Motivation, Emotion and the Role of Functional Circuits in Autonomous Agent Design Methodology

Erich Prem
The Austrian Research Institute for Artificial Intelligence*
erich@ai.univie.ac.at

Abstract

This paper discusses a new methodological approach to designing software for autonomous agents. For real autonomy such systems must be equipped with a motivational subsystem that drives the agent and selects among its possible behaviors. We present a methodology that supports the design of such a system and discuss its relation to theoretical biology, particularly the work of Jakob von Uexküll. Another issue which is treated more briefly here, is the role of emotion in such an agent, particularly the communicative function of showing emotions.

These issues are discussed in the context of behavior-based control circuits.

1 Introduction

1.1 Autonomy in agents

Despite (or maybe because) of the recent hype in autonomous agents research it often remains unclear why such agents are not treated as conventional computer software. The answer lies in the notion of *autonomy.* Whereas a conventional computer program is sometimes called "agent", the term "autonomous agent" is reserved for software that should exhibit special characteristics. It should be

- self-governing,
- reacting independently, and
- anthropomorphic interpretable or understandable.

Whereas a conventional computer program is often run in order to serve a specific purpose in a given context, e.g. calculate a numerical function, manipulate a database, etc., an autonoumous agent is usually operating in a more independent fashion. The idea is often to have an agent act whenever it "wants"

*During this work Erich Prem was also affiliated with the MIT Artificial Intelligence Lab. The Austrian Research Institute for AI is sponsored by the Austrian Federal Ministry for Science, Research and the Arts.

to do so, not when it is being called explicitely. Moreover, the agent also decides what to do and how to do it. A typical example would be a software agent that selects incoming electronic mail according to criteria it has autonomously established, for example learned on the basis of reinforcement or regularities in a person's mail reading behavior.

Autonomy therefore is concerned with control and independence. From a viewpoint that is oriented on software design this basically means a shift of focus in software development. It is straightforward to see that the design of an agent which is supposed to act independently of explicit calls by a program user must find methods for letting the agent act at appropriate times and having it do the right thing. In order to do this the *goals* of the system user must be well understood, i.e. it must be clear to which ends an agent should act to support the user optimally.

One way to talk about these "goals" is to turn them into goals, needs, and drives for the agent. The agent then, simply by having the "motivation" to act in accordance with a user's goals, chooses the right action at the right point in time.

1.2 Motivation and emotion: terminological pitfalls

Unfortunately, the notions of emotion and motivation are far from being uniquely defined and well-understood. (See also [Read & Sloman 93].) The main source for these problems is the opaqueness of the concepts, i.e. the fact that they can only be observed in a behavioristic, functionalist, or introspective fashion.

The problem with the term *emotion* is that it can mean several very distinct phenomena. Often it is used to describe an "inner" feeling (e.g. *hate*), sometimes authors refer more to the "outer" expressions (e.g. facial expression of *disgust*). The problem with *motivation* is more that its physiological correlates are far from clear. At the current state of research our knowledge about where motivation really comes from is very poor.

This is why, in this paper, *emotion* and *motivation* are regarded as useful metaphors in the design of autonomous agents. Here, we are not suggesting that we can contribute to a deeper understanding of these terms or the associated phenomena. Rather, we make use of these terms in order to support the design of autonomous agents.

1.3 Methodology

When discussing questions concerning parts of the control systems for autonomous agents it must be ensured that it is–at least in principle–possible to talk about these parts independently of the rest of the architecture. It is, however, not completely clear that questions concerning a motivational system can be successfully addressed without having a concrete control system in mind. To date, many such different control systems have been proposed. They range from centralized rule-based methods to systems of independent behaviors [Connell]. (An overview can be found in [Trappl & Petta 95].) The emphasis in

this paper will be on *reactive control systems* as they have been prototypically suggested by [Brooks 89].

This variety is one of the reasons why this paper is concerned with methodological aspects of designing the control system for an agent. Here we are trying to support the *designer* of an agent in her design process of developing (reactive) control systems for autonomous systems, we are not proposing one special control architecture.

2 Physiological Control Systems

Since motivation and emotion are concepts which originated in the domain of biological systems, a brief look at the differences between real physiological control systems and the metaphoric ones in artificially autonomous systems seems practical.

2.1 Physiological Psychology

In [Rosenzweig 89] motivation is considered as a means of controling behavioral states:

> ...how the organism selects among the many options that these systems provide. ...responding to one motive often precludes satisfying another at the same time, and different motives may have to be satisfied in successation. *How bodily systems function so that all the basic needs are satisfied is the overall question of motivation ...*

Rosenzweig and Leiman discuss the following phenomena in the context of motivation:

- sex
- heating/cooling
- drinking
- eating
- regulation of energy
- biological rhythms
- sleeping/waking
- emotions

Very often needs and drives are partitioned in homeostatic and non-homeostatic drives.

Homeostatic drives are less dependent on environmental conditions or learned features of the agent-environment interaction. They depend more on the deviation of certain values in homeostatic processes and are usually characterised by specific optimal values that must lie within exact boundaries. Examples are temperature regulation, hunger, sleep etc.

Non-homeostatic drives possess variable optimal values which often strongly depend on learning and environmental variations like triggers or availability. Examples are sexuality, exploratory drive, emotions.

Other possible characterizations are noxious, cyclical, default, exploratory, and anticipatory drives.

2.2 Differences in needs for agents and people

Obviously, the above list is closely related to a biological system's needs. Agents, however, do not have any implicit needs that could be derived from the fact that they are agents. Virtually *any* goal, need and drive that is introduced to control the agent's behavior will be artificial and can at least, in principle, be selected at the designer's will.[1]

At this point, there seem to be two basic strategies for designing a solution to the motivation problem.

1. teleomimesis
2. teleometaphoresis

Teleomimesis (i) would mean to simulate as much of the biological phenomena as possible. In this case, the designer tries to actually simulate needs to drink, biological rhythms, etc. This strong form of biocentrism, however, does not come without problems. Perhaps most strikingly, such an approach seems extremely artifical. Agents do not have to eat, why then should hunger be simulated?

There is another strategy, something I have preliminarily called *teleometaphoresis.* In this strategy the agent would *not* be equipped with goals that are immediately useless for it, i.e. we would not simulate the need to drink or sleep. We would, however, try to make the agent understand what it means to be thirsty by introducing a sense for diminishing energy sources etc. It must be understood, however, that this still means to exploit only metaphoric similarities between both systems. There are no direct connections between the simulated motivational system and the simulation. Especially, this means we cannot gain any insights into biological motivational systems by studying such a simulating computer program.

Commonalities Although (as mentioned before) needs for agents will always have to be introduced artificially, there are some which do not seem to share the same awkardness as thirst or hunger. One such basic need could be the need to make social contacts.

[1] This, of course, is not to imply any claims about "natural" goals, needs, or drives for human beings.

2.3 Motivation as a Control System

In this paper *motivation* and *needs* are used without their psychological connotations. They are referred to only as elements in a method which supports the design of control structures for agents. The design problem here is the following:

> How can a motivational system be used to control an agent's behavior?

This, as I will show in this paper, leads to the following question.

> How can a system of needs be designed such that the desired behavior will emerge?

Among the reasons for using a motivational subsystem we can find

1. a certain behavioral autonomy
2. and maybe immediately understandable (intuitive) behavior.

Both requirements touch upon different aspects of goals. Requirement 1 (autonomy) implies that the system is capable of pursuing its goals by autonomous established actions. Requirement 2 means that these actions make sense to a (human) observer. We have argued elsewhere [Prem 95] that understanding often means being able to understand the goals with which another system's actions are undertaken. It is, however, not clear how abstract goals can be integrated with (reactive) architectures for autonomous agents. One approach to this problem, the construction of a motivational system and behavioral engagements is discussed below.

2.4 Design steps

The design of an autonmous agent's motivational system according to the method proposed here happens in the following steps.

1. specification of behavior
2. reformulation in teleological terms
3. specification of needs and drives
4. design of the motivational system

The interesting aspect of this methodology is that the list is somewhat reverse to conventional thinking and behavioral science. The design starts with the specification of desired behavior and from this derives a list of needs and drives that will eventually produce the desired behavior when interacting with its environment.

It will be argued below that this change in point of view is central to designing autonomous agents and has severe consequences for the design process as well as the designed architecture itself. This claim will be motivated by comparing artificial agents research with their natural counterpart, i.e. biology and ethology. This will be done in the next section.

3 Theoretical Biology and Teleology

The engineer who tries to develop an autonomous system (a mobile robot or a software agent) is faced with an overwhelming amount of initial design decisions. Some of these concern system architecture, sensors, effectors, methods for solving the task, learning and adaptation methods, ontological aspects, etc. It is among the goals of this paper to show that the engineer also needs to define the autonomous system's goals (or needs). More precisely, the robotic engineer must develop a whole of functional circuits for the artificial system. This functional world guides the design decisions mentioned above.

In what follows we shall distinguish two different problems here:

- design of a motivational system that controls behaviors (controls competition between functional circuits)
- design of a sequence of actions that lead to the satisfaction of a given goal or need (controls one functional circuit)

In order to motivate this approach better, I will use the work of Jakob von Uexküll, an early biologist who is nowadays often regarded as the father of modern ethology.

As soon as 1930 Jakob von Uexküll described a view of biology which bases the study of animals on the animal's view of the world rather than on a scientist's "objective" view of the animal and its environment. The goal with which such a turn in perspective is undertaken is the exact description of phenomena encountered in the real world in a way which allows us to better understand what is going on in nature. Among other things, this means to be able to produce better predictions of how an animal will behave in a given context.

As an example consider the difference between the two following descriptions of the tick's feeding behavior:

1. The tick attacks warm-blooded animals like humans or deer when they make contact with the trees or grass inhabited by the tick.
2. The tick bites when making contact with anything which has a superficial temperature of 37 C and emits a specific chemical substance.

While the first description is immediately easy to understand, the second certainly has a higher predictive value. The analysis which is necessary to come up with the second way of describing the tick behavior consists in a careful study of a tick's sensory organs and reflexes. In fact, the second version is more a description of *how the tick sees the world* in human terms. For the tick there are no humans, deer, trees, grass, etc. All that governs the tick behavior in the feeding context are specific features of two environmental qualities: temperature and chemical concentration.

The next chapter introduces the work of von Uexküll in more detail.

3.1 The worlds of biology

> All efforts to find the reality behind the world of appearances, i.e. with disregard to the subject, have always been doomed to failure, because the subject plays an essential role in the construction of the world of appearances and there is no world beyond this world of appearances. [...] All reality is subjective appearance. [Jakob von Uexküll]

A major part of a biologist's endeavor involves understanding the behavior of animals, besides studying their morphology, physiology, etc. A successful study of animals produces knowledge which will allow us to predict animal behavior in a given environment. Therefore, in turn, it is necessary to understand the environment of the animal. In fact, a great part of biology can be seen as the study of the relations between the animal and its environment. Consequently, the questions what the environment of an animal actually is, how its structure can be determined, and how it should be described are essential to biology.

It was already around the beginning of our century that these questions received the attention they deserve. Jakob von Uexküll developed "Theoretical Biology" as a general framework of how to proceed in the study of the animal, which for him already meant how to proceed in the study of the structure of the animal's environment. Equating these seemingly different subjects of research is based on the idea that the relevant subject matter must be the relation between animal and environment. One depends on the other; there is no possibility to understand an animal's behavior without its environment, nor the environment without any animals put therein.

The study of this relation, however, goes much deeper than it may seem at the first point. It was Uexküll's original proposal that knowledge about this relation can only be gained appropriately, if we gain insight in *how the animal sees the world.*

As a first motivation of this viewpoint take the example of a jumping spider. Careful analysis (as performed by [Dress 52]) reveals that the behavior of the male jumping spider when engaged in either mating or feeding can be described by the following simple rule:

> If it moves, find out whether it has legs in the right places; if it does, mate or avoid it; if it doesn't, catch it. [Land 72]

Interestingly, the stimuli which can be used to produce prey-catching behavior range from small black squares to black circles, crosses, and many other patterns. Mating behavior, however, is only produced by black circles which exhibit "legs" (lines), the more the better.

It is easy to see that it does not make sense to speak about a fly as being caught by the spider. It is more appropriate to describe the set of objects which will cause a jumping reaction and those which will generate a sexual response. To fully account for the animal's behavior it is also necessary to understand *why* a specific behavior is undertaken, e.g. why the spider jumps to a pattern in the set. A possible answer would be: the spider jumps to the pattern in order to try to feed on it.

Uexüll's biological methodology is therefore concerned with the following questions.

- How does the animal perceive its environment?
- In the case of learning: How does the animal construct its reaction?
- What is the purpose of an animal's action?

Among other tasks necessary to answer the first question, the biologist will be forced to study the animal's sensory organs. The second question refers to the problems as to how an animal constructs knowledge about sensory signals and how it learns to react to those signals. The third research task consists in finding the meaning of the actions to the animal, i.e. their purposes. These three domains are described in more detail below.

3.2 Following Kant: environmental perception

> The sensations of the mind become properties of things during the construction of the world, or, one could also say, the subjective qualities construct the objective world. [J.v.Uexküll][2]

To a large extent it was Uexküll's intention to not only follow Kant in his search for the "conditions of the possibility of object constitution," but to also extend this epistemology. Such an extension could only be based on a deeper understanding of what the actual ways are in which we acquire "knowledge" (in the broader meaning of sensory quantity or quality) about the world. It is therefore not surprising that Uexküll's prevailing modes of recognizing the world are the same as Kant's: space and time.

However, at the point where Kant's considerations lead to a discussion of categories as the final set of tools of reason to bring the "manifoldness of experience into the unity of concepts", von Uexküll develops descriptions of sensor spaces. The intention is to describe, how the

> Marking signs of our attention turn into marks of the world.

At first, this investigation is still very much based on a traditional account of how external stimuli are mapped onto entities in the mind. Von Uexküll's task consists in describing every single perceptual quality as perceived by humans. For example, besides of being able to perceive a haptic stimulus as pressure, the human mind can also add a location to this stimulus. A single tactile quality, however, is hardly ever perceived. What we usually feel are neighbor relations of a whole set of tactile stimuli. This is Uexküll's transformation of Kant's statement that space is only the *form* of our perceptive faculty.

> It is not the quality of the local-signs but the form of their arrangement which allows us to perceive haptic sensory signals as extended.

[2] All Uexküll and Kant quotes are author's translations from German.

In a similar approach, Uexküll distinguishes qualities for body parts, directions, distances, etc. as basic elements of tactile and visual perception ("tactile and visual space").

Despite a certain psychological and maybe even ethological turn, this analysis is still very traditional up to the point where Uexküll states that the unmoved eye does not produce a visual space but only a visual area and the unmoved body produces no tactile space but only a tactile area. Only when our own muscles begin to work the area is extended to space. For example, the lens muscles move the visual area back and forth or the arm muscles add a spatial dimension to the tactile area.

In the end, such an analysis results in a description of an individual's sensations of its environment. The important turn is, however, that subjective qualities turn into marks of the world. Since it is usually impossible to know what these qualities are for an animal, i.e. how it perceives its environment, is it the observer's task to discover which qualities of *our* world of appearances are marks in the animal's environment.

3.3 From perception to action

A standard perceptive experience of objects tells us nothing about how to use things and, to Uexküll's opinion, nothing about what these objects really are. In order to arrive at an understanding of what the things mean to the autonomous system, it is necessary to study the system's actions.

Consider the case where a system action is triggered by some environmental situation: The animal subject selects features in the environment and responds with a certain action or behavior. This action in turn influences the environment, which produces new signals for the system.

Therefore, the most important source of knowledge about the things around us is acting with these things. The feedback loop forms a "rule" about how to use an object appropriately. Contrary to what people in the field of Artificial Intelligence have proposed (most prominently maybe M. Minsky in [Minsky 85]) "functions" may not be some additional property attached to an object, but at the very heart of what things actually are.

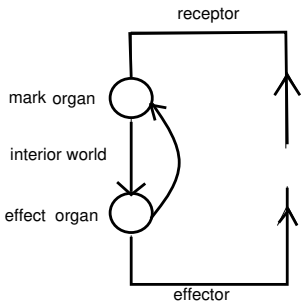

Figure 1: Action circuit as described by J. von Uexküll (1923)

Uexküll motivates this viewpoint with a human example: "A young and

skillful Negro whom I brought from inner Africa to the coast was unable to use a short ladder, because he did not know which sort of *thing* a ladder was. 'I can only see bars and holes', he said. After someone showed him how to climb the ladder, he was very good at using it. Although the ladder could be clearly seen, it was no thing for him, only a meaningless object without purpose. The rule of climbing immediately ordered and formed the ladder."

The following table summarizes notions which are used by Uexküll in a very specific way to distinguish the system's view of the world from the observer's world.

here	*von Uexküll*	*meaning*
object	Ding	objective entity as encountered by the human
thing	Gegenstand	subjective entity as used/learned/recognized by the animal
environment	Umgebung	set and structure of objects around animals as perceived by humans
world	Umwelt	the environment according to the animal

3.4 Functions

In the last example it was not just the act of using the ladder which contributed to constructing the ladder-thing. It is equally important that the usage itself has a specific purpose. Such a purpose turns the object from a collection of merely causally operating parts of physical entities into a meaningful assembly of things which are integrated in a purposeful whole. The essential point is *to understand how the thing is embedded in an action and how this action is embedded in a purposeful interaction with the world.*

The following figure summarizes Uexküll's view of a functional circuit. In the context of this circuit, the sensory impressions of the animal form the mark world (world of marks, *Merkwelt*). The effects which the animal produces in the environment make up the effective world (world of effects, *Wirkwelt*). The world which is formed by the animal due to its control of the world is called the interior world. Here, I have chosen to translate Uexküll's *Umwelt* as "the world" of an animal, which consists of mark world and effective world. Every object within the functional circuit of the animal is regarded from the viewpoint of function, therefore we call it a "thing" – *not* an object.

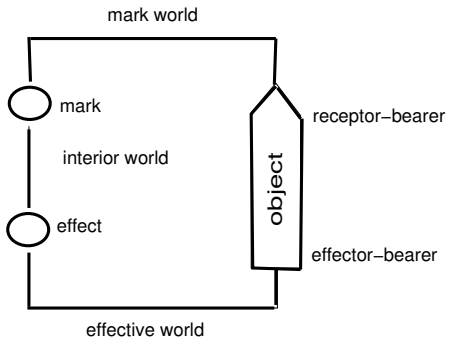

Figure 2: Functional circuit as described by J. von Uexküll (1923)

In order to fully understand the system's world, our task consists in the dissection of the functional world (i.e. the whole of the subject's functional circuits). Note that although this is widely acknowledged, it is usually not described that way. Typical examples for a simple animal are the following functional circuits:

medium The animal responds to deviations of its usual environment.

feeding The animal seeks for food and consumes it.

foe The animal flees any predator.

sex The animal searches for a sexual partner and mates.

New in this view here is that things *outside* the animal must be treated as full of purpose to the animal, *not* as simply causal. Everything must be viewed from the point of function, everything is a *thing*, not a mere object.

There remains the question as to how these different actions are coordinated, i.e. how are single actions put together into one meaningful whole? For Uexküll, every system action contains the plan to annihilate the trigger which caused the system's response (action). For example, a dog is *not* simply running (nor is a bug), it is running *away* or *to* the food; although an observer can maybe never sure about which goal it is that drives the animal.

> Every action goes from mark to mark, from one immaterial factor to the next, which always lies in the future.

To see how markworld and effectors cooperate, the sequence of actions which are controlled in this way can be described by coloring functional circuits, mark bearers and the corresponding worlds of the animal:

To take one of Uexküll's original examples, consider a monkey discovering an apple, taking and eating it. First the apple is seen (optical stimulus), next it is felt (tactile stimulus), finally it is tasted and swallowed. All these different marks are carried by the same object, hence the expression "mark carrier" for the apple.

Let us now try to color the functional circuits and the marks of the object associated with the whole process. The optical marks of the apple make it a part of the "blue" functional circuit; the apple gets ("is") a blue mark, and the corresponding mark signs in the animal's markworld are blue. But also the effective signs in the effector world are blue, which let the monkey grasp the object. The tactile marks, however, with their mark signs and the effector signs for taking the thing to the mouth are, say, red. Taste signs are yellow and finally lead to the activation of (yellow) swallowing. The overall action therefore goes from blue to red to yellow. Again, in every circuit, the mark-thing is nihilated by the effector-thing.

It is important to note that the same object can be a completely different thing at different points in time. The fly for the spider is a blue thing in the

net, but it may be red when flying by.

Summarizing Uexküll's position, there can be no understanding of animals without clarifying how they see the world, or better, what makes up the animal's world. Most notably, no such understanding seems possibly without having gained insight into the animal's meaningful whole of functional circuits.

4 The construction of autonomous systems

The view proposed by theoretical biology (and taken up by ethology later) motivates that the construction of an autonomous system must happen by reversing the process proposed by Jakob von Uexküll. Given a description of some system behavior, we want to arrive at a description of the system's sensory devices, methods for recognizing things in the environment, the generation of necessary behaviors, etc.

The conditions of the possibility of object constitution [Kant] are, of course, constrained by the sensory system. More importantly, they must reflect the system's ability to recognize *things*, not objects. Knowledge about what these things are can only be gained by understanding the different actions of the system. The actions, and with them the behaviors, must be based on understanding functional circuits. It is the primary task of the systems designer to develop this functional world of the autonomous system, and with it, its things, the recognition procedures, and the sensory devices. The designer is essentially left to develop functional circuits based on his own creativity and intuition.

4.1 Functional circuits in artificial systems

A question which we have not discussed so far is, what the relation between the goals of the system and our own goals is. It seems necessary for the designer to select a set of adaptation goals and/or to develop a world of functional circuits for the autonomous system. (*Auto-nomy,* hence, is the wrong term. There is only *hetero-nomy* in such a system.)

These goals will touch upon the human world in a very specific way: they tend to reflect human purposes. An artefact does not and indeed cannot have any original goals of its own. A robot, for example, simply does not need to do anything at all. Whatever the special goal or need is that it may try to pursue, it will be given to it. This may happen implicitly through the design of behavior systems or explicitly through the development of functional circuits.

Mobile and non-mobile robots are discussed next, autonomous software agents are considered in sec. 4.3.

4.1.1 Mobile robots

In the design of the mobile robot *Herbert* [Brooks etal. 88] a methodology of "minimal requirements" was followed. This is similar to what is proposed here; we shall therefore take a closer look at Herbert. The whole of the functional world for Herbert can be described as *soda can collection and delivery.* Let us

now briefly try to dissect Herbert's functional world into appropriate circuits and from this derive the robot's design.

In general, Herbert will wander around and look for soda-cans, or better: s-things which are detected by the laser-scanner. Let us take a look at what happens when Herbert "sees" a soda-can. In Uexküll's terminology, the system finds a thing which carries the mark, for which Herbert possesses a mark-detector. Soda cans are a part of Herbert's mark-world.

Using the coloring scheme which has been introduced above we could also say that Herbert discovers a blue mark-bearer. The blue mark triggers a blue action: reach. The blue reach action turns into a yellow grasp action once the soda can is positioned in between Herbert's grippers. Once grasped, the soda-can bears a red mark, and since the Herbert's motors also do, the robot wanders off.

Accordingly, we find the following things in the world according to Herbert: s-things which are to be reached for, i-things which are to be graped, and finally red t-things which are to be disposed...

4.1.2 Non-mobile autonomous robots

In the case of Herbert, the development of functional circuit descriptions is supported by the fact that there is a rather clear and specific task which Herbert has to fulfill. However, this is not the case for an autonomous humanoid robot. As an example for such a robot, we take Cog, which is currently under development at the MIT AI laboratory [Brooks & Stein 94]. The robot is mounted to an immovable platform. It is supposed to resemble a human from the waist up, torso and head each having three degrees of freedom. Currently, one arm (6 degrees of freedom) is mounted to the robot's body [Williamson 94]. A special humal-like actuator system used for moving the arm consists of spring-coupled (series elastic) actuators [Pratt & Williamson 95]. The hand consists of a 4-finger manipulator with four motors, 36 exteroceptor and proprioceptor sensors [Matsuoka 95] controlled by an on-palm microcontroller. Cog's head carries four black and white cameras which together form an active vision system, each eye having two degrees of freedom. The system can be best described as consisting of two pairs of cameras, one pair forming one "eye". Cog is also equipped with an auditory perception system [Irie 95]. The control system for Cog consists of a large number of microcontroller boards [Kapogiannis 94].

First of all, Cog cannot move much. For this reason and for the reason of being supposed to be humanoid it will have functional circuits which are very different from the ones we find in biological systems. In physiological psychology, for example, the following items are typically (among others) associated with basic *human* needs [Rosenzweig & Leiman]:

- eating & drinking
- heating/cooling
- biological rhythms
- sex

Obviously, none of these needs, which could easily be turned into functional circuits, makes immediate sense for an artificial humanoid robot. For the development of Cog a different approach is necessary: Given the behaviors which we would like to see in Cog, a set of functional circuits should be described such that trying to develop a robot which follows these circuits will exhibit the desired behaviors.

The following circuits could form an example engagement that Cog autonomously pursues.

touch When something is in Cog's hand and can be easily moved it is taken to Cog's body.

grasp When someone is standing in front of Cog with a toy it tries to grasp the toy object.

reach When someone is near try to elucidate the object.

attract Try to attract a person's attention.

follow Try to follow the person's movements.

strive Look for people.

give A toy in the hand, once it has touched the body, should be given back.

From these circuits things (mark-bearers) can be derived in a straigthforward fashion.

4.1.3 Things

The following things are in the world according to Cog. Descriptions are given from mixed points of view, but mainly according to Cog.

people-sounds Things bearing a people-sound mark should generate search behavior. In human terms this might be practically any sound which is generated in Cog's environment, except for continuous background sounds (radio) or perpetually recurring sounds (clock chimes).

people-image People-images should be followed to find attract-able people (those which look at Cog). In general, every movement in the environment may be regarded as a people-image in Cog's environment.

a-person An attract-able person is one who is likely to be attracted by one of Cog's movements. According to Cog, anything resembling a person looking to Cog or simply near Cog may turn out to be an a-person. The goal of attracting a-persons is to annihilate a-person marks and turn them into r-marks.

r-mark The r-mark is typically worn by a person-object directly in front of Cog. It will generate an immediate elucidation or begging behavior. The r-mark could be a friendly smiling face, a known face, a childish sound ("blu-blu"), etc. The goal is again, to get rid of the r-mark and replace it by the v-toy.

v-toy A v-toy mark is typically carried by toy-like objects in a person's hand. It will generate grasping behavior to annihilate the v-toy.

t-toy A t-toy mostly is a toy-object which can be moved around easily. Eventually, it might turn into a f-toy, which is free to be brought in contact with the body. Otherwise, t-toys should simply be tried to be turned into f-toys by appropriate actions. For example, such an action could consist in trying to free the t-toy by moving it around.

A necessary critical question to be asked in the present context is whether the approach to the design of autonomous systems is distinct from behavior-based robotics. The greatest difference is certainly the role which the functional circuits play in our view. The desired behavior is not simply an emergent phenomenon of different interacting behavior modules which have been designed by a clever engineer. Instead, these behaviors are a means of implementing the desired functional circuits. While behaviors may change, while parameters in such behaviors may adapt over time, the overall functional world of the system will not.

Also, as opposed to what some proponents of embodied AI seem to suggest, it is not enough to add the action information to the knowledge about the objects around us to arrive at an understanding of what these objects really are. In contrast to this view is it necessary to have the whole functional circuit at hand, because it is only the role of the tool in the circuit that allows us to understand what the tool does. In the end, it must be clear that not only the artificial man-made tools around us but the whole of existing objects is nothing but the result of the human functional world, indeed it *is* this world.

The list above shows how functional circuits can support the design of minimal requirements for an autonomous system. From the specification of an overall purpose, via the design of functional circuits we arrived at an (implicit) specification of the sensory systems which are necessary to achieve the overall task of the system. Let us now take a look at non-robotic software systems.

4.2 Autonomous software agents

In principle, the view proposed above can also be used for the design of an autonomous software agent. However, there are a few important differences which lie in the fact that robots are physically embodied systems. This means that the environment for robotic systems can never be fully known. This does not only mean that it is impossible to account for absolutely everything which might happen in the surroundings of a robot. It is even more important that the exact space of this environment is hard to predict. This, however, is a major difference to software agents: It is (at least theoretically) a priori clear what the state-space of the environment of such an agent can be. (This might be slightly objected for systems that explore the internet. The important point for what follows is not that this space is absolutely fixed, rather it is highly constrained and can be relatively easily put into syntactic descriptions. This is not the case for real physical environments.)

The second major difference is that in real-world environments the description (or development) of things (instead of objects) happens because of the limited perceptive abilities of the agent. In formal domains it is much easier to attribute every "perceptual" quality (e.g. a mouse click in the delete box of a mailtool) a specific meaning and purpose. And this is exactly how such systems have been designed to date. Take the development of a mail reading agent as an example. Such an agent is supposed to support a system user by presenting mail in an ordered fashion, by deleting unimportant mail, etc. Usually, the characteristics of such an agent are adaptive and adaptation happens on the basis of positively or negatively reinforced agent actions (see e.g. [Maes 94]). The purely perceptive problem, however, is solved by the system designer who predefines the meaning of certain actions.

However, this is only true for a simplistic view of computer environments. In reality, there are many different (or infinitely many possible) actions that have the same result or a result which is sufficiently satisfying for the user. In the mail example, there are many more ways of deleting mail then clicking the button in a specific mailtool. Moreover, and maybe more importantly, an a priori attribution of meaning to user actions restricts the discovery space of an adaptive agent severly. Real autonomy, however, means that the agent is not restricted to a designer's view of the user's actions. Instead, the autonomous system should be equipped with its own goals, i.e. regard user actions as a part of events that happen in the agent's environment and can become parts of functional circuits.

Another complication consists in the fact that the mouse-click approach is a rather simple scenario as far as the meaning of user actions is concerned. One single user action can often be part of several possible meaningful (user intended) actions. Just consider, how often one types "ls" (or "dir") on a conventional computer system. And although the meaning of this action is, of course, to list a directory on the screen, the intention of the user can be a rather different one ranging from deleting the current screen to finding a lost file, etc. The development of learning methods which discover such contexts that will make the meaning of user actions unique is therefore a major endeavour in current autonomous agent research.

From the viewpoint of functional circuits, however, the user actions are all events that can become triggers for certain behaviors. It is the conformity of the system view that allows the designer to concentrate on the most important aspect of autonomy, namely a system's goals and the motivation to pursue these goals. Therefore, motivation is discussed next.

5 Motivation and Functional Circuits

We have now pointed out why the notion of motivation is of central importance to autonomous systems. It has also been explained, what functional circuits are and how they can help in designing and explaining such systems (artificial or natural). However, the question how these circuits and the motivational issues cooperate, still remains to be shown.

5.1 Motivation and goals

Unlike the simple example we have given above, the usual autonomous system will possess functional circuits for several rather different "engagements". Typical animal examples for these engagements are grooming, mating, feeding, or hunting. Each of these "engagements" in turn may consists of several functional circuits. Whereas the control of functional circuits within one engagement is rather straightforward, the overall control of engagements is more complicated.

One solution to this problem that diffentiates well between functional circuits, motivations, and drives has been given by [Balkenius 95]. The basic problem consists in developing a selection scheme that works well with functional circuits, i.e. behaviors in this case. The version proposed below is a central selection scheme that is driven by a motivational indicator.

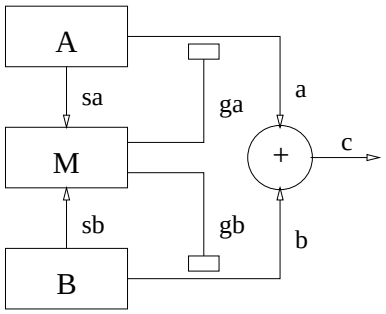

Figure 3: Central selection after [Balkenius 95]. The outputs a and b originate from corresponding behavior modules A and B. The result of the arbitration is c. Strength outputs sa and sb are used to reinforce the motivational state and the gating is performed by signals ga and gb.

In the selection scheme depticted in fig. 3 the gating signals (gating the behavior outputs) are functions of a central motivational state M. This state, in turn depends on the strength output from A and B. The important point in Balkenius' work is that the motivational states themselves are a function of internal and external reinforcements (internal reinforcement as depicted in fig. 3) and of a (set of) *needs*.

In this sense, *motivations* are concerned with what the system should do based on its needs. An example for how needs and motivations could become integrated is given in fig. 4. In this architecture *needs* (or drives) represent the relative importance or urgency of an engagement at a certain time. These needs drive the engagements.

The important aspect of this scheme is that it not only allows the control of a set of behaviors through the use of motivations. It also shows a way to cope with different needs that are coordinated by means of competition among the motivations. Coming back to the design of functional circuits, as we have described it previously, it can be seen that motivations can be used to control such circuits. Motivations here can be regarded as a form of representing a system's goals. On the other hand, drives as they are used in the architecture from fig. 4 are an easy and straightforward way of deciding what to do next, i.e. of

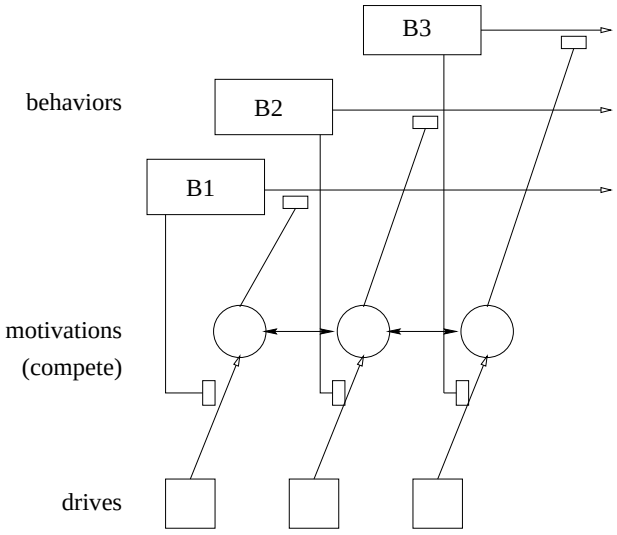

Figure 4: An example for a motivational/drive system, inspired by [Balkenius 95]. The figure depicts the control of several behaviors through competing motivations. The motivations themselves are activated by (non-competing) drives.

implementing a control scheme that negotiates between competing goals. The feedback loops between behaviors and motivations ensure that an engagement is active for a short while and not immediately stopped as soon as another need becomes more important.

Coming back to design issues we can now describe the necessary steps in designing a truly autonomous agent in more detail. Our previous list has been:

1. specification of behavior
2. reformulation in teleological terms
3. design of the motivational system
4. specification of needs and drives

Step 2 means the design of functional circuits as they have been described here. Together with the circuits themselves the things in the agent's environment are developed. This means the construction of the autonomous system's ontology and is further described in the discussion section.

Step 3 correlates with the development of engagements (groups of behaviors) and can be based on multiple and competing goals of a system.

Step 4 is the development of a system that is able to solve the contradictions among competing goals by using needs, drives and the corresponding incentives external and internal to the agent.

The most important turn in perspective is the concentration on functional circuits which support

- a minimalist system development
- and a goal-oriented redescription of the system behavior.

5.2 Emotions

Since emotions are very often discussed in the context of motivation we take a brief look at emotions in the context of the ideas presented here.

5.2.1 What are emotions?

First, a clarification of terms is necessary. Emotions can be/mean

- private subjective feelings,
- display of somatic responses,
- states of actions (e.g. defend, attack, etc.).

As with motivation and goals it must be asked what the meaning of "emotion" can be in the context of functional circuits. This means we have to assign a function to the emotional phenomena.

5.2.2 Emotions as subjective feelings

The notion of "private subjective feelings" comes with many problems due to their relation to consciousness. Several authors (e.g. [Balkenius 95]) have proposed that this inherently private aspect of emotions is associated with a teaching or reinforcement signal. To Balkenius' opinion emotions are concerned with what the system should have done. Emotions are therefore dependent on reinforcement and based on expectations. For example, relief is an emotion which often occurs when an expected negative reinforcement (e.g. punishment) does not take place. Fear, on the other hand, may be triggered by unexpected punishment and produce an incentive for some general avoidance behavior. In this view, the dimensions of emotion are derived from the presentation, omission or termination of reinforcing stimuli.

Since several authors have discussed the control aspect of emotions (see e.g. [Maes 91, O'Rorke & Ortony 94, Sloman 93, Wright 94]), we will concentrate on the *communicative* functions of emotions here and neglect the more general aspect of emotion as an (internal) control system.

5.2.3 Emotion and communication

Some functional circuits associated with a communicative function are

- social behavior (e.g. show compassion),
- (asking for) support,
- cooperation.

In [Maes 94] simple caricatures of faces are used to convey the state of the agent to the user. This is a simple form of how emotions can be used to support communicative functional circuits. In [Maes 94], however, the function of these caricatures is more a conventional control-oriented one, i.e. the user takes advantage of knowing the state of the agent in order to control its behavior. It is worth noting that the expressions implemented in this system are in close correspondance to Balkenius' dimensions of emotions. The reason for this lies in the fact that both systems are used to control learning in the agents. The difference is, however, that Maes' agents are asking the user for help, whereas the emotion in Balkenius' work drives the learning in agents without any social aspects.

In its more original version, the communicative function of emotion must, again, be viewed from within the system itself. An expression of doubt, for example, may reduce the danger of receiving strong negative reinforcement. An expression of joy or compassion, in turn, may result in positive reinforcement. The interesting aspect with respect to the description of functional circuits is that the role of the triggering stimulus is in a sense directly opposite with emotion as it was with external events. The trigger in communicative emotional behavior very often is something *internal* to the agent that can be annihilated (in the sense of chapter 3) by an emotional communicative action.

Again, this turn in viewing the circuit as primordial results in a tool-oriented view of the agent's environment. The observer of the agent's emotional state turns into a tool for the agent. The observer is communicated with *in order to* annihilate the triggering situation.

5.3 Motivation, Emotion, and Reactive Behavior

The motivational and emotional systems which have been presented above all work with control sytems that have traditionally been termed "reactive". In fact, they have been designed so as to coordinate among several reactive behavior systems or engagements. This, of course, comes with a departure from a purely reactive system. Motivation and, even more, emotion contribute to a system behavior which will be less predictable then their purely reactive counterpart.

The deeper reason for this phenomenon lies in the fact that motivational as well as emotional control systems introduce a number of state variables into any system. I.e. these control systems themselves are not purely reactive with respect to signals coming from the agent's environment. On the contrary, emotional systems tend to process information that is only derived from observing internal states of the agent. In the case of motivations, the input can also come from external incentives, e.g. food in front of the agent and thus drive the behavior on a more reactive basis.

Nevertheless, the introduction of emotional control systems should not be seen as a departure from the credo of behavior-oriented design. It seems to be a natural choice for a mechanism which is able to control among competing behaviors. Moreover, such systems become increasingly important with an increase in the agent's adaptivity. In Balkenius' work, for instance, emotions

are primarily used to control motivation and learning. And finally, as we have seen above, emotion plays an important functional role in the communication with other (emotional) agents.

6 Discussion

6.1 Ontology

In this paper I have tried to give an account of the role that motivation plays in the design of an autonomous system. I have introduced the notion of a functional circuit to negotiate between the teleological parts (goals, motivations, and drives) and the more reactive parts (behaviors). I will now discuss how this view changes the system onotology. As I have shown above, the circuits can be used to design what *is* in an autonomous system's world or, in the terms introduced above, what the world *is* according to the system. Traditionally, the science which tries to give a systematic account of *what is* is called "ontology". Ontology deals with the being qua being, i.e. it does not deal with objects as the subject matter of natural science or psychology. Instead, it describes what being is with respect to the fact that it is.

We have described a world which is structured by a humanoid robot's interaction with this world. As we have seen, everything there is in this world mirrors a functional interaction of Cog with its world. The functions, in turn, constitute the system's essential nature in the sense of being the only constants in its behavior. The functional world of the system thus forms the nature of this being and the things in its world reflect this essence of the system.

The ontological position which I have described here is so surprisingly similar to the existential-ontological philosophy of Martin Heidegger [Heidegger 27] that it is worth describing a few points of contact between both ontologies. This seems all the more worthwhile, because one of the most important critics of Artificial Intelligence, Hubert Dreyfus [Dreyfus 72, Dreyfus 91], bases his attacks on (symbolic) AI on the works of Heidegger. With the argumentation in this paper, Cog may result in the first experiment of implementing a Heideggerian ontology although it has not originally been designed that way. It also shows a way to better understanding Heidegger's conception of what the nature of (human) Being is.[3]

Our notion of *things* can be best compared to what Heidegger calls *equipment*. In the human Being's everyday practices things in our world make sense because we can use them.

> We shall call those entities which we encounter in concern *"equipment"*. In our dealings we come across equipment for writing, sewing, working, transportation, measurement. The kind of being which equipment possesses must be exhibited.[4]

[3] [Wheeler 95] has argumented in this direction before, but here I add a robotic example. Also, the approach considered here is not so much concerned with the notion of a "background."

[4] [Heidegger 62, p.97], [Heidegger 27, p.68], quoted after [Dreyfus 91].

The entities so encountered are not objects in the above sense to which we simply add a functional predicate. *Dealing* with them is our primordial way of having them, not some bare perceptual cognition. To paraphrase Heidegger, "hammering" does not know about this property of being a tool. Instead the more we are immediately engaged in coping with the problem of fixing something, the less the hammer is taken as an object which can be used in-order-to hammer [Heidegger 27, p.69]. Strictly speaking, for Heidegger there is nothing like one equipment in this sense, because anything which we are using is embedded in a whole of multiple references to other tools and purposes. The hammer thereby refers to nails, tables, wood, etc. i.e. a whole world of equipment and also of meaningful coping with the world. As long as we are engaged in "hammering", in a purposeful dealing with equipment and this equipment simply is "available", we do not even think about it. In such a situation the tools are simply "ready-at-hand".

We have already seen that this is exactly how Cog sees its world. Obviously, since Cog's way of structuring the environment is inherently tied to functions, every thing will immediately show up as "ready-at-hand", the world will consequently make up one meaningful whole.

> The world presents itself in the equipmental nexus, in the reference to a previously seen whole. [Heidegger 27, p.75, my translation]

The world does not consist of things which are "ready-at-hand", because it is only in situations of breakdown that the equipment can be recognized as one thing primarily identified by its sensory or physical properties. In these situations the things are deprived of being "ready-at-hand", creating mere *occurrentness*.

For Heidegger then, the fact that the world usually does not present itself as a world is the "condition of the possibility of the non-entering of the available from the inconspicuous phenomenal structure of this being-in-itself." [Heidegger 27, p.75, my translation]

This view opposes any tradition which believs that things can be identified with reference to their sensory properties.

Basically this is based on the Cartesian assumption that extension would be essential characteristic of substance.

> [...] that *Descartes* is not merely giving an ontological misconception of the world, but that his interpretation and their basis have lead to *skipping* the phenomenon of the world as well as the being of the [...] innerworldy being. [Heidegger 27, p.95, my translation]

In the end, this is one of the main sources for the problems of traditional approaches to robotics. From the idea that sensory and physical properties would be primordial it follows that a physical theory must be used to decide upon objects encountered in the world. Moreover, such a theoretical approach must be used to find out, whether a table could also be used as a chair.

Any usage of tools, any way of dealing with the world therefore has to be explained with respect to those senory qualities. In a (remotely) existenial-ontological view this problem simply does not arise in this way because dealing

with the things for a specific purpose is the prevailing mode of encountering them, or rather to create them. The argument, therefore, is not that theoretical objects would not exists, but that their properties must remain unaccessible if not taken as the basis, as the primordial source of creating things.

6.2 Conclusion

In this paper, we have presented a new approach to autonomous systems. Beginning with a discussion of the notion of autonomy, which we believe to be central for the development of intelligent agents, we have described the functional circuit as a fundamental building block in agent design methodology. We have outlined how the understanding of functional circuits helps to correctly predict and describe animal behavior and how this analytic procedure can be turned into a synthetic method for the construction of agents. Finally, we have shown how motivation and emotion are in close connection with the notion of a functional circuit and how they form the basis for a new system ontology of agents. In this new ontology, aspects of objects as they are relevant to the autonomous system are important rather then the more conventional properties of things as selected and described by the system developer.

Of course, thus a new difficulty arises, which we have not mentioned so far: How then is it that all these different views of one object become integrated? How is it that the hammer lying over there and the one in my hand become recognized as one hammer-thing?

An answer to this important question could lie in the communicative aspect which we have begun to discuss in the sections on emotions. As soon as there is a need for communication, the need for reference to external objects may arise, too. The next step here would be to describe language as a set of functional circuits. The need for reference to several different aspects of one thing by means of one easily recognizable symbol could mean the urge to develop internal and external signs for different things which thus become objects in our terminology.

Acknowledgments

The author gratefully acknowledges valuable discussions with Rodney Brooks, Polly Pook, and Matthew Williamson.

This research was made possible through support by the Austrian Federal Ministry of Science, Research, and the Arts and the University of Technology, Vienna, Austria.

References

[Balkenius 95] Balkenius C.: Natural Intelligence in Artificial Creatures. Cognitive Studies 37, Lund University, Lund, Sweden, 1995.

[Brooks 89] Brooks R.A.: A Robot That Walks: Emergent Behaviors from a Carefully Evolved Network, AI-Laboratory, Massachusetts Institute of Technology, Cambridge, MA, AI-Memo 1091, 1989.

[Brooks etal. 88] Brooks, R.A.: Connell, J.H., Ning, P.: Herbert: A Second Generation Mobile Robot. AI Memo 1016, MIT Artif. Intell. Laboratory, 1988.

[Brooks & Stein 94] Brooks R.A., Stein, L.A.: Building Brains for Bodies, Autonomous Robots, 1, 7–25, 1994.

[Connell] Connell, J.H. Minimalist Mobile Robotics. A Colony-style Architecture for an Artificial Creature. Academic Press, San Diego, CA, 1990.

[Dreyfus 72] Dreyfus, H.L.: What Computers Can't Do: a critique of artificial reason. Harper & Row, New York, 1972.

[Dreyfus 91] Dreyfus, H.L.: Being-in-the-world. MIT Press, Cambridge, MA, 1991.

[Dress 52] Dress, O.: Untersuchungen über die angeborenen Verhaltensweisen bei Springspinnen (Salticidae). Z.Tierpsychol. 9, 169–207, 1952.

[Heidegger 27] Heidegger, M.: Sein und Zeit. Max Niemeyer Verlag, Tübingen, 1927. (16th ed., 1986.)

[Heidegger 62] Heidegger, M.: Being and Time. Translated by Macquarrie, J. and Robinson, E., Harper & Row, New York, 1962.

[Irie 95] Irie, R.E.: Robust Sound Localization: An Application of an Auditory Perception System for a Humanoid Robot. Master's thesis, MIT AI Lab, Cambridge, MA, 1995.

[Land 72] Land, M.F.: Mechanisms of Orientation and Pattern Recognition by Jumping Spiders (Salticidae). In: Information Processing in the Visual System of Arthropods, R. Wehner (ed.), Springer-Verlag, 1972.

[Kant] Kant, I.: Kritik der reinen Vernunft. *(A critique of pure reason.)* Riga, 1787.

[Kapogiannis 94] Kapogiannis, E.: Design of a large scale MIMD computer. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 1994.

[Maes 91] Maes P.: A Bottom-Up Mechanism for Behavior Selection in an Artificial Creature. In Meyer J.-A. & Wilson S.W.(eds.), From Animals to Animats, A Bradford Book, MIT Press, Cambridge, MA, 238–246, 1991.

[Maes 94] Maes P.: Agent that Reduce Work and Information Overload. In: Communications of the ACM, Vol. 37, Number 7, pp. 31–40, 1994.

[Marjanovic 95] Marjanovic, M.: Learning Maps Between Sensorimotor Systems on a Humanoid Robot. Master's Thesis, MIT AI Lab, Cambridge, MA, 1995.

[Matsuoka 95] Matsuoka, Y.: Embodiment and Manipulation Learning Process for a Humanoid Robot. Master's thesis, MIT AI Lab, Cambridge, MA, 1995.

[Minsky 85] Minsky M.: The Society of the Mind. Simon & Schuster, New York, 1985.

[O'Rorke & Ortony 94] O'Rorke P., Ortony A.: Explaining Emotions, Cognitive Science, 18(2), 283-323, 1994.

[Pratt & Williamson 95] Pratt, G.A., Williamson, M.W.: Series Elastic Actuators. In Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), August 5-9, Pittsburgh, PA. Vol 2, pp. 399–406, 1995.

[Prem 95] Prem E.: Understanding Complex Systems: What Can the Speaking Lion Tell Us? In Steels L.(ed.), The Biology and Technology of Autonomous Agents, Springer, Berlin Heidelberg New York, NATO ASI Series F, Vol. 144, 1995.

[Read & Sloman 93] Read T., Sloman A.: The Terminological Pitfalls of Studying Emotion, Cognitive Science Research Centre, School of Computer Science, Univ. of Birmingham, Birmingham, UK, 1993.

[Rosenzweig 89] Rosenzweig M.R., Leiman A.L., Physiological Psychology. Random House, NY, 2nd Ed., 1989.

[Sloman 93] Sloman A.: The Mind as a Control System. In (Hookway C. and Peterson D., eds.) Proc. 1992 Royal Institute of Philosophy Conf. 'Philosophy and the Cognitive Sciences', Cambridge University Press, 1993, pp. 69–110, 1993.

[Trappl & Petta 95] Trappl R., Petta P.: What Governs Autonomous Agents?, in Magnenat-Thalmann N.M. & Thalmann D.(eds.), Proceedings Computer Animation '95, April 19-21, Geneva, Switzerland, IEEE Computer Society, Los Alamitos, CA, pp. 1–10, 1995.

[Uexküll 28] Uexküll, J. von: Theoretische Biologie. *(Theoretical Biology.)* Suhrkamp, Frankfurt am Main, 1973, (1928).

[Wheeler 95] Wheeler, M.: Escaping from the Cartesian Mind-Set: Heidegger and Artificial Life. In: Morán, F. et al. (ed.), Advances in Artificial Life, LNAI 929, Springer, Berlin, 1995.

[Williamson 94] Williamson, M.W.: Series Elastic Actuators. Master's thesis, MIT AI Lab, Cambridge, MA, 1995.

[Wright 94] Wright I.: An Emotional Agent. The Cognition and Affect Group, Cognitive Science Research Centre, School of Computer Science, Univ. of Birmingham, Birmingham, UK, 1994.