Building "Fungus Eaters": Design Principles of Autonomous Agents

Rolf Pfeifer
AI Lab, Computer Science Department
University of Zurich, Winterthurerstrasse 190
CH-8057, Zurich, Switzerland
pfeifer@ifi.unizh.ch

To appear in: Proc. SAB'96

Abstract

We describe a set of design principles for building "Fungus Eaters". "Fungus Eaters" are complete autonomous systems. The goal is to extract and describe in a compact way a large part of the insights which have been acquired in the animats field. The principles have been developed from a cognitive science perspective. Although they represent only a very modest beginning, they make immediately clear what sort of ideas about intelligence and cognition they endorse. They all contrast sharply with classical thinking. Moreover, they provide powerful heuristics for design.

1 Introduction

In their review paper of the first SAB conference in 1990, Jean-Arcady Meyer and Agnès Guillot argue that the animat approach will play an important role in resolving some of the fundamental controversies in the study of intelligence or cognition (Meyer and Guillot, 1991). Four years later, at the third SAB conference, they propose three types of goals for animat research, short term, intermediate term, and ultimate goals. In the short term it is the discovery and exploration "… of architectures and working principles that allow a real animal, a simulated animal, or a robot to exhibit a behavior that solves a specific problem of adaptation in a specific environment." (Meyer and Guillot, 1994, p. 7). In the intermediate term, it is the generalization of this knowledge in order to better understand the relation between architectures, working principles and adaptive performance vis-à-vis different types of environments. The ultimate goal, then, is to understand the adaptive value and working principles of human cognition. They conclude by stating that "… the domain is in definite need of theoretical advances that could provide useful generalizations of still highly disparate pieces of knowledge" (p. 8). This paper is an attempt to make a—however modest—contribution towards generalization. The contribution will be in the form of a set of design principles of autonomous agents.

Currently, there is no generally accepted theoretical framework. Although there have been some efforts at developing overarching theories, they are typically only recognized and taken up by a small part of the community. Examples are the "Behavioral Economics" approach (McFarland and Bösser, 1993), the dynamical systems approach (Beer, in press; Steinhage and Schöner, in press), and the evolutionary approach (for a review, see Harvey et al., in press). Maes, in a review paper, tries to capture some general principles contrasting the traditional and the animat approach (Maes, 1992).

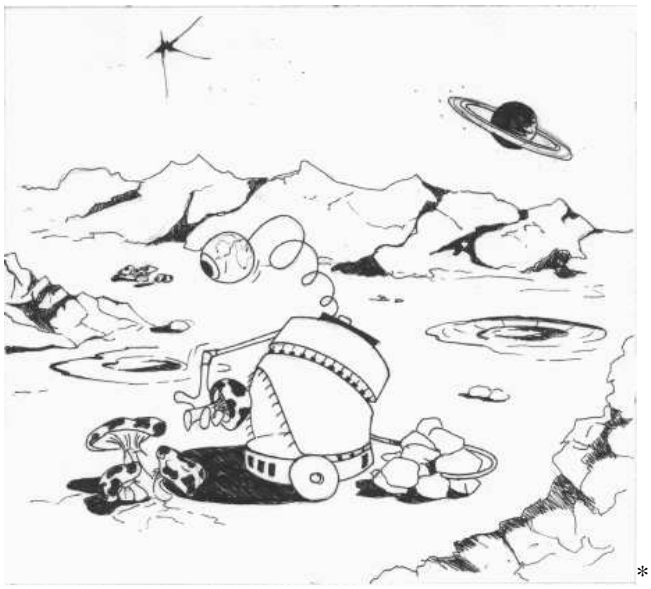

Figure 1: A "Fungus Eater" ingesting fungus on a distant planet. It has to perform its task autonomously while maintaining its energy supply (Cartoon by Isabelle Follath, Zurich).

In this paper we will focus on "Fungus Eaters" or real-world autonomous agents. We do not further discuss work that involves simulation only. "Fungus Eaters" are a particular species of animats. The term is inspired by Masanao Toda's seminal book entitled "Man, Robot, and Society" (Toda, 1982). Briefly, "Fungus Eaters" are complete autonomous creatures, sent to a distant planet for collecting uranium ore. They have to worry about energy supply—they feed on a particular type of fungus that grows on the planet—and predators, while performing their task (figure 1). Toda suggested the study of "Fungus Eaters" out

of a dissatisfaction with the way psychology, in particular cognitive psychology, was going at the time. The main point was that we should study "complete" systems, however simple, rather than only isolated faculties like planning, memory, or decision making.

One of the reasons for the lack of consensus is that the field is very new. Another one is that what we consider to be a good design of an autonomous agent or an interesting and valuable theory, depends on the goals we have in mind. If we want to build robots that collect garbage we are looking for something very different than the biologist who is trying to understand evolution. Or the computer scientist who is interested in the power of evolutionary algorithms is after something else than the psychologist trying to understand cognition. The interdisciplinary nature of the field adds to its diversity.

Although this diversity bears a lot of creative potential, it might nevertheless be useful to try and ferret out some of the accumulated insight and represent it in a compact form. We have tried to capture some of the results as a set of "design principles of autonomous agents". The principles presented here have emerged out of five years of intensive research on various aspects of the animat field, or "New AI", and prior to that over 10 years of research in traditional AI and psychological modeling (e.g. Pfeifer, 1988; Pfeifer, 1994; Pfeifer, 1995; Pfeifer et al., 1989; Pfeifer and Verschure, 1992, 1995). They do not constitute a "theory", but they could represent a starting point for discussion. Ultimately, the idea would be to discover the "theory" from which these principles can be derived. We put "theory" in quotes to indicate that it is an entirely open question whether there will ever be one unifying theory of intelligence or cognition. We are not saying that everyone should agree with these principles. But we do hope that they will help to focus a debate about the underlying principles of naturally intelligent systems.

We begin with a short argument of why we chose the form of "design principles". We then present some reflections on the design process from a cognitive science and an engineering perspective. Then we describe a set of design principles. We conclude with some comments on what we have achieved and what should be done next.

2 Why "Design principles"?

The short answer to this question is that the design perspective is highly productive. The animat approach is by definition synthetic. The underlying slogan is "understanding by building". Design principles provide guidance on how to build animats. The way we build our animats is a manifestation of our views of intelligence. One purpose of the design principles is to make this knowledge explicit. The great advantage of the synthetic approach is, of course, that we have built the agents ourselves, i.e. we know what is in our systems, and that we can experiment with alternatives as much as we like. This experimental freedom also accounts for the popularity of computer simulation models.

In this paper we do not want to study simulation but "Fungus Eaters". "Fungus Eaters" are "complete" in the sense that everything needed for behaving in the real world has to be there. It is not sufficient to model only one aspect, say its memory or its perceptual system. On the one hand, this makes it harder, but on the other, it constitutes the real power of the approach .

There is an additional point that makes the design perspective especially attractive. Natural animats, i.e. animals, can be productively viewed from a designer's perspective: evolution as a designer, perhaps a blind and slow one, but nevertheless a designer, and a good one at that (e.g. Dawkins, 1986). McFarland's "animal robotics" approach capitalizes on this point (e.g. McFarland and Bösser, 1993).

Before discussing the design principles, let us briefly look at some issues in design.

3 Engineering and cognitive science

Assume that the task is to build a robot that collects ping-pong balls in a particular room as quickly as possible. Figure 2 illustrates two alternative designs. The solution on the left shows a powerful vacuum cleaner, sucking in the ping-pong balls at great speed. On the right, there is a robot with sensors and manipulators, and with mechanisms that enable the robot to learn distinctions between different kinds of objects and to learn grasping and carrying light, delicate objects without hurting them. From an engineering perspective, the robot on the left is perfect. The only considerations are performance and price. The performance criterion in this case is obvious, namely the number of balls collected per unit time. It turns out to be much harder to evaluate the performance of autonomous agents. There are promising first attempts (e.g. Gat, in press; Hemelrijk and Lambrinos, 1994; Mataric, 1995; Smithers, 1995), but there is no consensus. Now, the design principles may also be used to assess whether a particular design is of potential interest from a cognitive science point of view.

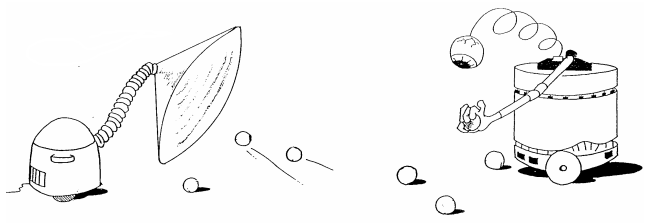

Figure 2: One the left the engineering solution, on the right, the cognitive science solution.

In a cognitive science context, what matters is of a completely different nature than what is relevant for engineering. While performance is certainly a criterion, it is

by no means the only one, nor is it the most important one. In cognitive science the important question is what we can learn about intelligence from our robot. And it seems that the robot on the right in figure 2 can teach us more. The kinds of behaviors it can display are more interesting (flexibility and adaptivity). From the robot on the left we can learn about good engineering, but only little about cognitive science.

4 Design principles of autonomous agents

4.1 Types of explanations

There is a kind of "meta principle" that has to be endorsed if the design principles are to make sense. The meta principle suggests that designs for "Fungus Eaters" always be evaluated from three different perspectives, namely functional, learning and development, and evolutionary. Since our goal is to understand intelligence, we should always keep these three types of explanations in mind. Experience has shown that they contribute in complementary ways to our understanding. While we may choose to focus on one of them we should demonstrate compatibility with the others.

The *functional* perspective[1] explains why a particular behavior is displayed by an agent based on its current internal and sensory state, given its physical set-up. Often, this kind of explanation is used in engineering. But also in cognitive science it is highly productive. Just think of the creative nature of the Braitenberg vehicles, where it is surprising how seemingly sophisticated kinds of behavior result from very simple mechanisms (Braitenberg, 1984).

The *learning and developmental* perspectives not only resort to the current internal state, but to some events in the past in order to explain the current behavior. They provide explanations of how the actual behavior came about. The distinction between learning and development is that development includes maturation of the organism, whereas learning does not.

Evolutionary explanations put the agent into the context of an evolutionary process. The fact that we argue with evolutionary principles does not mean that we have to reproduce evolution in simulation. Biologists have talked for many years about evolution without making simulation models. We have included evolutionary consideration throughout the paper, but we do not specifically elaborate the design principles underlying simulated evolution. For an excellent review, see Harvey et al. (in press).

There is a somewhat orthogonal perspective on intelligent systems, namely the one of societies of agents. A satisfactory explanation of intelligence would also have to include aspects of social systems. While at the functional level, society is not an issue, in ontogenetic and phylogenetic development, it is an essential perspective. We will not go further into this aspect in this paper.

[1]The term "functional" is used in different ways in the literature. Here the term is used to distinguish one level of explanation from a learning/developmental and an evolutionary one.

4.2 Classes of principles

There are three classes of design principles. The first one concerns the kinds of agents and behaviors that are of interest from a cognitive science perspective. The second concerns the agent itself, its morphology, its sensors and effectors, its control architecture, and its internal mechanisms. The third class contains principles that have to do with ways of thinking and proceeding, with stances, attitudes, and strategies to be adopted in the design process. Because of space limitations we will focus on the first and the second class, and only briefly mention the third. An overview of the principles is given in table 1.

Table 1: Summary of design principles

Principle	Name
Types of agents of interest, ecological niche and tasks	
1	The "complete agents" principle
2	The "ecological niche" principle
Morphology, architecture, mechanism	
3	The principle of parallel, loosely coupled processes (the "anti - homunculus" principle)
4	The "value" principle
5	The principle of sensory-motor coordination
6	The principle of "ecological balance"
7	The principle of "cheap designs"
Strategies, heuristics, stances, metaphors	
8	"Frane-of-reference" principle
9	"Constraints" principles
10	Compliance with principles
	etc.

4.3 Type of agents, ecological niche, and tasks

Principle 1: The "Fungus Eaters" principle

As pointed out above, the kinds of agents of interest are the "Fungus Eaters". They are "complete systems", i.e. systems capable of performing a set of tasks in the real world independently and without human intervention. In other words, they are (a) autonomous, (b) self-sufficient, (c) embodied, and (d) situated.

Principle 1a: The agents must be *autonomous*, i.e. they have to be able to function without human intervention, supervision, or instruction.

Principle 1b: The agents must be s*elf-sufficient,* i.e. they have to be able to sustain themselves over extended periods of time. They have to be able to perform a set of tasks, including maintaining themselves (keeping a sufficient

energy level, keeping clean, lubricated, undamaged, etc.), without incurring an irrecoverable deficit in any of its resources. This principle imposes constraints on the architecture (see below).

Principle 1c: The agents must be *embodied*, i.e. they must be realized as a physical system capable of acting in the real world. Although simulation studies can be extremely helpful in designing agents, building them physically typically leads to surprising new insights. This point has been forcefully made by Brooks (1991). Physical realization often facilitates solutions which might seem hard if considered only in an information processing context. An agent existing only in simulation would not be complete.

Principle 1d: The agents must be *situated*, i.e. the whole interaction with the environment must be controlled by the agent itself, i.e. the world must always be seen from the perspective of the agent. Moreover, the agent has to be able to bring in its own experience in dealing with the current situation.

The "Fungus Eater" perspective implies that the agent be studied over extended periods of time. This time span is relevant because we are specifically interested in how agents evolve over time, either on an ontogenetic time scale (which includes learning), or an evolutionary one. In physical agents there is at best a learning perspective—developmental and evolutionary ones are confined to simulation because of technological problems.

An example of what a "Fungus Eater" might look like is shown in the cartoon in figure 1. True "Fungus Eaters" that fulfill all the criteria of principle 1 still do not exist.

Note the contrast to classical views of intelligence where often only performance on one particular problem solving task was at issue.

References supporting this principle include: Brooks, 1991; McFarland and Bösser, 1993; Toda, 1982.

Principle 2: The principle of the ecological niche

There is no universality in the real world. Animats are always designed for a particular niche. The concepts of autonomy, self-sufficiency, deficits, etc. only make sense with respect to an ecological niche. This implies defining all the tasks the agent has to fulfill. Note that the definition of the task is, in a sense, independent of the agent itself. The designer decides what the tasks of the agent are and he will design it such that it will accomplish them. This does not mean that there must be an explicit representation of the task within the agent. This point is nicely illustrated by one of Maja Mataric's remarks about the behavior of her robots: "They're flocking, but that's not what they think they are doing" (quoted in Dennett, in press).

The description of the ecological niche also includes the kinds of possible competition (e.g. McFarland, 1991). A garbage collecting robot has to compete with other robots, but also with other machines and with humans. Moreover, environments may be characterized formally. A well-known example is the characterization in terms of so-called sensory-state machines, as class 0, 1 or 2 environments (Wilson, 1991).

The definition of the ecological niche provides useful constraints for the design of the agent. An illustrative example is Ian Horsewill's robot Polly. It is based on a cheap vision system that exploits the fact that office floors are flat. If the floors are flat, a higher y-coordinate implies that the object is further away (given the object is standing on the ground). This is illustrated in figure 3. In addition, learning problems that are intractable if considered from a purely computational view, often turn out to be benign, if the constraints of a particular ecological niche are taken into account (see below, principle 7).

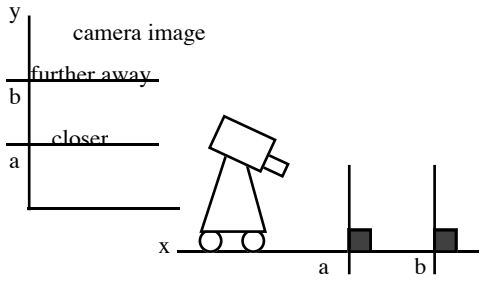

Figure 3: Ian Horsewill's robot, Polly, exploiting constraints of an ecological niche (flat floors in offices).

Evaluation of agents always has to be done with respect to a particular ecological niche, otherwise a comparison of performance is not possible.

Classical views of intelligence do not include the notion of an econiche because programs are restricted to virtual spaces.

References supporting this principle include: Horsewill, 1992; McFarland, 1991; McFarland and Bösser, 1993.

4.4 Morphology, architecture, mechanism

Principle 3: The principle of parallel, loosely coupled processes

In essence, this principle states that intelligence or cognition is emergent from a large number of parallel, loosely coupled processes. These processes run asynchronously and are largely peripheral, requiring little or no centralized resources. Principle 3 could also be called the "anti-homunculus" principle. It is directly motivated from biology.

A strong proponent of this principle is Brooks (e.g. 1991). The Braitenberg vehicles (Braitenberg, 1984), the extended Braitenberg architectures (e.g. Scheier and Pfeifer, 1995), Action Selection Dynamics (Maes, 1991), and PDL (Steels, 1992) can be seen in the same spirit. In all of these approaches, there is no "faculty" deciding on what to do next, i.e. there is no centralized action selection mechanism.

One of the main claims made here is that coherent behavior can be achieved without central control. A beautiful example that fully endorses this principle is the Cog project (Brooks, 1994; Brooks and Stein, 1993). In our own work we have applied this principle in all our agents. They employ an Extended Braitenberg Architecture (EBA), a straightforward generalization of standard Braitenberg architectures (e.g. Lambrinos, 1995; Scheier and Pfeifer, 1995).

While this principle is accepted by many researchers where lower levels of intelligence (e.g. insects, reptiles) are concerned, it is often contested when applied to human cognition. We feel that the principle should be maintained much longer and not given up until there is unequivocal evidence for the need of "higher" processes (Pfeifer, 1995).

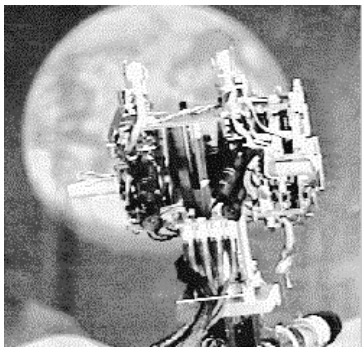

Figure 4: Cog—a robot displaying coherent behavior based on many parallel, loosely coupled processes.

The principle of parallel, loosely coupled processes contrasts sharply with classical thinking where a centralized seat of intelligence is assumed. Classical thinking does not object to parallel processes. The objection is that coherence cannot be achieved unless there is central integration.

References supporting this principle include: Braitenberg, 1984; Brooks, 1991; Brooks and Stein, 1993; Maes, 1991; Steels, 1992; Scheier and Pfeifer, 1995; Pfeifer and Scheier, in press

Principle 4: The "value" principle

This principle states that the agent has to be embedded in a value system, and that it must be based on self-supervised learning mechanisms employing principles of self-organization. If the agent is to be autonomous and situated it has to have a means to judge what is good for it and what isn't. This is achieved by a value system, a fundamental aspect of every "Fungus Eater" and more generally of every animat.

There is an implicit and an explicit aspect of the value system. In a sense, the whole set-up of the agent constitutes value: the designer decides that it is good for the agent to have a certain kind of locomotion (e.g. wheels), certain sensors (e.g. IR sensors), certain reflexes (e.g. turn away from objects), certain learning mechanisms (e.g. selectionist learning), etc. These values are implicit. They are not represented explicitly in the system. To illustrate the point, let us look at reflexes for a moment. Assume that a garbage collecting robot has the task to collect only small pegs and not large ones. Moreover, it should learn this distinction from its own perspective. The agent is equipped with a number of reflexes: turning away from objects, turning towards an object, and grasping if there has been lateral sensory stimulation over a certain period of time. The value of the first reflex is that the agent should not get damaged. The second and the third reflex increase the probability of an interesting interaction. Note that this interpretation in terms of value is only in the eye of the designer—the agent will simply execute the reflexes.

These reflexes introduce a bias. The purpose of this bias is to speed up the learning process because learning only takes place if a behavior is successful. If the behavior is successful, i.e. if the agent manages to pick up a peg, a value signal is generated. In this case, an *explicit* value system is required. In this way, the intuition that grasping is considered rewarding in itself, can be modeled. Figure 5 shows a learning robot, receiving a value signal because it has successfully grasped an object.

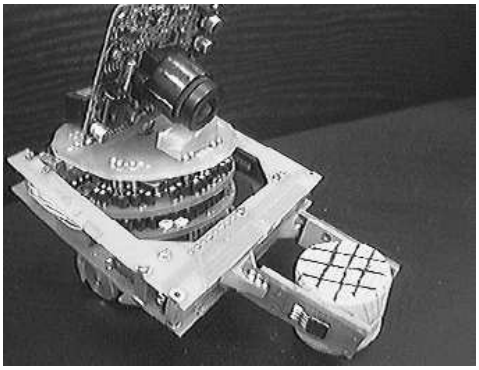

Figure 5: The garbage collecting robot has succeeded in grasping a small peg. An explicit reinforcement signal is generated which enables the robot to eventually learn the distinction between small (graspable) and large (non-graspable) pegs.

According to the "value" principle, the learning mechanisms have to be based on principles of self-organization, since the categories to be formed are not known to the agent beforehand. Examples are competitive schemes (e.g. Kohonen, 1988; Martinetz, 1994), or selectionist ones (Edelman, 1987).

If we were only interested in performance, there would be easier solutions. But the processes described are fundamental for understanding cognition. There is increasing evidence that categorization and concept formation in human infants is strongly based on value

systems and processes of self-organization (Thelen and Smith, 1994).

This view of value systems and self-organization contrasts with classical thinking. The metaphor of information processing that underlies traditional AI and cognitive science, cannot accommodate self-organization. The "value" principle is closely related to the principle of sensory-motor coordination and ecological balance.

References supporting this principle include: Edelman, 1987; Pfeifer and Verschure, 1992; Pfeifer and Scheier, in press; Thelen and Smith, 1994;

Principle 5: The principle of sensory-motor coordination

This principle states that the interaction with the environment is to be conceived as a sensory-motor coordination. Sensory-motor coordination involves the sensors, the control architecture, the effectors, and the agent as a whole. A consequence of this principle is that classification, perception, and memory should be viewed as sensory-motor coordinations rather than as individual modules (e.g. Dewey, 1896; Douglas, 1993; Edelman, 1987).

Normally, perception is viewed as a process of mapping a proximal (sensory) stimulus onto some kind of internal representation. The enormous difficulties of classical computer vision to come to grips with the problem of invariances suggests that there may be some fundamental problems involved. Viewing perception as sensory-motor coordination has a number of important consequences.

From an information theoretic view, the sensory-motor coordination leads to a dimensionality reduction of the high-dimensional sensory-motor space (Pfeifer and Scheier, in press). This reduction allows learning to take place even if the agent moves. In fact, movement itself is beneficial since through its own movement, the agent *generates* correlations in the interaction with the environment. The second important aspect of sensory-motor coordination is the generation of cross-modal associations, including proprioceptive cues originating from the motor system (Thelen and Smith, 1984; Scheier and Lambrinos, 1996).

Additional support for the principle of sensory-motor coordination comes from developmental studies. There is a lot of evidence that concept formation in human infants is directly based on sensory-motor coordination (Thelen and Smith, 1984; Smith and Thelen, 1993; see figure 6). The concepts of humans are thus automatically "grounded". Similarly, if this principle is applied to artificial agents, the latter will only form fully grounded categories. The symbol grounding problem is really not an issue—anything the agent does will be grounded in its sensory-motor coordination. Note that the terms categorization and concept building are entirely observer-based. They relate only to the behavior of the infant, not to any sort of internal mechanism.

There is another kind of approach that closely relates to this principle, namely active vision (e.g. Ballard, 1991). Vision is not seen as something that concerns only input, but movement is considered to be an integral aspect.

Figure 6: Infant categorizing objects and building up concepts while engaged in sensory-motor coordination.

As already alluded to, this view contrasts with the traditional view of perception as a process of mapping a proximal stimulus onto an internal representation. In the view proposed here, the object representation is in the sensory-motor coordination. "Recognizing" an object implies re-enacting a sensory-motor coordination. Most objections to this view of perception have their basis in introspection. The latter has long ago been demonstrated to be a poor guide to research (Nisbett and Wilson, 1977).

References supporting this principle include: Ballard, 1991; Dewey, 1896; Douglas, 1993; Edelman, 1987; Thelen and Smith, 1994; Smith and Thelen, 1993; Scheier and Lambrinos, 1996; Pfeifer and Scheier, in press; Scheier and Pfeifer, 1995;

Principle 6: The principle of "ecological balance"

The principle of "ecological balance" states that there has to be a match between the "complexity" of the sensors, the actuators, and the neural substrate. Moreover, it states that the tasks have to be "ecologically" adequate. The way the term "complexity" is used here, appeals to our everyday understanding: a human hand is more complex than a forklift, a CCD camera more complex than an IR sensor.

From this principle we can get considerable leverage. Let us look at an example illustrating how *not* to proceed. Assume that we have a robot with two motors and a few IR sensors, say the robot Khepera™. In some sense, this design is balanced due to the intuition of the engineers that built it (except that its processor is too powerful if it is fully exploited). Assume further that some researchers have become frustrated because with the IRs they can only do very simple experiments. They would like to do more interesting things like landmark navigation.

The next logical step for them is to add a CCD-camera. It has many more dimensions than the few IR sensors. The rich information from the camera is transmitted to a central device where it is processed. This processing can, for example, consist in extracting categories. But the categories are formed as a consequence of a sensory-motor

coordination. Because the motor system of the agent is still the same, the resulting categories will not be much more interesting than before (although they may be somewhat different). Trying to build categories using only the visual stimulation from the camera (not as a sensory-motor coordination) would violate principle 5. Classical computer vision has violated this principle—and the problems are well-known. It would be a different story if, together with the CCD camera, additional motor capabilities would have been added to the robot, like a gripper or an arm of sorts. Figure 7 shows a balanced design on the left , an unbalanced one in the middle, and again a more balanced one on the right.

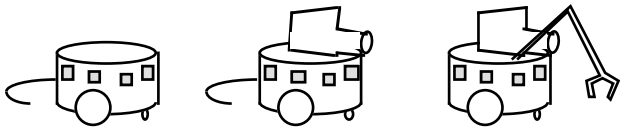

Figure 7: Balanced design on the left, unbalanced design in the middle, and again more balanced design on the right.

An approach that is fully compatible with the principle of "ecological balance" is again the Cog project. More sophistication on the sensor side (two eyes, each with a camera for peripheral and foveal vision), is balanced by more complexity on the motor side. The arm and the hand are quite sophisticated. Moreover, the head and the eyes can all move which leads to a system of a very large number of degrees of freedom. A lot of the processing is done peripherally, and the central processing capability is not inflated artificially. It is not surprising that Cog fulfills this design principle. It was Brooks who pointed out that tasks need to be ecologically appropriate (Brooks, 1990). In particular he argued that "elephants don't play chess." We couldn't agree more.

Important evidence for this principle comes also from studies in infant psychology by Bushnell and Boudreau (1993). Their results suggests that there is in fact a kind of co-evolution in the sensory-motor development of the infant. Roughly speaking, acuity of visual distinctions highly correlates with precision of motor movement.

Again, this view sharply contrasts with traditional AI and cognitive science, where intelligence was seen as centralized information processing, with no, or very little consideration given to the physical set-up. A concept like "ecological balance" would not make sense in that framework.

References supporting this principle include: Brooks, 1991, 1994; Pfeifer, 1995; Smith and Thelen, 1993; Bushnell and Boudreau, 1993.

Principle 7: The principle of "cheap design"

The principle of "cheap design" states that good designs are "cheap". This requires a bit of explanation. "Cheap", as used here, includes various components.

First, cheap means literally cheap. If the robot is built cheaply, we have to worry about physics. For example, the well-known Puma™ arm is not cheap. It is, in a way, "too good": it can be programmed without having to worry much about the real world. In a sense, the real world has largely been taken care of by the engineers. The programmer can simply choose the angles, and the arm will move to the requested position, as long as it is physically possible. If the arm is worse, if it is not so neatly engineered, the programmer has to worry much more about forces, about eigenfrequencies, about sensory-motor coordination, about interacting with the environment. Incorporating considerations about the physics into designs, typically leads to better and more robust designs. In this sense, cheap means capitalizing on the system-environment interaction.

A lovely illustration of exploitation of the physics is insect walking. Leg coordination in insects does not require a central controller. There is no internal process corresponding to global communication between the legs, they communicate only locally with each other (e.g. Cruse, 1991). But there *is* global communication between all the legs, namely through the environment. If the insect lifts one leg, the force on all other legs is changed instantaneously because of the weight of the insect. This communication is exploited for the purpose of coordination.

Second, cheap means parsimonious in the traditional sense of Occam's razor. If there are several models or designs achieving the same task performance, the most parsimonious model is considered the best. Even if the term "parsimonious" is subject to debate, the general idea is clear and generally accepted as a scientific principle. But of course, this depends on what we consider to be the task and the environment (principles 1 and 2). If the environment is subject to considerable change, and if the changes are unpredictable, it may be necessary to equip the agent with resources that it currently does not require to perform its task. Such a design would still be cheap in the sense used here. Depending on the adaptive requirements, even an Edelman-style system, based on selectionist mechanism would be considered cheap, much cheaper than preprogrammed systems, where all the potential situations would have to be foreseen. Having a system with general capabilities that can be exploited in specific situations, can be a cheap strategy.

And third, cheap means exploiting the constraints of the ecological niche. Above we have already seen that some behaviors can be achieved much more efficiently. An example is Horsewill's robot Polly, which exploits the fact that office floors are flat (see figure 3). Learning systems do not have to be universal: only very rarely is there a need in

the real world to learn something odd like an XOR function. It has even been experimentally shown, that natural systems perform poorly on XOR learning tasks (e.g. Thorpe and Imbert, 1989). And it is hard to think of natural situations in which the ability to solve an XOR problem would confer an advantage. "In general, if two cues both signal that food is about to arrive, when the two are present at the same time, the food is even more likely to appear!" (Thorpe and Imbert, 1989, p. 85). As a consequence much simpler neural networks may be used.

Focusing on cheap designs has the additional advantage that the limitations of a design, the ecological niches in which it will function, become immediately evident.

It is interesting to note that cheap designs in the sense discussed here imply ecological balance. Inflating one part, like building a huge brain while leaving sensors and effectors at the same level of complexity, will in any case be too expensive.

This view of cheap designs does not really have an analog in classical AI and cognitive science. There is no embodiment, there is no physics to be exploited, and there are no interesting interactions with the environment. The only overlap seems to be Occam's principle. But all the rest does not make sense in a classical perspective. Again, we see very clearly the fundamentally different view of intelligence endorsed here.

There is an interesting relation of the principle of "cheap design" to societies of animats. Often, tasks can be accomplished much cheaper by having a society of less sophisticated agents, rather than having one or only a few highly complex individuals (e.g. Mataric, 1995).

References supporting this principle include: Brooks, 1991; Horsewill, 1992; Franceschini et al., 1992; Pfeifer, 1993, 1995; Thorpe and Imbert, 1989.

4.5 Strategies, stances, metaphors

This category of design principles concerns the design process itself. Rather than constraining the designs of the agents directly as the set of principles outlined above, they provide suggestions on how to proceed. These principles are less well articulated and will not be discussed here. They include compliance with the design principles, taking the "frame-of-reference problem" into account (Clancey, 1991), incorporating constraints of the ecological niche, capitalizing on system-environment interaction, viewing the complete agent as a dynamical systems, etc.

5 Discussion

The design principles outlined above do not cover all the insights of the very rich field of animats. But we do believe that they capture a large part of the most essential aspects of what has emerged from pertinent research in the area. The principles described may seem somewhat vague and overly general, but they are enormously powerful as heuristics, providing guidelines as to what sorts of experiments to conduct next and what agents to design for future experiment. In order to achieve some degree of generality we have deliberately left out a lot of detail. These principles not only help us evaluate existing designs, but they get us to ask the right questions.

As mentioned initially we have not specifically discussed simulated evolution. Eventually, we may include some pertinent principles into our current set. In a number of places we have resorted to evolution for explanation. If evolutionary robotics gets to a stage where not only control architectures, but also morphology, sensors, effectors, and whole bodies can be evolved, it will be fascinating to see whether our principles also hold for these evolved creatures. A prerequisite is, of course, that the simulation environment reflects our laws of physics. Will these creatures also have value systems and self-organizing schemas? Will they also exploit the physics in interesting ways? If the creatures turned out to obey our design principles this would add additional force to them. But that remains to be seen.

In the future we might be looking for something more formal, than merely a set of verbally stated design principles. Eventually, this will certainly be necessary. As pointed out initially, the mathematical theory of dynamical systems is a promising candidate. But since we are dealing with "Fungus Eaters", i.e. complete multifaceted systems, it may be a while before we have a formal theory of "Fungus Eaters". This does by no means exclude productive use of formal methods to study specialized issues like learning algorithms, problems of mechatronics, etc.

What is needed right now is an in-depth discussion of these principles. They have to be revised and the list of principles has to be augmented. Moreover, an appropriate level of abstraction has to be found. It may turn out that the principles will be more useful if they are more concrete. However, that would imply the well-known trade-off between generality and direct applicability.

7 Conclusions

The design principles discussed in this paper communicate a view of intelligence and human cognition that is entirely different from the classical one endorsed by traditional AI and cognitive science. It seems that it would be premature to ask for a theory of autonomous agents. This is why we started with a set of principles that can help us formulate our beliefs about the nature of intelligence in a compact way.

Having a concise way of talking about our views of intelligence is extremely important since we want to convince researchers from other disciplines like psychology, biology, and neurobiology, that novel perspectives and directions can be expected from the animat field. The animate perspective may shed new light on old controversies. Examples are the conundrums involved in perception and categorization. While designing agents—as discussed above—is a fascinating and productive endeavor

in itself, it is a highly creative tool for other scientific disciplines involved in the study of intelligence.

Let us conclude by saying that we hope to have made a small contribution towards Jean-Arcady Meyer and Anne Guillot's quest for theoretical advances and generalizations

Acknowledgments

This research was supported in part by grant # 20-40581.94 of the Swiss National Science Foundation. I would like to thank Christian Scheier, Dimitri Lambrinos, and Ralf Salomon for their inspiring discussions and valuable comments on the manuscript.

References

Ballard, D.H. (1991). Animate vision. *Artificial Intelligence*, **48**, 57-86.

Beer, R. (in press). The dynamics of adaptive behavior: a research program. To appear in: *Robotics and Autonomous Systems, Special Issue on "Practice and Future of Autonomous Agents",* R. Pfeifer, and R. Brooks (eds.).

Braitenberg, V. (1984). *Vehicles: experiments in synthetic psychology*. Cambridge, Mass.: MIT Press.

Brooks, R.A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, **6,** 3-15.

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence*, **47**, 139-160.

Brooks, R.A. (1994). Coherent behavior from many adaptive processes. In: D. Cliff, P. Husbands, J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animats 3. Proc. SAB'94,* 22-29.

Brooks, R.A., and Stein, L.A. (1993). Building brains for bodies. Memo 1439, MIT Artificial Intelligence Laboratory, Cambridge, Mass.

Bushnell, E.M. and Boudreau, J.P. (1993). Motor development in the mind: The potential role of motor abilities as a determinant of aspects of perceptual development. *Child Development*, **64**, 1005-1021.

Clancey, W.J. (1991). The frame of reference problem in the design of intelligent machines. In K. van Lehn (ed.). *Architectures for intelligence*. Hillsdale, N.J.: Erlbaum.

Cruse, H. (1991). Coordination of leg movement in walking animals. In: J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animat 1. Proc. SAB'90,* 105-119.

Dawkins, R. (1986). *The blind watchmaker*. London, UK: Penuin Books (1988) (first published by Longman).

Dennett, D. (in press). Cog as a thought experiment. *Robotics and Autonomous Systems, Special Issue on "Practice and Future of Autonomous Agents"*, R. Pfeifer, and R. Brooks (eds.).

Dewey, J. (1896). The reflex arc concept in psychology. *Psychol. Rev.*, **3** (1981) 357-370; Reprinted in: J.J. McDermott (ed.) *The Philosophy of John Dewey*. Chicago, IL: University of Chicago Press, 136-148.

Douglas, R.J., Martin, K.A.C., and Nelson, J.C. (1993). The neurobiology of primate vision. *Bailliere's Clinical Neurology*, **2**, No. 2, 191 - 225.

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

Franceschini, N., Pichon, J.M., and Blanes, C. (1992). From insect vision to robot vision. *Phil. Trans. R. Soc. Lond. B*, **337**, 283-294.

Gat, E. (in press). Towards principled experimental study of autonomous mobile robots. To appear in *Autonomous Robots*.

Harvey, I., Husbands, P., Cliff, D., Thompson, A., Jakobi, N. (1996). Evolutionary robotics: the Sussex approach. *Robotics and Autonomous Systems, Special Issue on "Practice and Future of Autonomous Agents"*, R. Pfeifer, and R. Brooks (eds.).

Hemelrijk, C.K., and Lambrinos, D. (1994). Peformance of two homing strategies in environments with differently distributed obstacles. *Perac-94*, 352-355.

Horsewill, I. (1992). A simple, cheap, and robust visual navigation system. In: J.-A. Meyer, H.L. Roitblat, and S.W. Wilson (eds.). *From animals to animats 2. Proc. SAB'92,* 129-137.

Kohonen, T. (1988). *Self-organization and associative memory*. Berlin: Springer.

Lambrinos, D. (1995). Navigating with an adaptive light compass. *Proc. ECAL-95,* 602-613.

Maes, P. (1991). A bottom-up mechanism for behavior selection in an artificial creature. In: J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animats 1. Proc. SAB'90,* 238-246.

Maes, P. (1992). Behavior-based artificial intelligence. In: J.-A. Meyer, H.L. Roitblat, and S.W. Wilson (eds.). *From animals to animats 2. Proc. SAB'92*, 2-10.

Martinetz, T. (1994). Topology representing networks. *Neural Networks*, **7**, 505-522.

Mataric, M. (1995). Evaluation of learning performance of situated embodied agents. *ECAL-95*, 579-589.

McFarland, D. (1991). What it means for robot behavior to be adaptive. In: J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animats 1. Proc. SAB'90*, 2-14.

McFarland, D., and Bösser, M. (1993). *Intelligent behavior in animals and robots*. MIT Press.

Meyer, J.-A., and Guillot, A. (1991). Simulation of adaptive behavior in animats: review and prospect. In: J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animats 1. Proc. SAB'90*, 2-14.

Meyer, J.-A., and Guillot, A. (1994). From SAB90 to SAB94: four years of animat research. In: D. Cliff, P.

Husbands, J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animats 3. Proc. SAB'94*, 2-11.

Nisbett, and Wilson (1977). Telling more than we can know: verbal reports of mental processes. *Psychological Review*, **84,** 231-259.

Pfeifer, R. (1988). Artificial intelligence models of emotion. In V. Hamilton, G.E. Bower, and N. Frijda (eds.). *Cognitive perspectives on emotion and motivation* (Proc. NATO Advanced Research Workshop). Amsterdam: Kluwer, 287-320.

Pfeifer, R. (1993). Cheap designs: exploiting the dynamics of the system-environment interaction. Three case studies on navigation. In: Conference on *"Prerational Intelligence"*. Center for Interdisciplinary Research, University of Bielefeld, 81-91.

Pfeifer, R. (1994). The "Fungus Eater" approach to the study of emotion. *Cognitive Studies,* **1**, 42-57 (in Japanese). English version: The ``Fungus Eater'' Approach to Emotion: A View from Artificial Intelligence. Artificial Intelligence Laboratory, University of Zurich, Techreport #95.04.

Pfeifer, R. (1995). Cognition — perspectives from autonomous agents. *Robotics and Autonomous Systems*, **15**, 47-70.

Pfeifer, R., Schreter, Z., Fogelman-Soulie, F., and Steels L. (eds.) (1989). *Connectionism in perspective*. Amsterdam: Elsevier.

Pfeifer, R., and Scheier, C. (in press). Sensory-motor coordination: the metaphor and beyond. To appear in: *Robotics and Autonomous Systems, Special Issue on "Practice and Future of Autonomous Agents",* R. Pfeifer, and R. Brooks).

Pfeifer, R., and Verschure, P.F.M.J. (1992). Distributed adaptive control: a paradigm for designing autonomous agents, in F.J. Varela, and P. Bourgine (eds.) *Proc. ECAL-92*, 21-30.

Pfeifer, R., and Verschure, P.F.M.J. (1995). The challenge of autonomous systems: pitfalls and how to avoid them. In L. Steels and R. Brooks (eds.). *The Artificial Life Route to Artificial Intelligence*. Hillsdale, N.J.: Erlbaum, 237-263.

Scheier, C., and Lambrinos, D. (1996). Categorization in a real-world agent using haptic exploration and active vision. *Proc. SAB'96.*

Scheier, C., and Pfeifer, R. (1995). Classification as sensory-motor coordination: a case study on autonomous agents. *Proc. ECAL-95,* 657-667.

Smith, L.B., and Thelen, E. (eds.) (1993). *A dynamic systems approach to development. Applications*. Cambridge, Mass.: MIT Press, Bradford Books.

Smithers, T. (1995). On quantitative performance measure of robot behaviour. *Robotics and Autonomous Systems*, **15**, 7-133.

Steels, L. (1992). The PDL reference manual. VUB AI Lab memo 92-5.

Steinhage, A., and Schöner, G. (in press). Self-calibration based on invariant view recognition: Dynamic approach to navigation. To appear in: *Robotics and Autonomous Systems, Special Issue on "Practice and Future of Autonomous Agents",* R. Pfeifer, and R. Brooks (eds.).

Thelen, E. and Smith, L. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, Mass.: MIT Press, Bradford Books.

Thorpe, S.J., and Imbert, M. (1989). Biological constraints on connectionist modelling. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, and L. Steels (eds.). *Connectionism in Perspective*. Amsterdam: North-Holland, 63-92.

Toda, M. (1982). *Man, robot, and society*. The Hague, Nijhoff.

Webb, B. (1994). Robotic experiments in cricket phonotaxis. In. D. Cliff, P. Husbands, J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animats 3. Proc. SAB'94,* 45-54.

Wilson, S.W. (1991). The animat path to AI. In: J.-A. Meyer, and S.W. Wilson (eds.). *From animals to animats 1. Proc. SAB'90*, 15-28.