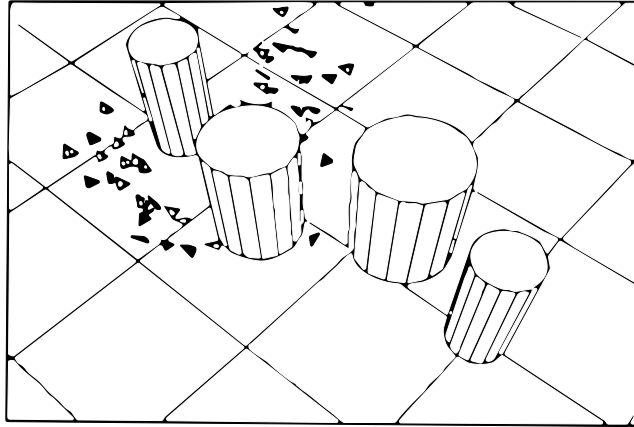


Figure 10: Flocking boids. The flock has split up to avoid the obstacles in the flight path, and will reassemble after the obstacles. Reprinted with permission from Craig W. Reynolds.

describe how the agent deals with these situations is to consider the agent and environment in combination, and to say that after some initial “confusion”, the agent’s distributed, continuous-valued internal state becomes once again “entrained” with the environment.

4.3 Collective behavior

As a final example of a line of research in the adaptive behavior community, let us have a look at work that is concerned not just with one agent, but with multiple agents interacting with each other. Interestingly, complex collective behavior can arise from the interaction of simple agents. This can be viewed as the previously discussed principle of self-organization in systems with decentralized control, but applied on a larger scale.

As early as 1950, Grey Walter experimented with a pair of very simple robots interacting with each other. He noted that the resultant behavior from the robots could become surprisingly complex: “Crude though they are, they give an eerie impression of purposefulness, independence, and spontaneity” (Walter, 1950). Reynolds (1987) showed how the adaptive and well-coordinated behavior exhibited by flocking birds could be replicated by agents, “boids”, that follow very simple rules based on the distance to their immediate neighbors (see figure 10). The resultant behavior is very smooth, adaptive, and life-like.

Work by Steels and co-workers demonstrates cooperation in learning agents. In one study (Steels, 1995), there is potentially mutual benefit for robots to cooperate, because they are hindered by parasites in obtaining energy. Even though in principle the robots compete for the same energy source, and cooperation behavior is neither programmed in beforehand nor suggested by explicit instruction, cooperation emerges spontaneously. Even stronger forms of altruism can arise. Brinkers & den Dulk (1999) investigated groups of evolving agents changing over generations because of simulated evolution. The experiment was set up such that some members of a group of agents will do much better if

some other members sacrifice their “lives”. Even though agents are selected on an individual basis, as is the case in natural selection, and therefore selfish behavior is the default expectation, such radically altruistic agents evolve and the altruistic behavior is evolutionarily stable.

In many animals, cooperation is accomplished with the help of communication. In a simulation study by de Jong (1999), agents can potentially benefit from warnings by other agents that a certain type of predator is present. Each type of predator makes a specific location unsafe, e.g. the presence of a snake makes it unsafe to stay on the ground. The warning signals are learned in a bottom-up way. The agents start out using different, random warning signals for situations where different predators are present. Eventually they converge to common, reliable signals, to the point where they learn to rely on the other agents’ warnings and avoid an unsafe location even if their own perception indicates that this location is safe.

These were all examples of cooperative behavior. The darker side of nature can be replicated as well. Nolfi & Floreano (1998) show, both in computer simulations and in robots, how co-evolution of predators and preys may yield relatively quick bottom-up development of complex behavioral strategies that oppose each other. The evolution of a slightly smarter strategy on one side pushes the other side to evolve a strategy that can cope with that, which in turn pushes the first side to develop an even smarter strategy, etc. This can be described as an “arms race”. The predator species developed intricate pursuit and ambush strategies, while the prey species developed effective escape and avoidance strategies. These strategies, which betray sophisticated anticipation capabilities, are encoded as distributed information in artificial neural networks. Once again, the bottom-up engineering approach leads to a system without distinct or central functional components for storing the strategies, for planning the behavior according to the strategy, or for explicit anticipation representation.

5 Discussion

The arguments and examples presented in this paper showed that the field of adaptive behavior does not provide mere implementations of the abstract theories of mainstream cognitive psychology. In contrast, in many cases successful agents are controlled by mechanisms that are very different in architecture and functioning from what was or would be expected given the abstract theories. Understanding those mechanisms often required completely new types of explanation, rather than explanations derived from those abstract theories. On the other hand, direct attempts of straightforward implementation of ideas from the abstract theories turned out to be problematic or unnecessary.

What does this mean? It was argued that implementation is neither trivial nor irrelevant for cognitive psychology. In contrast, if the implementation tells a different story than the abstract theory, suggesting that the abstract theory cannot be implemented or that some behavior is best achieved using other mechanisms than was anticipated, this has implications for the abstract theory. In this case, implementation of agents suggests that certain standard ideas in cognitive psychology, such as centralized control, cognitive maps, separation between working memory and long-term memory, and in general decomposition into isolated functional components, are in need of revision or, at the very least,

should not be taken for granted as much as they are now.

Many of these standard ideas in cognitive psychology were arrived at and supported using the standard methodology of collecting data in controlled experiments. The fact that now some of those standard ideas are rejected prompts rethinking of that methodology. Experiments are important, but perhaps not for the same purpose as much of mainstream cognitive psychology has it. In some cases, there is an overly strong preoccupation with finding effects, without worrying what the effects mean. It seems to me that certain types of data—reaction time, strength, frequency, and accuracy of responses—are measures that are usually too indirect to straightforwardly derive from them reliable conclusions about the underlying mechanisms, and collecting additional indirect data does not help (they may have much practical significance, though, if they can be applied to industrial, clinical, or educational settings).

The most interesting data from experiments, in terms of relevance for models of underlying mechanisms, may be data that indicate a beforehand unknown capability that humans or animals have, that map out what the capabilities are, or that disprove a previously assumed capability. These data have direct implications for models of the mechanisms underlying behavior, in that they say what those models should be able to do and what they need not do. In that way, cognitive psychology heavily constrains adaptive behavior research. The exact speed and accuracy of the behavior are constraints that become important only later on, once we have successful models that account for the behavior itself. Adaptive behavior research, in turn, constrains cognitive psychology in the kinds of systems that are proposed as models. Certain kinds of systems work very well and other kinds of systems cannot be made to work at all. The test of implementation is crucial and should be brought in to theorizing as early as possible. In addition, from successful agents the adaptive behavior approach can derive hypotheses about the biological mechanisms, which can subsequently be tested by experimentalists.

It was shown in adaptive behavior research that multiple agents may benefit from cooperating with each other. The same may be true for the two approaches discussed here. A stronger relationship between cognitive psychology and adaptive behavior research may be mutually beneficial.

6 Acknowledgments

I am grateful to Paul den Dulk, Martijn Brinkers, Edwin de Jong, Michiel de Jong, Arjan de Boer, Fred Keijzer, Antonino Raffone, Patrick Hudson, Bernhard Hommel, Dagmar van der Neut, and Robert Griffioen for fruitful discussions and suggestions.

7 References

- Arbib, M.A. (1989). *The metaphorical brain 2: Neural networks and beyond*. New York: John Wiley & Sons.
- Ashcraft, M.H. (1998). *Fundamentals of cognition*. New York: Addison-Wesley.
- Baddeley, A. (1990). *Human Memory*. London: Lawrence Erlbaum Associates.

- Bakker, B., & de Jong, M. (2000). The epsilon state count. In J.-A. Meyer, A. Berthoz, D. Floreano, H. Roitblat, & S.W. Wilson (Eds.), *From Animals to Animals 6: Proceedings of The Sixth International Conference on Simulation of Adaptive Behavior*, 51–60, Cambridge, MA: MIT Press.
- Bakker, B., & van der Voort van der Kleij, G. (2000). Trading off perception with internal state: Reinforcement learning and analysis of Q-Elman networks in a Markovian task. In S.-I. Amari, C.L. Giles, M. Gori, & V. Piuri (Eds.), *Proceedings of the International Joint Conference on Neural Networks 2000, Vol. III*, 213–218.
- Beer, R.D. (1990). *Intelligence as adaptive behavior: An experiment in computational neuroethology*. San Diego, CA: Academic Press.
- Beer, R.D. (1995). Computational and dynamical languages for autonomous agents. In: R. Port & T. van Gelder (Eds.), *Mind as Motion*, Cambridge, MA: MIT Press.
- Beer, R.D., Chiel, H., Quinn, K., Espenschied, S., & Larsson, P. (1992). A distributed neural network architecture for hexapod robot locomotion. *Neural Computation*, 4, no. 3:356–365.
- Beer, R.D., & Gallagher, J.C. (1992). Evolving dynamical neural networks for adaptive behavior. *Adaptive Behavior*, 1, 91–122.
- Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.
- Brinkers, M., & den Dulk, P. (1999). The evolution of non-reciprocal altruism. In D. Floreano, J.-D. Nicoud, & F. Mondada (Eds.), *Advances in Artificial Life, ECAL'99*, 499–503, Berlin: Springer.
- Brooks, R.A. (1989). A robot that walks: Emergent behaviors from a carefully evolved network. *Neural Computation*, 1, 253–262.
- Brooks, R.A. (1991). Intelligence without reason. In: *Proceedings of the International Joint Conference on Artificial Intelligence*, 569–595.
- Chalmers, D.J (1990). Syntactic transformations on distributed representations. *Connection Science*, 2, 53–62.
- Churchland, P.S. (1986). *Neurophilosophy: Toward a Unified Science of Mind-Brain*. Cambridge, MA: MIT Press.
- Crick, F. (1988). *What mad pursuit: A personal view of scientific discovery*. New York: Basic Books.
- de Jong, E.D. (1999). Autonomous concept formation. In T. Dean (ed.), *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence IJCAI'99*, 344–349. San Francisco, CA: Morgan Kaufmann.
- Dennett, D.C. (1991). *Consciousness explained*. Boston: Little, Brown.

- Dennett, D.C. (1994). Cognitive science as reverse engineering: Several meanings of “top-down” and “bottom-up”. In: D. Prawitz, B. Skyrms, & D. Westerstahl (Eds.), *Proceedings of the 9th International Congress of Logic, Methodology, and Philosophy of Science*.
- Donders, F.C. (1862). Die Schnelligkeit Psychischer Prozesse. [The speed of psychological processes], *Arch. Anat. Physiol.*, 657–681.
- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.
- Glaser, M.O., & Dungelhoff, F.-J. (1984). The time-course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 640–654.
- Keijzer, F.A. (in press). *Representation and behavior*. Cambridge, MA: MIT Press.
- Krotkov, E.P. & Simmons, R.G. (1996). Perception, planning and control for autonomous walking with the Ambler planetary rover. *International Journal of Robotics Research*, 15, 155–180.
- Lenat, D.B. & Guha, R.V. (1990). *Building large knowledge-based systems*. Reading, MA: Addison-Wesley.
- Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–75.
- Marder, E. & Calabrese, R.L. (1996). Principles of rhythmic motor pattern generation. *Physiological Reviews*, 76, 687–717.
- Mataric, M.J. (1991). Navigating with a rat brain: A neurobiologically-inspired model for robot spatial representation. In: J.-A. Meyer & S. Wilson (Eds.) *Proceedings of the First International Conference on Simulation of Adaptive Behavior*, Cambridge, MA: MIT Press.
- Moravec, H.P. (1982). The Stanford Cart and the CMU Rover. In: *Proceedings of the IEEE*, 71 (7), 872–884.
- Newell, A. & Simon, H.A. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the Association for Computing Machinery*, 19, 113-126.
- Nilsson, N.J. (1984). Shakey the robot. In: N.J. Nilsson (Ed.), *SRI A.I. Technical Note 323*, April.
- Nolfi, S., & Floreano, D. (1998). Co-evolving predator and prey robots: Do ‘arms races’ arise in artificial evolution? *Artificial Life*, 4(4).
- Pylyshyn, Z.W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.
- Reynolds, C.W. (1987). Flocks, herds, and schools: A distributed behavior model. *Computer Graphics*, 21(4), 25–34.

- Rumelhart, D.E., & McClelland, J.L. (1986). On learning the past tense of English verbs. In: J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models.*, 216–271, Cambridge, MA: MIT Press.
- Simmons, P. & Young, D. (1999). *Nerve cells and animal behavior*. Cambridge: Cambridge University Press.
- Steels, L. (1995). Intelligence - dynamics and representations. In: L. Steels (Ed.), *The Biology and Technology of Intelligent Autonomous Agents*. Berlin: Springer-Verlag.
- Steels, L. (1997). Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. In: J. Hurford, C. Knight and M. Studdert-Kennedy (Eds.), *Evolution of Human Language*. Edinburgh: Edinburgh Univ. Press.
- Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.
- Sutton, R.S. & Barto, A.G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55, 189–208.
- Walter, W.G. (1950). An imitation of life. *Scientific American*, 182(5), 42–45.
- Webb, B. (1994). Robotic experiments in cricket phonotaxis. In: D. Cliff, P. Husbands, J.-A. Meyer, & S. Wilson (Eds.) *Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, Cambridge, MA: MIT Press.
- Werner, G.M. (1994). Using second order neural connections for motivation of behavioral choices. In: D. Cliff, P. Husbands, J.-A. Meyer, & S. Wilson (Eds.) *Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, Cambridge, MA: MIT Press.
- Wilson, S.W. (1991). The animat path to AI. In: J.-A. Meyer & S. Wilson (Eds.) *Proceedings of the First International Conference on Simulation of Adaptive Behavior*, Cambridge, MA: MIT Press.