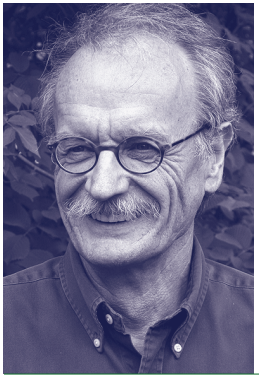

THE TALKING STICK: A COGNITIVE SYSTEM IN A NUTSHELL

HOLK CRUSE

Born in 1942 in Stuttgart, Germany. *Staatsexamen* at the Albert-Ludwig-Universität Freiburg (Biology, Mathematics, Physics) in 1968: dissertation at the University of Stuttgart in 1972; habilitation (Zoology) at the University of Kaiserslautern in 1976; Professor of Biology at the University of Bielefeld, Department of Biological Cybernetics starting in 1981 and continuing to the present. Research stays at the Max Planck Institute for Biological Cybernetics, Tübingen 1971–1973; at the Marine Biological Lab Archachon in 1982; and at the University of California, Berkeley in 1985. Member of the directorial board of the Institute for Interdisciplinary Research (ZIF) at the University of Bielefeld 1989–1997; awarded the Körber Prize in 1993; Fellow of the Wissenschaftskolleg zu Berlin 1995–1996; member of the Excellenz Cluster Cognitive Interaction Technology from 2007 to the present. Research interests: motor control on the reactive and cognitive level, experimental and simulation studies. – Address: Fakultät für Biologie / Biologische Kybernetik, Universität Bielefeld, Postfach 100131, 33501 Bielefeld.
E-mail: holk.cruse@uni-bielefeld.de

During the academic year 2008/09, the Wissenschaftskolleg supported the focus group "Understanding the Brain – an Attempt to Unify Language Production, Reasoning and Motor Control". Its members were Lisa Aziz-Zadeh, Thomas Metzinger, Srinivas Narayanan, Rafael Núñez, Luc Steels, and me. In what follows I will present a brief report on the work we performed during the year.

Matter can assume different states, such as solid, fluid or gaseous. Matter can also aggregate to form living systems. Probably the most miraculous state formed by matter is

cognitive systems. Although there are varying views of what is meant by cognition, most people agree that matter has to adopt a specific structure to allow cognitive properties to arise. There are examples of cognitive systems in specific neuronal systems, but we still only possess preliminary ideas as to how a neural system should be organized in order to allow for cognition.

So, what exactly is meant by cognition? It is said that if you pose this question to six cognitive scientists, you will receive at least seven different answers. In other words, there is no generally accepted definition of the term in the way that there is, for example, in physics. This problem concerns not only the term cognition but more or less all related notions, such as "schema", "attention" or "declarative memory", not to mention "consciousness". But this is not the result of cognitive scientists being less interested in applying rigorous scientific methods; rather, it is owing to the fact that the entire field is still in a comparatively early stage of development. In a rapidly growing research area, the definitions of terms and of problems are to a large extent dependent on individual intuitions that often contain unexplained, implicit assumptions. Moreover, and even more importantly, this situation results from the fact that research covered by what is called cognitive science is conducted in very different disciplinary domains – behavioral biology, computer science, linguistics, neurology, neurophysiology, philosophy of the mind and psychology. Researchers in the different domains may be trying to understand the same phenomenon while yet investigating this phenomenon at different descriptive levels, meaning that they make varying use of similar terminology. The challenge is in finding a link between these diverse levels. But this is a goal not easily achieved, with some researchers even denying that links can be forged at all between, for example, the activation patterns of a neural network and a thought, or a reason, or a subjective experience.

Here we would like to propose a way of not only finding some common ground for the shared concepts used in cognition research but also in helping to understand the mechanisms underlying basic cognitive abilities. To this end, we have attempted to construct an artificial agent consisting of a body with nontrivial morphology (in this case with six 3-jointed legs) and a "brain" comprised of an artificial neural network. All relevant physical and computational properties of this system are known and therefore the causes of specific behavioral properties can be specified. In other words, we are taking to heart Feynman's statement (see Hawkin 2001) that we can only understand that which we are able to construct (see also Vico 1710); or as Kawato (2008) put it, "understanding the brain by creating the brain". If this agent displays certain nontrivial behavior, we may

then be able to compare this behavior with similar conduct observed in animals or human beings. We can then ask whether the concepts we are applying to describe the mechanisms possibly underlying the behavior of a biological system (e. g. attention, mental workspace, procedural and declarative memory, spotlight, metaphor) may also be applied to those of the artificial system. If such an assumption can be confirmed, we will then have found, in the form of this artificial agent, a quantitatively defined hypothesis forming a clear definition of the corresponding concepts.

The structure of the network should be as simple as possible in order to be manageable. We will therefore focus on a system that is first of all able to deal with a specific area of behavior, for instance walking in an unpredictable environment, climbing over large gaps, orientation with respect to landmarks, and the ability to manipulate objects. Furthermore, the agent should be able to plan ahead in order to solve problems for which no solution is actually available, and the agent should be able to understand and produce verbal (in the sense of propositional) expressions that consist of single words or short sentences, what Bickerton (1990) called "protolanguage". As language should be "grounded" in the certain behavior, this language ability mainly focuses on items related to the sphere of motor control.

In order for this approach to make sense we will not be searching for a number of separate solutions to the different problems but rather for one unique system that is capable of coping with all (or as many as possible) of the features defining a cognitive system. In other words, we are making headway toward an autonomous agent.

While traditional A(rtificial)I(ntelligence) attempted to approach cognition by dealing with the abstract manipulation of symbols, in the past two decades views have decisively changed. It is now considered crucial that cognitive abilities be "grounded" in a body and in environmental situations (Steels 2008). Taking an evolutionary view, our assumption is that cognitive systems do not (and cannot) exist per se but always require an embodied "reactive" system as a basis on which to operate. In our attempt to construct a cognitive system, this means that some kind of reactive system is first required – a reactive system capable of controlling certain nontrivial behaviors, including the ability to adapt somewhat to environmental changes.

As mentioned above, an essential aspect of cognition refers to the capacity for planning ahead (McFarland and Boesser 1993). In order to implement this ability, the neuronal system must be equipped with a representation of various parts of the environment. As has been argued with regard to the brain, the body is the most important part of this

environment (Cruse 2003), and so a neural representation of it is the first step to be taken (for details concerning this body-model see Cruse, Schilling 2010).

In the course of this year, we have continued to develop our network based on the reactive controller Walknet (Dürr et al. 2004). It has been augmented by introducing new behavioral modules as well as a body-model. This body-model answers different purposes. In particular, it internally simulates behavior. This simulation may lead to finding a new solution to a problem detected by problem-sensors, and should a solution indeed be found then this behavior will be performed and the corresponding network stored as an element of long-term memory. Specifically, the reactive network is equipped with a new type of winner-take-all net. This network forms the core of the "cognitive expansion", which cannot function by itself but, like a parasite, operates on the reactive system.

Thus far we have been considering an agent that is able to display reactive, or automatic, behavior and that, in addition, is capable of motor planning (*probehandeln*) through its accomplishment of imagined behavior. As a first step, we will be extending the network in order to show the capability of low-level acoustic communication. Two modes have to be distinguished: (i) the capacity for behavioral reactions as a response to verbally administered commands (react-to-command) and (ii) a faculty for verbally reporting on its internal states (i. e. on actual or simulated behavior) as triggered by a verbally posed question (react-to-question). To this end, and inspired by the work of Narayanan (1997), we introduced some basic capabilities related to the understanding of language and its production. How these questions have been approached will be explained by one specific example. In Fig. 1, left side, there is shown a section of the walking controller that contains local networks able, for example, to control the swing or stance movement of a specific leg (see the boxes: swing, stance); on the right side there are other networks that represent the corresponding words (see boxes with "swing" and "stance" in quotation marks), these latter networks suited for both production and an understanding of the word. The semantically corresponding partner networks are coupled in a way that is indicated by the double-headed, dashed arrows. What is the function of this network? If the agent is in the react-to-command mode and the word "swing" is issued, the agent will hear the word and may then activate a swing movement (using the dashed arrow pointing from right to left) and thereby showing that it has understood the meaning of this word. If the word "swing" is issued in the react-to-question mode, the dashed connection in the opposite direction will be activated. As a consequence, when the leg is actually swinging, the wordnet "swing" is activated and the agent will utter the word "swing". This means

that the agent will report on its internal state, whether it be referring to an actual or an internally simulated behavior or not (i. e. during *probehandeln*). By extending these simple "one behavior"-"one word" connections, the network can also treat combinations of words like "left-front-leg".

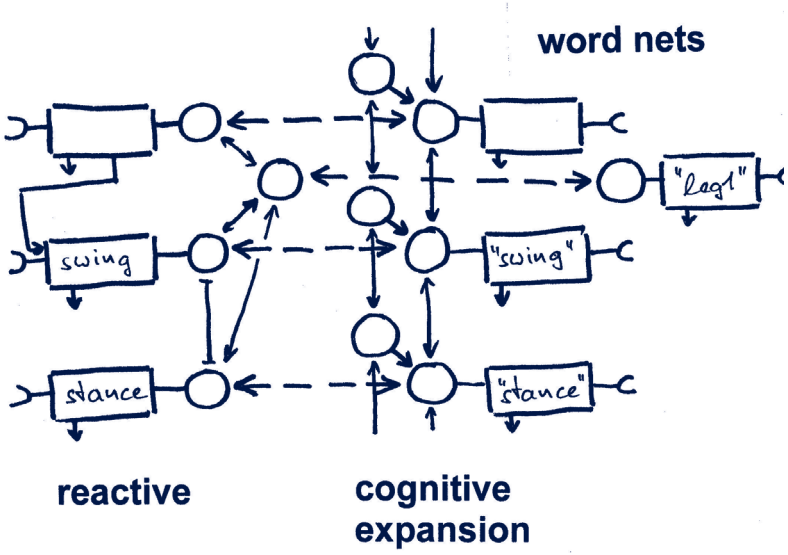

Fig. 1. A section of the network showing the reactive nets (left, e. g. swing), the cognitive expansion, and some wordnets (right, e. g. "swing").

The network as described thus far only displays hardwired connections representing elements of long-term memory. Expanding things yet further, therefore, we have introduced variable connections that allow dynamical and temporal grouping of the memory elements through use of a new version of a winner-take-all network. This coupling of networks may be regarded as a contribution to short-term memory.

As an example of how this grouping might be employed, we introduced the so-called role-units. These units are particularly helpful when it comes to describing dynamic situations. Assume that the agent is watching a red ball moving toward a blue ball, with the red ball finally rolling up against and pushing the blue one. In describing this dynamic

situation, we require more than a mere static representation of the objects: the red ball may be ascribed the role of an agent (actor), and the blue ball that of a patient. Furthermore, there is an action (push), which means that we need role-recognizers (agent, patient) and action-recognizers (e. g. push). Indeed, a net can be trained if, before learning, the three role-units as well as the pair red-blue are connected by mutual inhibition – connections that represent elements of long-term memory.

The availability of a quantitatively defined network able to control the specific behavior of an agent allows for a discussion of the extent to which certain basic concepts developed in behavioral biology, psychology and philosophy of mind might be meaningfully applied to processes running on our network.

In consideration of the complete network, the following characteristics can be ascertained: Using its procedural memory, the net (together with the body) is capable of carrying on various types of reactive or "automatic" behavior. Several of these types of behavior can be performed in parallel. To run these types of automatic behavior, no "central" controller is required to influence the details of the given behavior; rather, the "decisions" result from the dynamics of a self-organizing system.

Independent of the faculty to simultaneously run various types of automatic behavior, the network can select one specific behavior for special treatment. This selection is made when a problem occurs, having been defined by the activation of specific "problem sensors". If we were to describe the function of this system in psychological terms, the "cognitive expansion" (see Fig. 1) might be called an attention-controller that "concentrates on" or "attends" a specific behavior. This focusing mechanism would appear to functionally correspond to what has been described by the "spotlight" metaphor (e. g. Baars and Franklin 2007).

Attention is generally equated with conscious awareness, or is at least considered its necessary prerequisite (Dehaene and Naccache 2001). Hence, the question arises as to what extent the properties of our network might be compared with those of biological networks displaying consciousness. Many authors state that the concept of consciousness is far too complex to be approached on a purely neuronal basis. Other authors have claimed that a network showing the property of consciousness can only be obtained through the use of large-scale systems, whereas the attempt to connect simple networks (like the one being investigated here) with such a high-level concept as consciousness is doomed to failure from the start; however, small-scale networks might well allow for interesting cognitive properties (Menzel and Giurfa 2006). Therefore, we are unable to re-

sist the temptation of discussing to what extent our network may be related to at least some aspects of consciousness.

According to Cleeremans (2005), the lowest level has been termed "Access Consciousness". Access-consciousness refers to the ability of a system to plan and guide actions, to reason, and to report verbally on the content of the corresponding representations; by contrast, unconscious representations cannot be used in this way. Another difference between unconscious and conscious representations is that the latter are characterized as being "globally accessible" or "globally available". This means that many (though probably not all) of the representations stored in memory can become conscious representations. Another aspect is the ability to allow for the communication between different processes, a property of the so-called "unified neural workspace" (Dehaene and Naccache 2001), this being closely related to what Baars and Franklin (2007), at a more abstract level, have called "global workspace". According to these authors, consciousness uses this "neural workspace" in allowing the brain to avoid possibly hazardous actions by simulating them instead.

I should like to argue that most of those properties listed by Cleeremans (2005) as constitutive of access-consciousness can indeed be found in our network. If the agent is engaging in automatic behavior, for example walking on an uneven surface, the behavior can be driven by direct (and therefore rapid) application of local modules belonging to the procedural memory; the attention-network is not activated. According to Cleeremans' definition (2005), these elements are active but unconscious; under specific conditions many of these elements can, in principle, be accessed by attention; therefore, all these modules are "globally accessible". And further properties attributed to consciousness can be found: the network allows for "strategic control" because it can "deliberately" use – or not use – available knowledge (Seth et al. 2008) that can be employed in volitional behavior, i.e. behavior not directly triggered by the environment; or, in the words of Dehaene and Naccache (2001), the network is able "to inhibit the automatic stream of processes and deploy novel strategy". Reportability is a property of our system, too.

Could access-consciousness possibly be localized anywhere in our network? The answer is a resounding no. In our model the neural workspace does not constitute a separate "theater" where the content of the memory elements are re-represented; instead, pre-existing modules of the procedural memory are coupled via the loop through the model of the body and the environment, thus forming a second-order embodiment (see Metzinger 2004) version of global workspace.

In summary, apart from the ability to reason, all requirements for access-consciousness as listed above would appear to be fulfilled. Furthermore, in the course of this year we intensively discussed relations between our network and the properties of transparency, global availability and presence – properties that, according to Metzinger (2009), form necessary and sufficient conditions for a minimalist concept of consciousness. This includes speculation regarding possible connections between certain states of the network and those of phenomenal consciousness.

As a final example, I would like to briefly mention that our network may also be used to provide a specification of the traditional distinction between procedural memory and declarative memory, terms broadly used in the psychology of memory and learning. Procedural memory is generally defined as memory content that cannot be easily verbalized, if at all, whereas the content of declarative memory can indeed be articulated. This definition implies a clear separation between these two types of memories, but to what extent can we apply these definitions to our system? Procedural memory is easily localized. For example, the local networks (at the left-hand side of Fig. 1) that receive sensory input and provide motor output are typical elements of procedural memory. But with respect to declarative memory, the situation is more delicate. According to our system, a procedural memory element may belong to declarative memory if a dotted-line connection is active. Therefore, in order to characterize declarative memory, a more complex network needs to be developed. In addition to procedural memories in the strict sense, the corresponding wordnet and the connections between this wordnet and its procedural partner must also be subsumed under the rubric of declarative memory. In other words, drawing a strict distinction between the two categories would appear to make no real sense.

The members of our focus group are very grateful for the extremely generous support provided by the Wissenschaftskolleg zu Berlin, which has enabled us to make an exciting and promising step forward.

References

Baars, B. J. and S. Franklin (2007). "An Architectural Model of Conscious and Unconscious Brain Functions: Global Workspace Theory and IDA." *Neural Networks* 20: 955–961.

Bickerton, D. (1990). *Language and Species*. Chicago: University of Chicago Press.

Cleeremans, A. (2005). "Computational Correlates of Consciousness." *Progress in Brain Research* 150: 81–98.

Cruse, H. (2003). "The Evolution of Cognition – A Hypothesis." *Cognitive Science* 27: 135–155.

Cruse, H. and M. Schilling (2010). "Getting Cognitive." In *The Neurocognition of Dance*, edited by B. Bläsing, M. Puttke, and T. Schack. Psychology Press (forthcoming).

Dehaene, S. and L. Naccache (2001). "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition* 79: 1–37.

Dürr, V., J. Schmitz and H. Cruse (2004). "Behaviour-based Modelling of Hexapod Locomotion: Linking Biology and Technical Application." *Arthropod Structure & Development* 33: 237–250.

Hawking, S. (2001). *The Universe in a Nutshell*. London: Bantam Press, 83.

Kawato, M. (2008). "From 'Understanding the Brain by Creating the Brain' toward Manipulative Neuroscience." *Philosophical Transactions of the Royal Society* B 363: 2201–2214.

McFarland, D. and T. Bösser (1993). *Intelligent Behavior in Animals and Robots*. Cambridge, Mass.: MIT Press.

Menzel R. and M. Giurfa (2006). "Dimensions of Cognition in an Insect, the Honeybee." *Behavioral and Cognitive Neuroscience Reviews* 5: 24–40.

Metzinger, T. (2004). "Précis of 'Being No One'." *PSYCHE – An Interdisciplinary Journal of Research on Consciousness* 11, 5: 1–35.

Metzinger, T. (2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.

Narayanan, S. (1997). "Talking the Talk is Like Walking the Walk: A Computational Model of Verbal Aspect." *Proceedings of the 19th Cognitive Science Society Conference*. New Jersey: Lawrence Earlbaum Press.

Seth, A. K., Z. Dienes, A. Cleeremans, M. Overgaard, and L. Pessoa (2008). "Measuring Consciousness: Relating Behavioral and Neurophysiological Approaches. *TICS* 12: 314–321.

Steels, L. (2008). "The Symbol Grounding Problem Has Been Solved: So What's Next?" In *Symbols, Embodiment and Meaning*, edited by A. Glenberg, A. Graesser, and M. de Vega, 506–557. Oxford: Oxford University Press.

Vico, Giambattista (1710). "De antiquissima italorum sapientia." In *Opere*, edited by Roberto Parenti. Naples: F. Rossi, 1972.