Mental States as Emergent Properties

From Walking to Consciousness

Holk Cruse & Malte Schilling

In this article we propose a bottom-up approach to higher-level mental states, such as emotions, attention, intention, volition, or consciousness. The idea behind this bottom-up approach is that higher-level properties may arise as emergent properties, i.e., occur without requiring explicit implementation of the phenomenon under examination. Using a neural architecture that shows the abilities of autonomous agents, we want to come up with quantitative hypotheses concerning cognitive mechanisms, i.e., to come up with testable predictions concerning the underlying structure and functioning of an autonomous system that can be tested in a robot-control system.

We do not want to build an artificial system that is, for example, conscious in the first place. On the contrary, we want to construct a system able to control behavior. Only then will this system be used as a tool to test to what extent descriptions of mental phenomena used in psychology or philosophy of mind may be applied to such an artificial system. Originally these phenomena are necessarily defined using verbal formulations that allow for interpreting them differently. A functional definition, in contrast, does not suffer from being ambiguous, because it can be expressed explicitly using mathematical formulations that can be tested, for example, in a quantitative simulation. It is important to note that we are not concerned with the “hard” problem of consciousness, i.e., the subjective aspect of mental phenomena. This approach is possible because, adopting a monist view, we assume that we can circumvent the “hard” problem without losing information concerning the possible function of these phenomena. In other words, we assume that phenomenality is an inherent property of both access consciousness and metacognition (or reflexive consciousness). Following these arguments, we claim that our network does not only show emergent properties on the reactive level; it also shows that mental states, such as emotions, attention, intention, volition, or consciousness can be observed, too. Concerning consciousness, we argue that properties assumed to partially constitute access consciousness are present in our network, including the property of global availability, which means that elements of the procedural memory can be addressed even if they do not belong to the current context. Further expansions are discussed that may allow for the recognition of properties attributed to metacognition or reflexive consciousness.

Keywords
Access consciousness | Consciousness | Emergent properties | Internal body model | Minimal cognitive system | Motor control | ReaCog | Recurrent neural networks | Reflexive consciousness | Robotic architecture

Authors

Holk Cruse
holk.cruse@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Malte Schilling
malte.schilling@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Commentator

Aaron Gutknecht
aaron-gutknecht@gmx.de
Goethe-Universität
Frankfurt a. M., Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

In this article we propose a bottom-up approach to higher-level mental states, such as, for example, consciousness. In contrast to most related approaches, we do not take consciousness as our point of departure, but rather aim, firstly, to construct a system that has basic properties of a reactive system. In a second step, this system will be expanded and will gain cognitive properties in the sense of being able to plan ahead. Only after this work is finished, we ask to what extent this system is equipped with higher-level properties as for example emotions or consciousness. While other approaches would require an exact definition of, for example, consciousness, in our case we do not have to start from a clear-cut definition and try to fit it into a model. We follow this alternative route because there are no generally accepted definitions concerning these higher-level phenomena. In this way we hope to identify the essential elements required to instantiate, for example, consciousness.

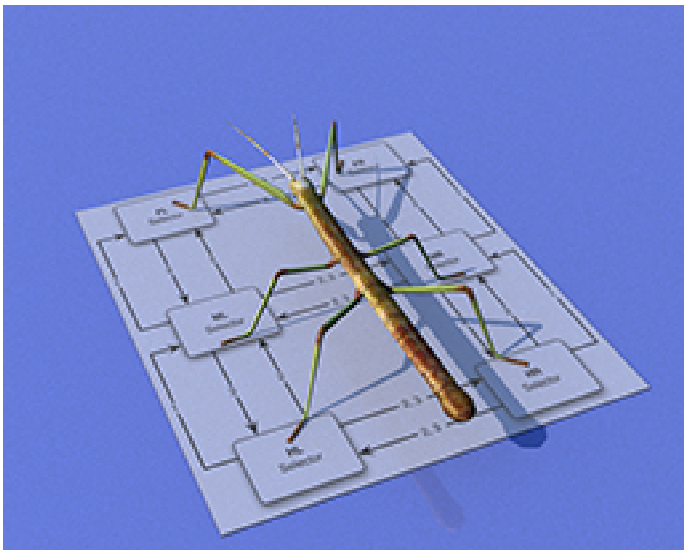

Figure 1: Arrangement of the leg-controllers (boxes: FL front left, ML middle left, HL hind left, FR front right, MR middle right, HL hind right) of the hexapod walker. The arrows show coordinating influences (1–4) that act between neighbouring leg-controllers.

The idea behind this approach is that higher-level properties may arise as emergent properties, i.e., may occur without requiring explicit implementation of the phenomenon under examination but instead arise from the cooperation of lower-level elements. Some authors distinguish between "strong" emergence and "weak" emergence (e.g., Laughlin & Pines 2000). Strong emergence means that there is principally no way to explain the emergent property by known properties of the elements of the system and their coupling. Here we are dealing with weak emergence. In this case, a property recognized when looking at the whole system can at first glance not be traced back (or perhaps only partially) to known properties of the elements and their couplings. Often, auxiliary assumptions are made to explain this property as a global property, i.e., as a property ascribed to the system as a whole. A more detailed inspection may, however, show that such auxiliary assumptions are not required. Instead, the emergent property follows from the properties of the elements and the specific ways in which they causally interact. This insight allows for an understanding of an emergent property in the sense that this property can be predicted, although we may not understand why it arises, and that one is able to construct a new system showing this property.

Following this approach, one crucial decision to be made at the beginning concerns the granularity of the lower-level elements. In our approach, we start from a behavioral perspective and focus on the nervous system as central to the control of action. Therefore, we use neuronal units as the basic elements for our modeling and for the analysis. Specifically, we use artificial neural network units with analogue activation values and dynamic (low-pass filter) properties[1]. That is, our neural elements are qualitatively comparable with non-spiking neurons. Although there are arguments that consciousness, in order to arise, might require synchronously oscillating spikes (Singer & Gray 1995), we claim that the level applied here is general enough to allow for an understanding of such mental processes. As a side effect, this level of abstraction covers different evolutionary groups, such as those represented by insects and mammals, for example. Though much of our discussion, below, focuses on the example of insects, we do not want to argue that insects have all the higher-level properties addressed later in this article, but only that they share the same fundamental functions used in motor control and have, on that level, a comparable structure.

Using these simple neural elements, we start by implementing very basic faculties that include the ability to move one's own (non-

[1] A low-pass filter is qualitatively characterized by an increase of output activation that, when excited by a constant stimulus, asymptotically approaches a given output value.

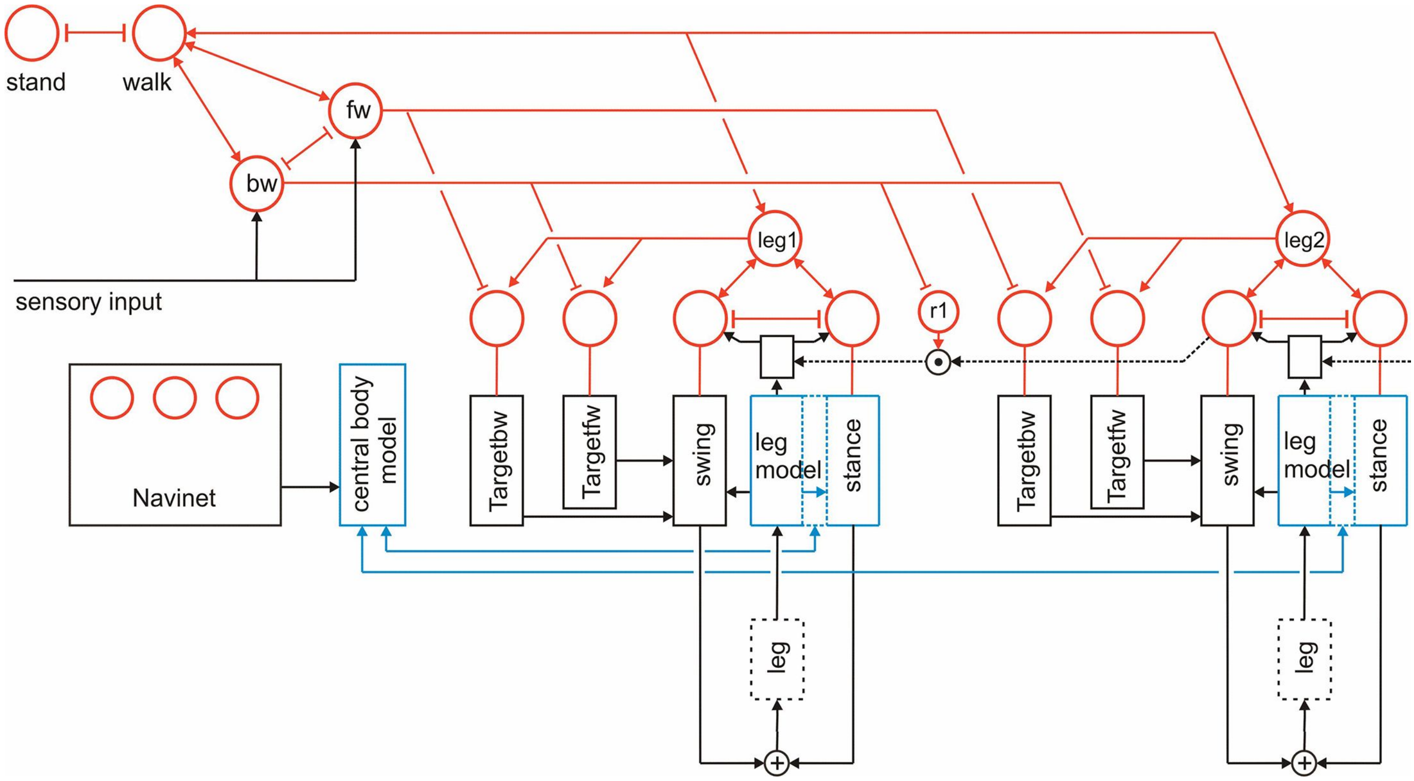

Figure 2: The network controlling the reactive system. Motivation units (red) form an RNN that can assume various attractor states (only two leg-controllers are shown). Arrows show excitatory influences, T-shaped connections show inhibitory influences (fw forward, bw backward, r1 coordination rule 1) The motivation units at the lower margin control procedures (boxes, e.g., swing, stance). The procedures include the internal body model (blue). The body is marked by dashed boxes ("leg"). Indicated here is the network Navinet that controls walking direction (see figure 4 for more details).

trivial[2]) body, and allow for orientation and navigation in an only partially known environment. To this end we use a body with six, insect-like legs. This means that we deal with at least eighteen active degrees of freedom (DoF) and not two—as is the case for many robots that are restricted to moving around on a two-dimensional plane. This means that the controller has to deal with a large number of redundant DoFs. To control the behavior of the robot we use a reactive and embodied neuronal controller, as it is available from earlier work on insect behavior (Schilling et al. 2013a). Later, a minor expansion of the network will allow for cognitive faculties.

What are the properties of the reactive/cognitive system considered here? The reactive system is called "Walknet" and it is based on biological insights from experiments on the walking behavior of stick insects (Dürr et al. 2004; Bläsing 2006; Schilling et al. 2013b). As will be explained in section 2, Walknet was set up as a system for controlling the walking behavior of a six-legged system in an unpredictable environment, e.g., on cluttered terrain or climbing over large gaps—which, when performed in a realistic, natural environment is a non-trivial task. Already on this level we can observe emergent properties. The number of legs on the ground differs depending on the velocity of the walker (for slower walking more legs are on the ground). As a consequence the phase relations between different legs differ depending on the velocity of the walker. Importantly, the resulting stepping patterns ("gaits") are not explicitly encoded in the control network, but are a result of the interaction of the control network with the environment as mediated through the body (1st order embodiment Metzinger 2014). In a further step, the reactive

[2] I.e., a body with redundant degrees of freedom arranged in both parallel and serial order.

controller is expanded to be able to deal with navigation tasks. This additional network, called “Navinet”, is able to simulate a number of experimental results observed in desert ants and honeybees, such as the capability of finding food sources using path integration and orientation with respect to visual landmarks.

Both networks are characterized by their decentralized nature. These networks consist of procedural, (reactive) elements, namely small neural networks that in general connect sensory input with motor output, thereby forming the procedural memory. Inspired by (Maes 1991), these procedural elements are coupled via a “motivation unit network”, a recurrent neural network (RNN) that forms the backbone of the complete system. This type of architecture has been termed MUBCA (for Motivation Unit Based Columnar Architecture (MUBCA), Schilling et al. 2013b). The motivation unit network allows for selection of different behaviors by adopting different attractor states, where each attractor represents a group of motivation units being activated, which in turn control the procedural elements. As the different groups do in part overlap, albeit in different ways, the network allows for the representation of a heterarchical structure (e.g., see left upper part of figure 2).

As a next “evolutionary” step, this reactive network will be expanded to be able to embrace cognitive properties (sects. 3 and 6). The notion of cognition is often used in a broad and sometimes unspecific way. In the following we will rely on the definition given by McFarland & Bösser (1993) who assume that *a cognitive system is characterized by the capability of planning ahead*. We prefer this clear-cut definition of cognition compared to many others found in the literature, as the latter are generally quite weak (in extreme cases cognitive properties are even attributed to bacteria, which, in our view, would make the term cognition meaningless). While such a specific definition might seem too narrow, in our understanding it captures the essence of cognition. Focusing on planning ahead being realized by mental simulation (Hesslow 2002) allows extending this notion of cognition to easily include other high-level phenomena, while still relying on the same internal mechanism. Therefore, in this article, apart from section 10.3 (Metacognition) we will use the term cognition in the strict sense as proposed by McFarland & Bösser (1993).

Being able to plan ahead implies the capability of being able to internally simulate behavior, which basically means to be able to simulate movements of one’s own body within a given environment. This faculty requires, as a first step, the availability of a flexible, “manipulable” internal body-model. Planning ahead is interesting in a situation where the actually carried out reactive behavior cannot reach the currently pending goal. Therefore, a further expansion is required that allows for the invention of new behaviors. Together with the faculty of planning ahead, the system can then test newly-invented behaviors by applying internal simulation (“internal trial-and-error”) in order to find a solution for novel problems for which no solution is currently known to the system.[3]

This system, called “reaCog”, represents a basic version of a cognitive system in the strict sense intended by McFarland & Bösser (1993). As such, cognitive expansion does not function by itself, but only, like a parasite, on top of the reactive structures—a view that has been supported for a long time (Norman & Shallice 1986). The cognitive system depends on its reactive basis (therefore it is called reaCog). Therefore, the evolution of cognitive abilities crucially requires a correspondingly rich (procedural) memory.

In order to increase the richness of the memory of the complete system, in section 5 we introduce perceptual memory and complete the system by implementing “Word-nets”, a specific form of procedural and perceptual memory. In this way, the whole system is equipped with aspects of semantic memory, and can be claimed to represent a minimal cognitive system. We do not deal with learning but only discuss the properties of the finished network. The learning

[3] Note that the term simulation is used here in two different ways. “Internal simulation” enables the agent to simulate behaviors internally, i.e. without actually performing them in reality. Simulation of an animal addresses the construction of an artificial agent. The agent may take the form of a software simulation or a hardware simulation (i.e., a physical robot).

of some aspects has, however, been treated earlier (Hoinville et al. 2012; Cruse & Schilling 2010a).

After having introduced reaCog in sections 2–6, we will, in sections 7–11, discuss how more abstract functions, such as those described in, e.g., psychology, can be based on such a simply-structured network.

A fundamental problem when aiming for an understanding of phenomena like emotions or consciousness concerns the phenomenal aspect. The phenomenal aspect, often characterized as the hard problem (Chalmers 1997), refers to the strange, unexplainable phenomenon that physical systems, in our case represented by specific dynamics of neuronal structures, can be accompanied by subjective experience. Basic examples are experiencing pain, a color, or the internal state of an emotion (e.g., joy, fear). In section 7 we discuss this aspect in some detail and postulate that phenomenality is an emergent property. As mentioned, we are not aiming to solve the "hard" problem (Chalmers 1997), but we argue that it is sufficient to concentrate on the functional aspect.

In particular, we focus on the phenomena of emotions and consciousness. According to a number of authors (e.g., Valdez & Mehrabian 1994), these are assumed to be an inherent property for some cognitive systems. Therefore, although we do not want to state that emotions (section 8), attention, volition, intention (section 9), and consciousness (section 10) should necessarily be attributed to our system in any sense, we want to discuss to what extent properties characterized by different levels of description can be observed in our model.

Considering emotions, these are defined on different levels in the literature, so that there is no clear, generally accepted distinction between concepts like emotions, moods, motivations, drives, etc., which appear to form a continuum of overlapping, not clearly separable concepts (Pérez et al. 2012). Focusing on selected examples, in section 8 we will show how these phenomena may be attributed to our system, for example by referring to basic emotions as proposed by Ekman (1999).

Concerning consciousness, as discussed by Cleeremans (2005), this phenomenon should be approached by differentiating different aspects and treating those aspects separately. To this end, following Block (1995, 2001), Cleeremans (2005), introduces a distinction between access consciousness, metacognition, and phenomenal consciousness. In sections 10.1 (access consciousness) and 10.3 (metacognition) we will focus on whether and how the presented model can be related to the different aspects that are described by Cleeremans (2005), such as access consciousness and metacognition. From our point of view the simple control structure presented does fulfill some aspects of both access consciousness and metacognition. We shall finish with discussion and conclusion in sects. 11, 12.[4]

2 Walknet

ReaCog is an expansion of a control system that has been realized as a neural network. The underlying system has been termed Walknet. Walknet is biologically inspired and is supposed to describe the results of many behavioral studies on the walking behavior of stick insects (Dürr et al. 2004; Schilling et al. 2013b). We will briefly sketch the properties of the network as far as is required for understanding the basic abilities considered here.

Overall, the controller has to deal with the difficult task of coordinating multiple degrees of freedom; in the case of the hexapod walker the body consists of twenty-two DoF. There are three DoF for each of the six legs and an additional four DoF are present in between the body segments. The system is redundant, as only six DoFs are needed to define a position and orientation in three-dimensional space. The controller therefore has to to deal with sixteen extra DoFs. The architecture of the Walknet controller is decentral. Each leg has an individual and more or less independent controller that decides which action to choose (two such leg-controllers are shown in figure 2, the black boxes in the lower part). A single leg

[4] This article comprises an essential extension of an earlier paper (Cruse & Schilling 2013).

controller consists of several procedures. In the figure, each procedure is represented as a single black box. In the basic system, the two important behaviors a leg can perform are the swing and stance movement. The procedures themselves are realized as artificial RNN. Examples are the two basic procedures: the “Swing-net”, which controls the swing movement, and the “Stance-net”, which controls the stance movement of the leg. Only two of the six leg-controllers are shown. These networks constitute the procedural memory of the system. The procedural modules receive direct sensory input and provide motor control commands as an output. But there are also modules that provide input to another module. The controller on the leg level determines which procedure should be actived at any given time, depending on the current state of the leg (swing or stance), as well as on sensory inputs (ground contact, position). In addition, controllers of neighboring legs can influence each other through a small number of connections between those controllers. These influences are explicitly derived from experiments on the coordination of legs in walking experiments on the stick insect.

As was found in the insects, during the swing movement (protraction) the legs aim towards a position at the front, close to the position of the anterior leg. Therefore, each leg possesses a so-called “target net” in order to produce these targeted movements. During forward walking the so-called “Target_fw-net” is responsible for this targeting. During backward walking “Target_bw-net” is used. Both directly influence the Swing-net. Procedures marked as blue boxes (“body model”, “leg model”) will be explained below (section 3.1).

ReaCog is expanded by an RNN, which consists of motivation units (figure 2, marked in red). This network allows the system to autonomously select one of the different possible behaviors. For example, the system may choose between forward or backward walking, or standing. A motivation unit is an artificial neuron with linear summation input and piecewise linear activation function, showing output values from zero to one. Applied to a procedure, for example Swing-net, a motivation unit determines the strength of the output of the corresponding procedural network (in a multiplicative way). As mentioned above, motivation units form a recurrent neural network and can influence each other through excitatory or inhibitory connections (as shown in figure 2).

In addition, there are sensory units that are part of this RNN and that can directly influence the motivation units’ activation, e.g., as shown in figure 2 for the “lower-level” units for Swing and Stance. There, an active ground-contact sensor of a leg reinforces the stance motivation unit for this leg. As the motivation unit network can be arbitrarily expanded, it allows to control of complex behaviors. To illustrate a small group of behaviors only, units as “walk”, “fw” (forward), “bw” (backward), “leg1” are depicted (for more examples see Schilling et al. 2013b; Cruse & Wehner 2011).

The network of motivation and sensory units does not have to form a simple, tree-like structure (see figure 2). It can constitute a heterarchy. Motivation units can be bi-directionally connected through positive (arrowheads) and negative (T-shaped heads) connections. As shown in the figure, this can lead to cycles. There are also different overlapping subnetworks, e.g., the “leg” units as well as the motivation unit for “walk” are active during backward and forward walking. But only one unit indicating the direction of walking can be active at any given time, i.e. either the unit “fw” or “bw” can be active. As a consequence, there are multiple stable attractor states formed through the combinations of excitatory and inhibitory connections. The stable “internal states” stabilize the behavior of the overall control system, as the system cannot be easily disturbed solely through inappropriate sensory inputs. For example, sensory inputs are treated differently depending on the current state (swing or stance) of the control system, and these internal states can be differentiated on a higher-level, e.g., into walking, standing, or feeding (for details see Schilling et al. 2013a; Schilling et al. 2013b).

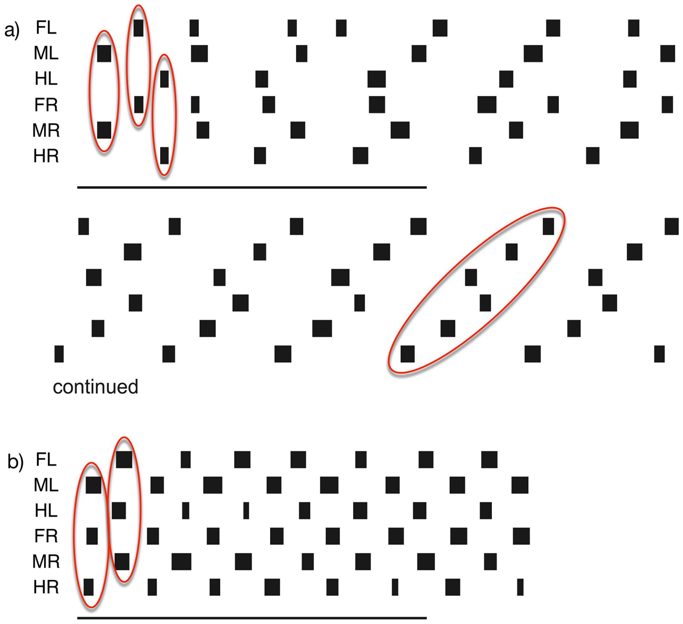

Figure 3: Step pattern arising from the decentralized leg-controllers connected by local rules and the environment. Abscissa is time; black bars indicate swing movement; the gaps represent stance movement of this leg (from top to bottom: front left leg (FL), middle left leg (ML), hind left leg (HL), correspondingly front right leg (FR), middle right leg (MR) and hind right leg (HR) for the right side). The lower bars indicate 500 iterations corresponding to 5s real time. These "foot-fall patterns" show various locally or globally stable patterns depending on walking velocity (a: slow, b: fast) and of starting position. In (a) the legs start with an "uncomfortable" leg configuration leading to a gallop-like pattern (indicated by the vertical ellipses) that after about six steps changes to the globally stable pattern, typical for slow insect walking (see inclined ellipses, step # 8). (b) shows fast walking leading to a tripod gait characterized by synchronous swing movements of ML, FR, HR and FL, HL, MR (see vertical ellipses).

For an RNN, maintaining a stable state is a non-trivial problem, in particular, when there are various disturbances. To illustrate the adaptability and at the same time the stability of the behavior controlled by such a motivation unit network, in figure 3 we show two cases of hexapod walking. Figure 3a shows an example of a slow walking speed where the legs begin from a difficult starting configuration (both front legs, both middle legs and both hind legs start from the same position, which is opposite to the coordination found in normal walking, where opposite legs alternate). Nonetheless, the agent is able to walk. After some steps, the agent reaches a temporally stable pattern corresponding to normal walking. Figure 3b shows a step pattern corresponding to high-speed walking, often termed "tripod gait". Although usually considered to be a regular pattern, detailed inspection shows that there are local temporal variations, but the overall pattern remains stable (for videos of further walking examples see Schilling et al. 2013b). It is important to note that none of these step-patterns are explicitly implemented, but arise as emergent properties (for details see Schilling et al. 2013a). As another impressive emergent property, Bläsing (2006) showed that, with some minor extensions, this walker is able to climb over large obstacles (which can be more than twice the normal step-width).

3 Internal representation

In addition to using the loop through the environment itself, some form of internalization is a prerequisite for any kind of planning. Therefore, specific internal representations[5] are necessary for a cognitive system. This is well in line with the embodied perspective, because from an evolutionary point of view internal models are not at first disconnectable from a very specific function, and they work in service of a specific behavior (Glenberg 1997). Internal models have, in this sense, co-evolved with behavior (Steels 2003). An early representation is the representation of one's own body, and such a representation becomes meaningful early on, in simple control tasks like targeted movements or sensor fusion.

3.1 Body model

In reaCog we introduced an internal model of the body. This model is realized as an RNN (Schilling 2011) and has a modular structure (Schilling & Cruse 2007; Schilling et al. 2012). The overall model consists of two different

[5] The term representation is used here in the broad sense of Steels (1995) "physical structures (for example electro-chemical states) which have correlations with aspects of the environment".

levels. On the top level the whole body and the structure of the insect are represented in an abstract way. Only on the lower level are the details filled in. The lower level consists of six leg networks. Here, for each leg the functional structure of the joints and the limb is captured. In this way this level of representation can be used for motor control and provides detailed information about joint movements. On the higher level, the structure of the body and the legs is represented in an abstract form, i.e., only the footholds of the legs appear on this level. Figure 2 shows the different parts of the body model (drawn in blue). The body model is modular. It comprises a holistic system that is realized as an RNN (figure 5, see Schilling 2011; Schilling et al. 2012 for details).

The body model is used during normal walking, meaning that the system is still in the reactive mode, in forward as well as backward walking or when negotiating curves. It coordinates the movement of the joints and delivers the appropriate control signals for the Stance-networks. As explained above, overall the system is redundant, with twenty-two DoFs in the whole body structure, and this makes deriving consistent control signals for all the joints a difficult problem that can't be computed directly, but rather requires application of additional criteria (e.g., for optimizing energy consumption). In our approach, which uses the internal body model, we employ the passive motion paradigm (von Kleist 1810; Mussa-Ivaldi et al. 1988; Loeb 2001). Consider the body model as a simulated puppet of the body (figure 5) that is pulled by its head in the direction of the goal (figure 5b, pull_fw). This information on the target direction could be provided by sensory input, e.g., from the antennae or vision, in the form of a target vector (figure 2, sensory input). When pulled in this direction, the whole model should take up this movement and therefore the individual legs currently in stance should follow the movement in an appropriate way. The induced changes in the joints can be read out and applied as motor commands in order to control the real joints. In backward or curved walking, the body model has only to be pulled into a corresponding direction (in backward walking using the vector attached to the back of the body model, pull_bw (figure 5b). In this way we obtain an easy solution to the inverse kinematic problem as the body-model represents the kinematical constraints of the body of the walker. It restrains the possible movements of the individual joints through these constraints, and only allows possible solutions for the legs standing on the ground, thereby providing coordinated movements in all the involved joints.

The body-model is also connected to the sensors of the walking system and integrates the incoming sensory information into the currently-assumed state of the body as represented in the body-model. In this way the body-model is able to correct noisy or incorrect sensory data (Schilling & Cruse 2012). Overall, the main task of the body model is pattern completion. It uses the current state and incoming sensory data to come up with the most likely state of the body that fulfils the encoded kinematic constraints. In this way, the model can also be used as a forward-model, meaning that, given specific joint configuration, the model can predict the three-dimensional arrangement of the body, for example the position of the leg tips. The predictive nature of the model is crucial as it allows exploiting the model for planning ahead (see below). It is important to note that while we do not want to claim the existence of such a model in insects, the functions of internal models are prediction, inverse function, and sensor fusion, and these can all already be found in insects.

3.2 Representation of the environment

Of course, internal representation should also contain information on the surroundings. We started with a focus on the body and want to extend this network in a way that reflects how the environment affords (Gibson 1979) itself to the body, i.e., a focus on interaction with the environment.

As an example of how the reaCog architecture could be extended to include representation of meaningful parts of the environment, we want to briefly sketch an expansion of Walknet that would allow for insect-like navigation ("Navinet" Cruse & Wehner 2011; Hoinville et

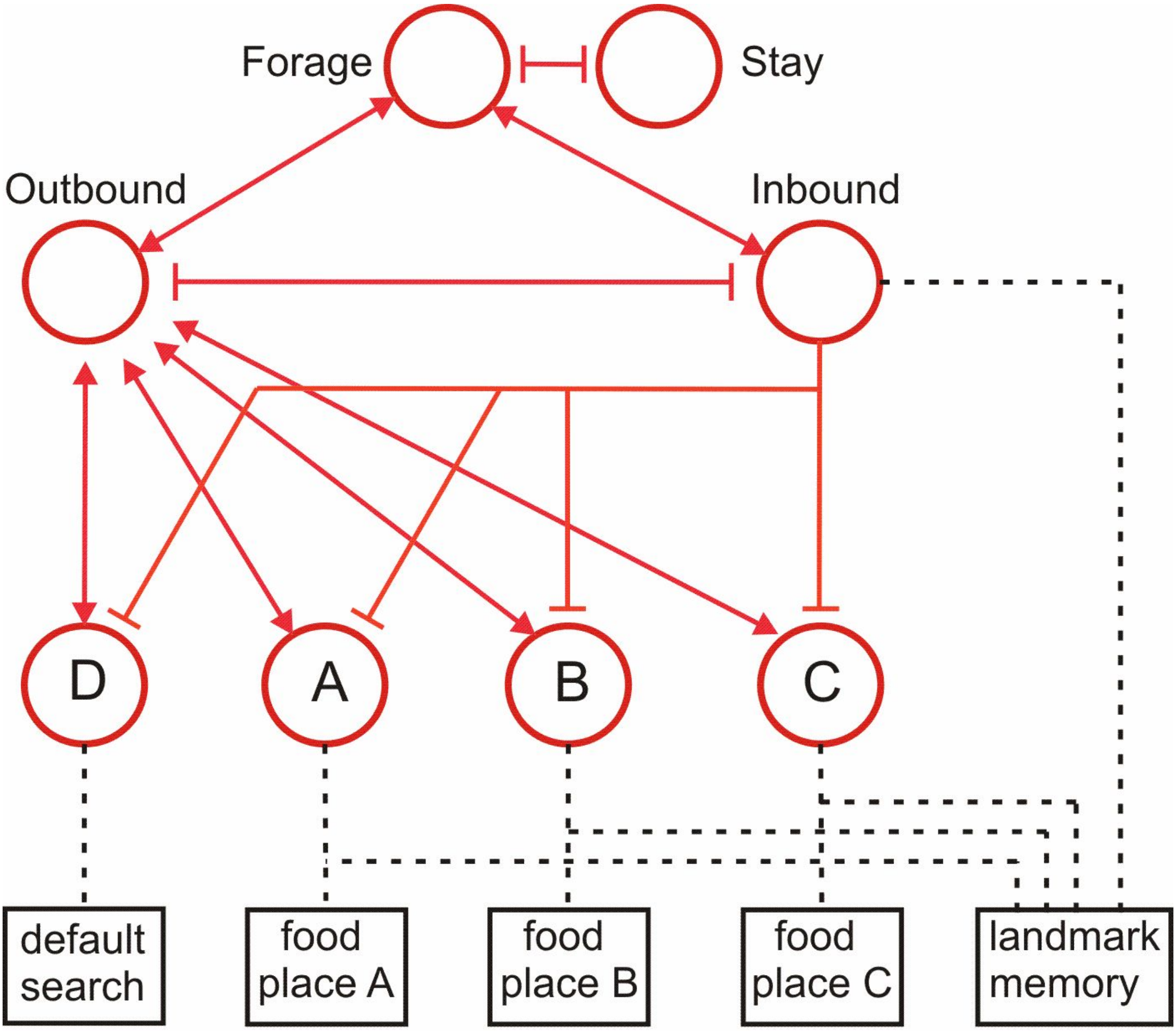

Figure 4: Motivation unit network of Navinet for the control of ant-like navigation. Unit Outbound controls travel from the home to a food source (A, B, C) or a default search for a new source (D). Unit Inbound controls travel back to the home. Memory elements (black boxes) contain position and quality of the food source (A, B, C) or information on visual landmarks (landmark memory).

al. 2012). Navinet provides an output that will be used by the body-model explained above to guide walking direction. Due to the network, the agent can make an informed decision about which learned food source she will visit (e.g., sources A, B or C), or if she is travelling back home or not (Outbound, Inbound, respectively). The output of Navinet is, in this way, on the one hand tightly coupled to the control of walking and the representation of the body. On the other hand, Navinet is constructed using motivation units in the same way as the walking controller, and those motivation units take part in the action-selection process. Importantly, Navinet (like desert ants) shows the capability of selective attention, since it is context dependent and only responds to learned visual landmarks in the appropriate context, i.e., when related to the current active target food source. The structure of the motivation-unit network is sketched in figure 4. Examples of possible stable internal states are (Forage – Outbound – source A – landmarks associated with source A) or (Inbound – landmarks associated with Inbound), for instance. As an interesting emergent property, Navinet does not presuppose an explicit "cognitive map". Such a map-like representation has been assumed necessary by several other authors (Cruse & Wehner 2011). How learning of food source positions and food quality is possible has been shown by Hoinville et al. (2012).

4 Planning ahead, cognition

Even though Walknet is set up as a fixed structure consisting of hard-wired connections of the RNN, it can flexibly adapt to disturbances in the environment as needed during, for instance, crossing large gaps (Bläsing 2006). Nonetheless, the system might of course run into novel situations that require an even higher degree of adaption, and as such will require novel behaviors. As an example, think of a situation in which all the legs except the right hind leg are in the anterior part of the working range. When the right hind leg is forced to lift from the ground as it approaches a position very far to the rear, the whole system will become unstable, as the center of gravity is positioned very far towards the rear of the animal. In this case, the center of gravity would not be supported by the other legs, nor by the right hind leg that tries to start a swing movement. As a consequence, the agent would fall over, backwards. This problem could be detected by "problem detectors", e.g., specific sensory input that reacts to the specific load distribution (a different solution is explained in section 8). In order to overcome this problem, the system would have to break out of its usual pattern of behavioral selection and try to select a different behavioral module that is usually not applicable in the given context. For instance, making a step backward with the right middle leg would be a possible solution, as this would provide support for the body and would afterwards allow going back to the normal walking behavior and the subsequent swing movement of the right hind leg. Usually, backward steps can only be selected in the context of backward walking.

Figure 6 shows an expansion that allows the system to search for solutions that are not connected to the current context. This expansion is termed the "attention controller". We introduce a third layer of units (figure 6, in green), that is essentially a recurrent winner-take-all network (WTA-net). For each motivation unit there is a corresponding partner unit in this WTA-network. Currently-active motivation units suppress their winner-take-all (WTA) partner units (T-shaped connections in figure 6). Therefore, a random activation of this WTA-net will lead to the activation of one single unit not belonging to the currently- activated context. The random activation will be induced by another parallel layer, the "Spreading Activation Layer" (not depicted in figure 6, further details are described in (Schilling & Cruse submitted). The winning unit of the WTA layer than activates its corresponding motivation unit. This triggers the connected behavior that can be tested as a solution to the problem at hand. The network follows a trial-and-error strategy as observed in, e.g., insects.

As has been proposed (Schilling & Cruse 2008), a further expansion of the system that is, most probably, not given in insects is not the testing of a behavior in reality, but instead the application of a newly-selected behavior on the body-model and the use of the model instead of the real body. The motor output is routed to the body-model instead of to the real body, and the real body is decoupled from the control system while testing new behaviors. Due to the predictive nature of the body-model, it can be used to predict possible consequences and to afterwards decide if a behavior solves the current problem and should be tried out on the real body. This procedure is called internal simulation and requires the introduction of switches that reroute motor output signals from the real body to the body model (figure 6, switch SW). Only after a successful internal simulation will the behavior be applied to the real body. McFarland & Bösser (1993) defined a cognitive system as a system that has the ability of planning ahead, i.e., that is able to perform internal simulations in order to predict possible outcomes of behaviors. Therefore, this latter expansion would make the control system cognitive (for details see Cruse & Schilling 2010b).

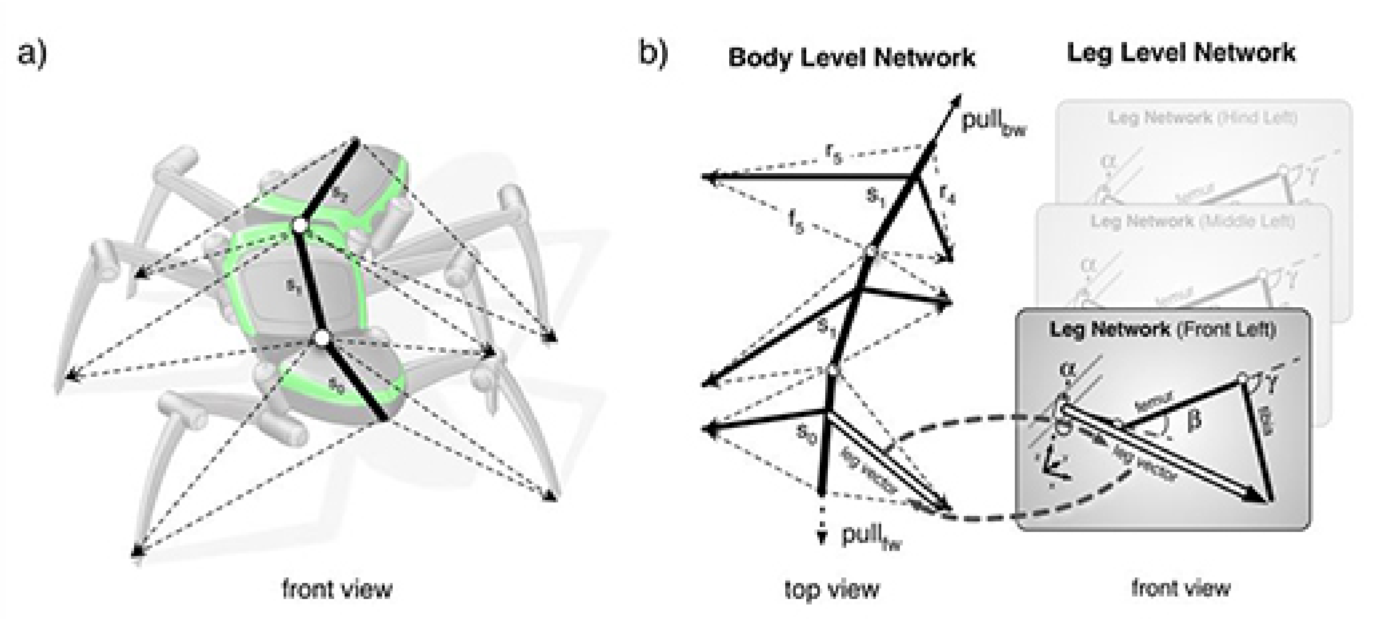

Figure 5: The body-model and its relation to the body of robot Hector (a). (b) shows the vectors forming the central body (left) and the vectors forming one leg model (right). The central model and the leg-models are connected via the shared “leg vector” (white arrows) that point from the hip to the tip of the leg (shown here for the left front leg only). Walking direction and velocity are controlled by the input vectors pull_fw (forward) or pull_bw (backward) provided by sensory input.

5 Word-net and perceptual memory

In our network, we have up to this point only dealt with procedural memories, i.e., memories representing the connections between specific sensorimotor elements that are able to control specific behaviors (e.g., Swingnet, landmark). As a final extension, we will now show how the network might also be equipped with some aspect of semantic memory, such that meaning can be attributed to verbal expressions. To this end, the network can be expanded through the introduction of another layer (not shown in figure 6). In this fourth layer, verbal expressions are stored as procedures or “Word-nets”. These procedures can either be used to pronounce a stored word or to comprehend it, i.e., they can be used for motor control and for auditory perception. As is the case for other procedures, each Word-net is equipped with a motivation unit. As the motivation units of Word-nets have a specific function, for an easier distinction we will call them word units (WU). Following Steels (2007; Steels & Belpaeme 2005) each Word-net is related to a corresponding unit of the motivation network that carries meaning (e.g., the motivation unit for walking is connected to a Word-net “walk”). The meaning of the Word-nets is in this way grounded in the behaviors of the corresponding motivation units. As an example, figure 7 shows a possible detail of such a network, including some elements of Walknet and Navinet. The motivation units of a procedure (e.g., Swing net) and its corresponding Word-net (e.g., “Swing”) are coupled via bidirectional connections (dashed double-headed arrows). The connections cannot be active at the same time, but depend on an overall state of the network, termed “Report” and “Perceive”. In the Perceive state, only connections from the word unit to the motivation unit of its non-word procedure can be activated (from top to bottom in figure 7), whereas in the Report state only the opposite connections can be activated. As can be seen in figure 7, Word-nets can not only be connected with motivation units of the sensorimotor nets, but also with motivation units that do not directly control a sensorimotor element (e.g., Walk, Outbound).

What might be the function of this extension by Word-nets? In the Perceive state (or react state), a perceived word, uttered by another agent, will activate, via its word unit, its partner's motivation unit, and thereby possibly influence behavior (depending on the actual internal state of the system and on the strength of the word input). When in the Report state, the actually active motivation units will in turn activate their corresponding word units, which may lead to an uttering of a word. As, of course, only one word can be activated at a given time, some kind of decision network (e.g., a WTA net) is required, though, for reasons of simplicity, not shown in figure 7. In any case, introduction of Word-nets allows for a very basic form of communication between the agent and any other partner, communication being limited to "one-word sentences".

As indicated on the left side of figure 7 (units "front", "left"), further motivation units might be introduced into the network that do not have a direct function within, in this case, the Walknet controller. Of course, these units may be connected to word units. (Note that we do not deal with the question how these units may be connected within the network through training).

This architecture combines sensorimotor procedures with Word-nets (which by themselves represent specific sensorimotor procedures). Together, they form a simple case of semantic memory, because procedural memory representing an action (e.g., Swing-net) is connected with a memory element representing verbal symbols.

To illustrate the versatility of this architecture, we will briefly address how it can also be applied in order to embrace perceptual memory. Following ideas of O'Connor et al. (2009), Cruse & Schilling (2010a) have shown how an RNN, using the same elements as applied here for the motivation unit network, could be used to construct a perceptual memory. This network does not only allow the representation of directly perceived perceptual elements (e.g., the colour or shape of an object), but also of superordinate concepts (e.g., Cow, Animal, four-legged). Note that "four-legged" might also be a feature of non-animals, e.g., a table. Therefore, the ability of our network to deal with heterarchical structures is advantageous for perceptual memory, too. Elements of such a distributed memory can also be connected to specific Word-nets (e.g., "red", "Cow", "animal"), as has been explained above for the sensorimotor motivation units. Correspondingly, activation of one memory element of this perceptual memory may elicit the uttering of the corresponding word, and, in turn, when in Perceive mode, the hearing of a word may activate various elements of the procedural memory that are associated with this word.

6 ReaCog: Emergent properties characterized by applying other levels of description

To summarise, the neural controller Walknet, (for details see Dürr et al. 2004; Schilling et al. 2013a) is an embodied control system (first-order embodiment, cf. Metzinger (2006, 2014). The reactive system can deal with varying unpredictable environments. It relies only on information that is available to the given mechanosensors, which is possible because both body and environment are integral to the overall computational system. In this way, the system is embodied. Of course, the system has a physical body, but even more, being embodied means that properties of the body (like its geometry) are exploited in computations of the controller. Using its own body as part of a loop through the world allows for dramatically simplifying computations (Schmitz et al. 2008). These properties are of course also present in the expanded version, reaCog. Even though in reaCog an internal body-model is introduced in order to control the high number of DoFs, reaCog still relies heavily on the cooperation of individual parts, i.e., the combination of couplings between body, environment, the internal body model, and the controller itself. In addition, this internal model of its own body is used for planning ahead. Such a network, following Metzinger (2006, 2014) represents a system that is characterized by second-order embodiment.

As shown in figure 2, the procedures forming the decentralized controller are basically arranged in parallel, i.e., each procedure obtains its own sensory input and provides a specific

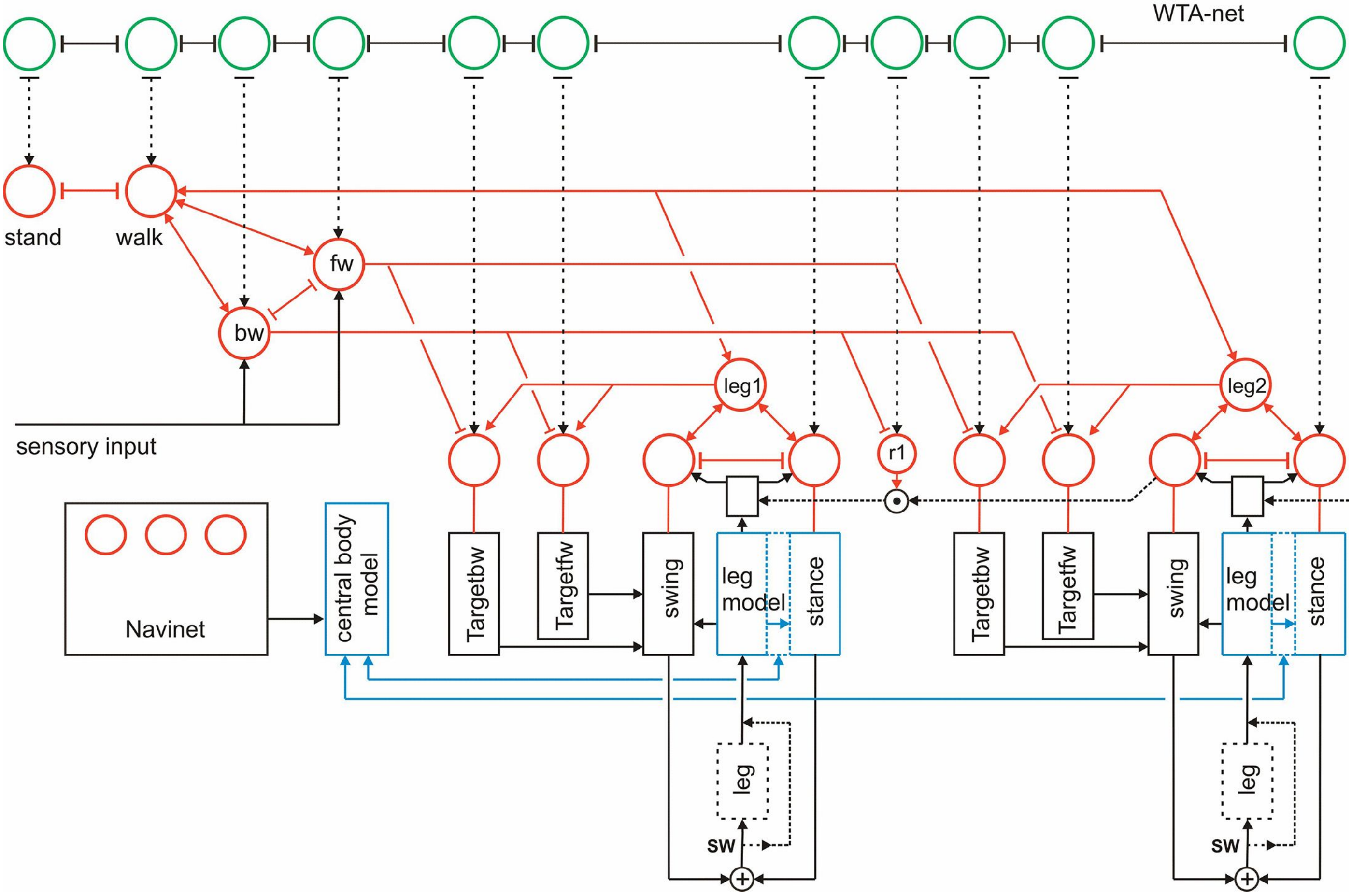

Figure 6: The controller of the reactive system as depicted in figure 2 expanded by a WTA-net (green units, not all connections are shown). Each WTA unit shows a bi-directional connection to a unit of the motivation unit network. This architecture provides the basis of reaCog, as explained in the main text. (for further explanations see figure 2).

motor output. But procedures can also receive input from other procedures and can provide output directly to other procedures. This relatively flat, heterarchical structure is also applied by the Word-nets and in perceptual memory (Cruse & Schilling 2010a).

ReaCog automatically selects actions on the lower reactive level. Several of these procedures can be performed in parallel. On the cognitive level, decisions about which action to choose are not based solely on sensory input, but are chosen depending on the imagined action, since there is a stochastic effect due to noise in the attention controller. The decision is afterwards tested by internal simulation before it is applied to the real system, and only after successful execution is the proposed behavior stored in long-term memory. Therefore, this decision process can be envisioned as a Darwinian type of selection that begins from stochastic "mutations" that are then tested for "fitness" and selected based on this fitness. Thus, reaCog is a minimally cognitive system in the sense of the definition given by McFarland & Bösser (1993).

After we have defined the control network quantitatively, we can use reaCog to analyze emergent properties, which haven't been implemented explicitly. As an example we have already considered a term like "tripod gait" that is sensible on a behavioral level in order to describe the emergent overall behavior of the walker. But on the control level there is no explicit tripod gait controller in reaCog (Schilling et al. 2008; Schilling et al. 2013a). The local influences coupling neighboring legs are responsible for overall coordinated walking behavior (different from many other hexapod controllers), and different gaits can emerge just by choosing different velocities. Therefore, apparent "gaits" or the observation that "cognitive maps" are required can be seen an emergent property of such a network.

In the following, we will turn to concepts that are usually applied in fields different from computer science or behavioral biology, like psychology and philosophy of mind. Choosing another level of description can help us gain a better understanding of the system on a more abstract level. In addition, this approach can lead to more operational definitions for concepts used in other disciplines. This is based on the assumption that many of the above-mentioned phenomena emerge (Vision 2011) and that they can be used as concepts only on a higher, more abstract level.

For some authors, consciousness is thought to be restricted to human beings. In contrast, other authors share the opinion that there are degrees of consciousness and that consciousness does occur, to a smaller degree, in lower-level animals (Dennett 1991). Showing that quite small and simplistic networks can allow for interesting cognitive properties (Chittka & Niven 2009; Menzel et al. 2007) supports such a view, as it provides a plausible evolutionary explanation for consciousness (or better degrees of consciousness). Agreeing with this basic assumption, we want to analyze to what extent our simple control network fulfils certain aspects of consciousness or emotions, even though we did not intend to realize this in our system in the beginning. The graded emergence of such high-level concepts would offer an evolutionary account and might allow us to address questions on the function, e.g., of consciousness, and explain how it relates to the control of behavior.

7 Phenomenality

Before concentrating on specific phenomena, such as emotions or consciousness, we would like to address a more fundamental aspect that appears to be relevant for all higher-level phenomena, namely the occurrence of subjective experience.

An example of subjective experience is pain. Even though it might be possible for us to closely attend to all neuronal activities of a human test subject while stimulating that person's skin with a needle, the observed data would be different from the experienced pain, which is only felt by that person. Nobody other than that person can feel the pain. This form of experiencing an internal perspective is therefore only accessible to us through self-observation. Intuitively, other systems—like non-living things or simple machines—lack such an internal perspective. But in many cases, like for animals, it is hard to determine whether they have subjective experience or are merely reflexive machines that do not possess an internal perspective.

This problem is also visible when we consider a human brain, in the contrasting states of being awake or asleep, for example. While in (dreamless) sleep or under anesthesia the same neuronal systems as in a wakeful state may be active, subjective experience is assumed not to be present. And even in a normal wakeful state, we are not aware of all the contents of the different neuronal activities that take place in our brain. Therefore, only a specific type of neuronal activity seems to be accompanied by subjective experience.

There is only indirect evidence on the conditions required for subjective experience. Libet et al. (1964) performed an early experiment, where the cortex of a human subject was directly stimulated, electronically. Only for stimuli longer than 500 ms did the subjects report a subjective experience. Bloch's law (Bloch 1885) formulates this connection more generally. The subjectively-experienced strength of a stimulus depends on the mathematical product of stimulus duration and stimulus intensity. In other words, a stimulus is only experienced subjectively when the temporally-integrated stimulus intensity surpasses a given threshold.

More recent experiments have studied the concurrent activation of different procedures that compete for becoming subjectively experienced. A basic experiment has been performed by Fehrer & Raab (1962), and has been followed by detailed later studies (Neumann & Klotz 1994). First, participants learned to press a button whenever a square was shown on a screen, but not when two squares were shown in a position on the screen flanking the first square. After the learning period was over, in the experiment the single square was presented for only a short period (about 30 ms), which was then followed by a longer presentation of the two squares. The participants did not report having seen the single square, but reported only having seen the two squares. Nonetheless, they pressed the button. This result shows, first, that the first procedure A ("stimulus single square-motor response"), can be executed without being accompanied by subjective experience of stimulus stimA, the single square. Second, procedure B ("stimulus double squares—no motor response") appears to influence how the first procedure is experienced, i.e., this procedure inhibits the subjective experience of stimulus stimA. Therefore, stimulus stimA is not subjectively experienced (the "masking" effect), but nonetheless triggers the motor reaction.

This situation can be interpreted in the following way (Figure 8, left). On the input side, each procedure shows temporal dynamics that are similar to that of a low-pass filter (LPF) (see footnote on page 2) followed by an integrator (IntA, IntB).[6] Stimulation of one procedure inhibits the representation of the other procedure for some limited time (figure 8, Δt). In addition, both integrators are coupled via mutual inhibition (in figure 8 depicted by separate units). In the masking experiment, the first stimulus (stimA) does not inhibit the second procedure (B), because the latter is not yet stimulated, as long as stimulus stimA is active. In contrast, when the second stimulus, stimB, is given, the representation of procedure A may be suppressed. The representation of the input given by units IntA and IntB activate the corresponding motivation units (MU) of the procedures, MUA and MUB, respectively. This could be explained if we assume two different thresholds. First, the motor command of a procedure can be elicited when a small threshold (thr1, figure 8) is reached. But, a second, larger threshold (thr2, figure 8) must be reached in order to have subjective experience. Then, in our paradigm, procedure A, which was activated first, may reach the level of thr1, which is sufficient to activate the motor output, but not thr2. Only the second procedure, B, has enough time to reach the state of subjective experience (thr2, figure 8, right), which allows the double square (stimB) to become subjectively experienced (however this comes about). The model therefore suffices to explain the basic properties characterizing the backward-masking experiment. As has been shown by Cruse & Schilling (2014), the structure depicted in figure 8 can also deal with a forward-masking paradigm, the so-called attentional blink effect (Schneider 2013). To further describe another experiment, showing the so called psychological refractory period (PRP) paradigm (e.g., Zylberberg et al. 2011), the motivation units (MUA, MUB) of procedure A and procedure B are connected in such a way as to inhibit each other. In other words, the motivation units of these procedures form a WTA network. In addition, each procedure inhibits its own motivation unit after its action has been completed.

From these observations we conclude that there are specific neuronal states that require time to be developed. While eliciting an output signal (like a motor command) is the basic function of the system, this can happen without accompanying subjective experiences. Only some procedures may give rise to such phenomenal experience and might, in addition, trigger subsequent functions in the neural system. For example, this procedure may be able to access more neuronal sources and perhaps allow faster storing of new information (e.g., for one-shot learning). In addition to such functional properties the network can endorse the (mental) property of showing subjective experience, i.e., entering the phenomenal state.

The experimental findings mentioned above support a non-dualist, or monist, view, which means that there are no separate domains (or "substances"), such as the mental and the physical domain, in the sense that there are causal influences from one domain to the other one as postulated by substance dualism. Rather, the impression of there being two "domains"—often characterized as being separated by an ex-

[6] An integrator performs a mathematical integration, i.e., it sums the input over time.

planatory gap (Levine 1983)—, results from using different levels of descriptions.[7]

An explanation of the necessary and sufficient conditions of neural networks that allow for subjective experience would be extremely interesting. Even though there currently exist only early insights or mere speculations, there has been a lot of progress during the past few years (review Schier 2009; Dehaene & Changeux 2011). The continuation of these research projects will hopefully yield a more detailed understanding. Using combinations of neurophysiological and behavioral studies may lead a better understanding of the physiological properties and functions of this state. It is, however, generally assumed that even if we knew the physical details at some future time, we would not understand why this state, which is characterized by physical properties, is accompanied by phenomenal experience. Here we propose another view. We assume that this problem will be "solved" such that the question concerning the explanatory gap will simply disappear, as has happened in the case of explaining the occurrence of life. Concerning the latter, there was an intensive debate between Vitalists and Mechanists at the beginning of the last century on how non-living matter could be transformed into living matter. The Vitalists argued that a special, unknown force, termed *vis vitalis*, was required. After many decades of intensive research, we are in a position where an internal model is available, which represents the observation that a specific collection and arrangement of molecules is endowed with the property of living. This and similar cases may be generalized as the following rule: If we have enough information, such that we can develop an internal model of the phenomena under examination, and if it is sufficiently detailed to allow the prediction of the properties of the system, we have the impression of having understood the system. In the case of life, indeed we do not need a *vis vitalis* any longer, but consider liveliness an emergent property. Correspondingly, we propose that if we knew the functional details and conditions that lead to matter having subjective experience well enough, so that the appearance of subjective experience can be predicted, we would have the impression of having understood the problem. Therefore, we assume that the question of the explanatory gap will disappear at some point, as was the case in the example of life.

Adopting a monist view allows us to concentrate on the functional aspects when trying to compare systems endowed with the phenomenality, i.e., human beings, with animals or artificial systems. According to this view, phenomenality is considered a property that is directly connected with specific functions of the network. This means that mental phenomena that are characterized by phenomenal properties—as are, for example, attention, intention, volition, emotion, and consciousness—can be examined by concentrating on the aspect of information processing (Neisser 1967).

To avoid possible misunderstandings, we want to stress that we do not mean that the phenomenal aspect does not have any function in the sense that the system would work in the same way if there was no such phenomenal properties. Since, according to our view, the phenomenality necessarily arises with such a system, a version of such a system showing exactly the same functions but not having the phenomenal aspect would not be possible. A change in the phenomenal properties of a system has to be accompanied by a change in its functional properties. Functional and phenomenal aspects are two sides of one coin. However, remaining on the functional side makes the discussion much easier.

To summarize, the content of any memory element may be subjectively experienced (or available to conscious awareness) if (1) the (unknown) neuronal structures that allow for the neural dynamics required for the phenomenal aspect to occur are given, and (2) the strength and duration of the activation of the memory element is large enough, provided the element is not inhibited by competing elements.

The question of how any system can possibly have subjective experience was famously called the "hard problem" by Chalmers (1997).

[7] There are various views adopting a monist approach, that differ in detail (epiphenomenalism, emergentism, property dualism and their many derivatives, see Vision 2011). We will not enter into this discussion here.

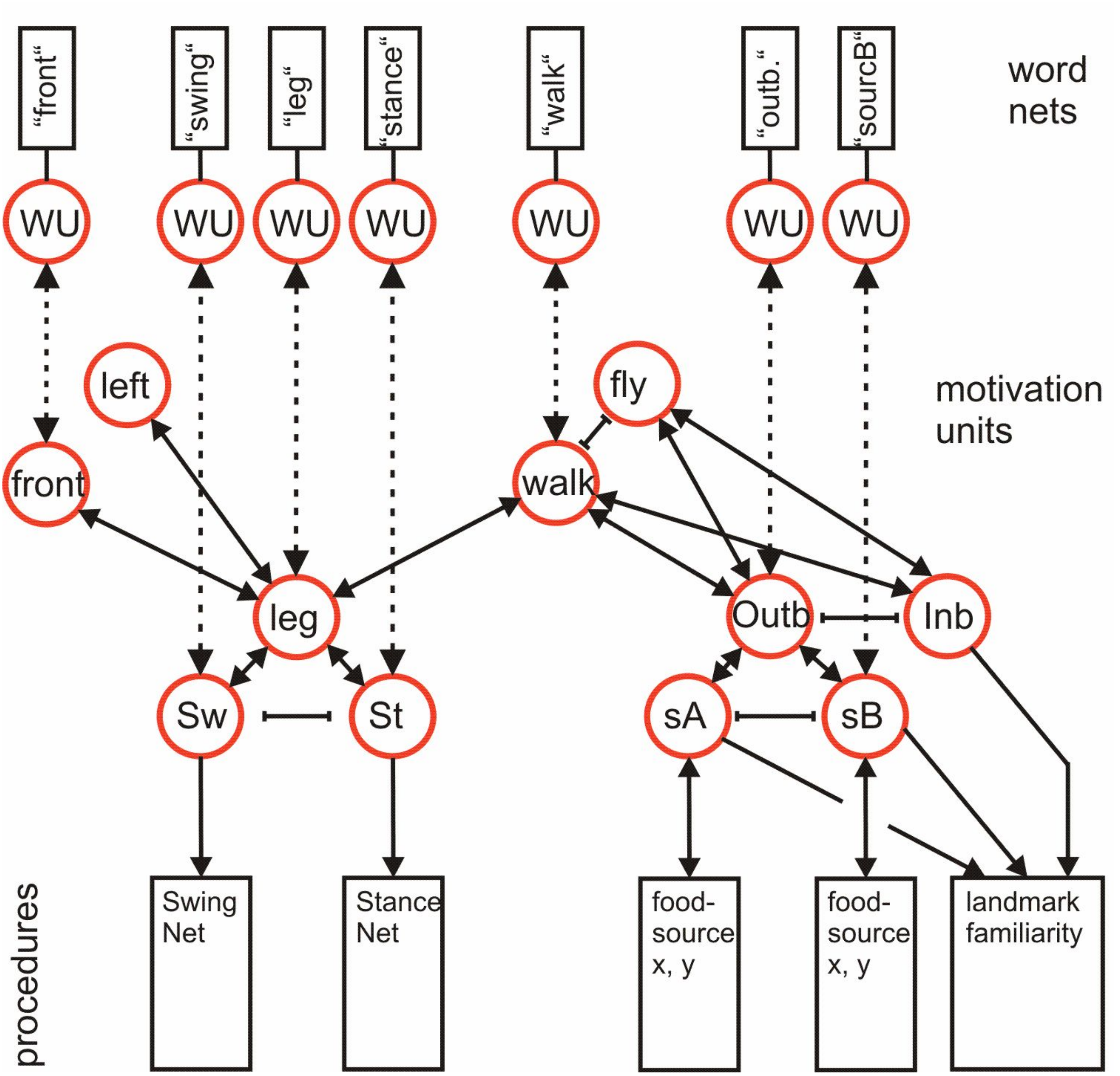

Figure 7: The reactive network expanded by a layer containing procedures that represent words (Word-net, upper row). The motivation unit of a Word-net (WU) is bi-directionally connected (dashed double-headed arrows) with the corresponding motivation unit of the reactive system containing procedural elements of Walknet (left, see figure 2) and of Navinet (right, see figure 4). The word stored in a Word-net is indicated as (" … "). Not all of these motivation units have to be connected with a Word-net.

Adopting a monist view, we can avoid this question and leave it open, as we are interested in understanding the functional aspects of consciousness (on the ethical implications of an artificial system having subjective experience implemented in appropriate neural dynamics see Metzinger 2009, 2013). Regarding what kind of dynamics could be thought of, it has been speculated that subjective experience might occur in a recurrent neural network that is equipped with attractor properties. Following this hypothesis, subjective experience would occur if such a network approached its attractor state (Cruse 2003). This assumption would mean that any system showing an attractor might be endowed with the phenomenon of subjective experience. It may, however, not have all the other properties characterizing consciousness. On the other hand, there might be systems in which the functional aspects currently attributed to consciousness are fulfilled, but where there is no subjective experience present. This case would imply that our list representing the functions of consciousness as given in section 10 below is not yet complete.

In the following two sections we shall briefly treat two phenomena—emotions and consciousness—and discuss how they might be related to the minimally-cognitive system as represented by reaCog.

8 Emotions

Most authors generally agree that emotions are accompanied by subjective experience and that they have the function of helping the subject respond adaptively to environmental pressures. So there is the phenomenal aspect of emotions as well as a functional aspect. As we have already treated the phenomenal aspect above, here we will put aside this aspect, i.e., how it feels to be happy, sad, etc., and concentrate on the functional aspect of emotions.

Even though several authors assume or even demand that emotions are already present in simple reactive systems, and that they are necessary for a cognitive system (Valdez & Mehrabian 1994), in our above description of the properties of the network reaCog, any emotional aspects have not been taken into account. We did not require the term "emotions" to explain our approach, nor have we built in any kind of explicit emotional system. However, we will argue that there are emerging properties that are comparable to what is usually ascribed to properties of emotional systems. In the following, we want to focus on which parts in our system take this role and how the functions of these parts can be described and related to attributes of emotional systems.

The attempt to relate the properties of our network with the concept of emotions appears not very promising at first sight, because a series of interrelated conceptual terms such as emotions, attitudes, motivations, sentiments, moods, drives, and feelings can be found in the literature, and are defined in different but partly overlapping ways by different authors (Pérez et al. 2012). The reason for this disagreement might be that there are indeed no clearly separable mechanisms underlying these phenomena but rather we are dealing with a holistic system, which makes separation into clear-cut concepts difficult, if not impossible. As mentioned, the problem of being confronted with heterarchical structures appeared when looking at the reactive level (and reappeared later when dealing with perceptual memory), which led us to the neutral term "motivation unit" for all "levels" of the heterarchy formed by the motivation unit network. To simplify matters, we will only deal with the term emotions in the following.

What might be possible functions of emotions? As follows from the examples of overlapping conceptual approaches found in the literature and mentioned below, emotions are attributed to various functions characterized by different levels of complexity. These range from enabling the agent to select sensory input (e.g., tunnel vision, Pérez et al. 2012) and activate different procedures, or, at a higher level, to select between different behavioral demands (e.g., hunger – thirst, flight – fight, Parisi & Petrosino 2010) up to more abstract states such as suffering from sadness or being in a state of happiness and controlling the corresponding behaviors (e.g., Ekman 1999). The lower-level decisions are well covered by our motivation unit network, and form a heterarchical system showing attractor states (e.g., swing – stance, Inbound – Outbound). These states allow for selection of sensory input and/or motor procedures that are stimulated by sensory input to specific motivation units. In the following, we therefore focus on higher-level states, such as, for example, emotions, as listed by Ekman (1999).

In general, and as discussed below, one can distinguish between prototypical approaches and reductionist approaches—the latter simplifying emotions down to just a few basic dimen-

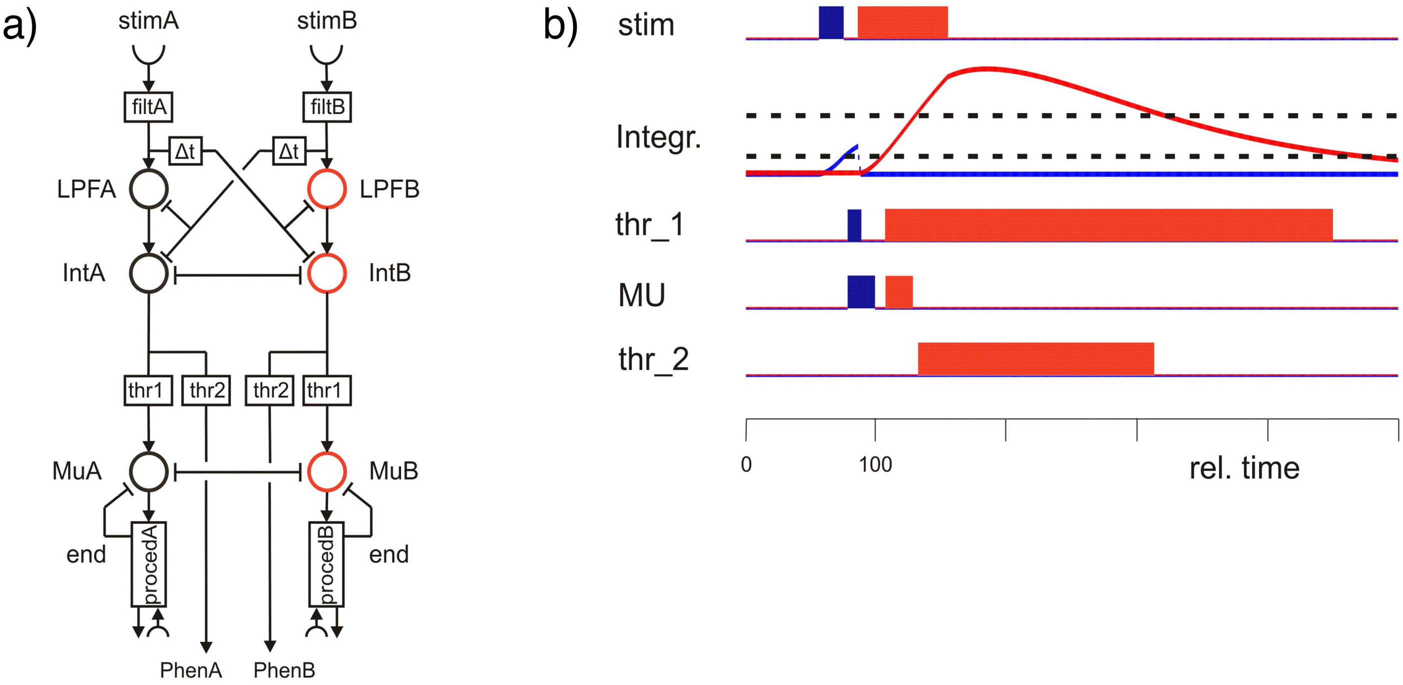

Figure 8: (a) A hypothetical network that is capable of dealing with some dual task experiments, for example the backward masking experiment. Stimulation of one of the procedures, A or B, activates a low-pass filter (LPFA, LPFB) followed by an integrator (IntA, IntB) and inhibits the corresponding units of the other procedure for a limited time (Δt). The integrators are coupled via mutual inhibition. After activation of one of the integrator units has reached threshold thr1 (lower dashed line), the corresponding motor motivation unit (MuA or MuB), coupled via mutual inhibition, is activated, which drives the behavior. If threshold thr2 (upper dashed line) is reached, the stimulus can be phenomenally experienced. A feedback from the procedure can provide an "end" signal to inhibit its own motivation unit. (b) Temporal development of the activation of some units (procedure A, blue, procedure B, red). Abscissa is relative time. If stimB follows briefly after stimA, the unit IntA may reach its motor threshold thr1, but not the threshold thr2 for eliciting the phenomenal experience. In contrast, stimB elicits both the motor output and the phenomenal experience that corresponds to the backward masking effect (for details see Cruse & Schilling 2014).

sions. In current research, both views appear to be justified as they both try to describe the phenomena observed, though at different levels of description.

Following the first approach, research tries to trace emotions back to a set of basic emotions, the combination of which can explain further derived emotions. This approach has been advocated by Plutchik (1980). A problem with such an approach is how to draw borders between emotions and what counts as a basic emotion. Ekman (1999) proposed a list of characteristics of similarity between emotions and came up with a set of fifteen basic emotions. Later on, based on their relation to facial expressions, he reduced this number to six. This set, which is now widely used as the basic set of emotions in many different contexts, consists of happiness, anger, disgust, sadness, fear, and surprise. As an example, let us consider happiness. Happiness is elicited when we are in a state of having had or expecting positive situations. The behavioral effect of happiness might be characterized as being open to new ideas, perhaps not being too critical and open to performing new, unconventional behaviors. How might such a phenomenon be represented in reaCog? First of all, a neuronal state of the motivation network would correspond to a specific emotion. Such a network state is usually triggered by some sensory stimulus eliciting an emotion. This stimulus activates specific, basically innate, networks which, when active, influence the system and put it into the respective emotional state. Such a network—which could, in the most reduced case, consist of just one neuronal unit—has not been introduced in reaCog, but if assumed as given, it may modulate meta-control parameters such as, for example, noise levels, thresholds, or learning rates (Doya 2000, 2002). To stick to our example of "happiness", activation of such a network, which represents stimulus situations considered to elicit this state may, within the Spreading Activation Layer, lead to a faster diffusion process, perhaps supported by stronger noise amplitude. Such a broadening of the attention range as a consequence of positive affects has been reported by Dreisbach & Goschke (2004). In addition, or as an alternative, the threshold for the problem detectors that we mentioned in section 4 might be increased. As a consequence, the system would take more risks. All these changes would lead to an increase in "creativity", i.e., the ability to find new ideas for possible solutions. Corresponding structures might be found in the other basic emotions listed by Ekman.

In the second group of approaches to characterizing the emotions, emotions are described through a set of dimensions that represent the emotional state. We will briefly sketch this seemingly alternative reductionist approach and will again draw parallels with reaCog. The connection to reaCog is made on a different level and is therefore not logically exclusive with respect to the former. Wundt (1863) was quite opposed to the idea of breaking down emotions into a set of basic emotions that serve as prototypes, mainly because he assumed that a set of emotions is better described by a continuum than by separable categories. This follows his idea of describing emotions through principal components leading to dimensional systems, like the pleasure-arousal-dominance (PAD) framework (Mehrabian 1996). In the PAD framework, three dimensions span the space of the emotions. The first describes the state pleasure–displeasure and corresponds to the affective state (excited – relaxed). Arousal, as the second dimension, represents the level of mental alertness and physical activity (tense – sleepy). The third axis describes the level of dominance–submissiveness, i.e., the feeling of being in control. The three factors of the PAD framework have successfully been employed as semantic differential factors to describe emotional states in different contexts, e.g., for describing postures, facial expressions, gestures, and vocal expression. The three dimensions appear to be sufficient as they capture large parts of the variance (Mehrabian 1996). Mehrabian has related the three traits—pleasure, arousal and dominance—to specific cognitive characteristics. First, pleasure-displeasure, according to Mehrabian, deals with the fulfillment of expectations. Fulfillment of an expectation (or not) occurs when, during a problematic situation, planning ahead is activated and after some time and searching a solution is found (or not)—a state that can be found in reaCog, too. But fulfillment of expectation might also occur at lower levels, when, for example, a simple procedure such as Swingnet is equipped with a target value and this goal is either reached or not. The error signal might then be used as a measure for fulfillment of expectation. For example, it might be used as problem detector in the case mentioned earlier, when a subject tries to lift a leg off the ground, but due to an inconvenient load distribution, the body falls down and the leg remains in contact with the substrate. The arousability trait, as introduced by Mehrabian (1996), was meant to incorporate the process of "stimulus screening". In short, "stimulus screening" is a process of attentional focusing. Such a process of focusing attention occurs in our system, too, as, on the one end of the spectrum, the system broadly attends to all environmental influences as perceived through its sensors, and this is characterized as its being in the "reactive state". At the other extreme, when a specific problematic situation occurs, it is necessary to focus attention and to guide the search for a solution towards specific modalities, parts of the body, etc. But even on the reactive level, attention selection can be observed, as we mentioned earlier. Finally, the dominance trait ("generalized expectations of control" Mehrabian 1996) concerns the extent to which the agent takes over in the actual situation and is not only responding and reacting, which agrees with the main thesis of our approach, namely that it is possible to switch between the reactive mode and the cognitive mode. Similarly, Russell &

Norvig (2003) have required an autonomous agent to be able to both react to known situations and to be in control of the situation itself (or as Russell and Norvig call it: being proactive).

Our approach, as we have mentioned, does not aim to build specific emotional properties, but tries to build a functional autonomous system and then to look at the aspects of emotional properties that might be found in the network or gained after some further functional expansion of the network. We have listed examples from different levels of description in psychology and point to related properties in our network. We are not arguing that reaCog has emotions (we are in any case agnostic with respect to the subjective aspect). Rather, we claim that by taking a network like reaCog as a scaffold, different conceptualizations of the functional aspects of emotions can be mapped onto such a quantitatively defined system and thus be considered emergent properties.

It might be added here that recent studies support the idea that emotion-like states do indeed occur in brains which are by far less complex than mammalian brains. Yang et al. (2014) could show that the concept of "learned uncontrollability", generally considered as an animal model for depression as observed in humans, can be found in Drosophila, too. For vertebrates, it is known that stress induces the state of fear or of anxiety, the latter being considered as a second order emotion. Fossat et al. (2014) could show that a crayfish treated by stressors (i) avoids illuminated parts of the environment and (b) shows an increased level of serotonin in the brain, as can be observed in vertebrates. As in vertebrates, the state of anxiety could be relieved by application of anxiolytic drugs. Both results have been interpreted such that the ability to adopt emotional states must have been evolved before the separation of the arthropods and vertebrates.

9 Attention, volition and intention

In the following section we want to turn to attention, intention, and volition. To what degree can those properties be attributed to our system? We start from the definitions of attention provided by Desimone & Duncan (1995), of intention from Pacherie (2006) and Goschke (2013), and of volition from Goschke (2013).

Attention is the ongoing selection process in perception. It can be driven bottom-up, i.e., by sensory influences, or it can be controlled by top-down influences (Desimone & Duncan 1995). Top-down driving of attention depends on the internal or emotional state and might depend on familiarity with the stimulus.

We can indeed find properties corresponding to attention in reaCog. The motivation network is constituted of local clusters of units that always compete on this local level and form in this way coalitions of units and small subclusters. As an example, we introduced the selection of procedures at the leg level. Either a swing or a stance motivation unit can be active and inhibits the other one. These two units compete for control of behavior. Sensory units can influence this competition. For example, an incoming ground-contact signal ends a swing movement and initiates stance activation. After activating the "Stance" unit only sensory input relevant to stance can be perceived by the system, but not inputs relevant to swing. Therefore, this case corresponds to bottom-up attention control.

Such competition can also be found on a global level, on which different behaviors can be chosen. The activation of these higher-level elements influences the lower level. This activation provides a context for the lower level, which guides the selection process on that level and decides which sensory inputs might be relevant. Thereby, more global clusters control the attention on the lower levels in a top-down fashion. Corresponding examples can be found in Navinet, which we mentioned earlier. Only visual signals concerning landmarks that belong to the current active context are considered and switching between contexts only becomes possible after the food source has been depleted and found empty.

The cognitive expansion of reaCog represents another case of top-down influence. This system comes up with new behaviors and probes them via internal simulation. As mentioned, there is a specific WTA layer that mirrors the arrangement of the lower motor control layer (figure 6, green units). This part of the controller can be called an "attention controller", as the explicit function of this layer is to narrow down the search for suitable behavior and to actively select a single one. We call this selection a cognitive decision, as the system is supposed to select a behavior that would not normally be triggered through the given context. In this way the system represents a special type of top-down attention. The focusing mechanism may correspond to what sometimes has been termed "spot light" (Baars & Franklin 2007 p. 955). Overall, we can therefore observe three different types of attentional influences in reaCog.

Volition is an umbrella term denoting mechanisms allowing for voluntary actions. The latter are "actions that are not fully determined by the immediate stimulus situation but depend on mental representations of intended goals and anticipated effects" (Goschke 2013). For an outside observer, voluntary actions cannot be predicted. As mentioned above, it is crucial for the cognitive expansion that it can select behaviors that are not triggered by the current situation. The system has to invent new behaviors. Even though the consequences of these behaviors are predicted, from the outside the finally chosen behavior is not predictable, as this invention and selection of new behaviors is stochastic to some extent. The application of internal simulation only guarantees that the proposed behavior will lead to a solution, but it does not give away which behavior will be chosen. To the contrary, the search space of possible solutions can easily become very large and has to be restricted. Such restrictions help to span a tractable space of possible solutions. In our example, reaCog looks first for solutions in the morphological neighborhood, i.e., it tries to use the neighboring legs to help find a solution for a locally-given problem. There are still many possible behaviors that must be tested in a somewhat random order. The system will end up with one that has been anticipated as a solution in internal simulation, but this solution is not selected through sensory inputs or the current context as such. Therefore, volition may be attributed to a system like reaCog.

Does an agent controlled by reaCog show intentions? Intentions are present when the controlled action is goal-directed. We are following Pacherie (2006), who proposes a differentiation of three types of intentions (based on Bratman's (1987) original differentiation into two such types). Pacherie distinguishes future-directed as well as present-directed intentions and introduces motor-intentions as a third type. Present-directed intentions are considered to be under "conscious" (or "rational") control. In contrast, motor intentions are related to lower-level function (Pacherie 2006). Defining for these types of intention is that they provide guidance for the function on the respective level. In reaCog, motor-intentions are realized by the fact that, on the reactive-control level, behaviors can be selected based on the context. Present-directed intentions can be found on the level of cognitive decision. Future-directed intensions are not treated by reaCog, because its architecture in the current version only allows for dealing with problems that occur in the context of current walking behavior. However, an expansion of reaCog that would include planning ahead using Navinet as a substrate would include future-directed intensions, too.

Goschke (2013) defines intentions as "causal preconditions explaining why a particular stimulus triggers a particular action (rather than a different action)" (Goschke 2013, p. 415). In other words, "intentions can be said to shape the "attractor landscape" of an agent's behavioral state space" (Kugler et al. 1990, ref. from Goschke 2013, p. 415). In reaCog, such an attractor landscape is described by the motivation unit network. As explained in the preceding paragraph on attention, the activation of a context guides, in a top-down fashion, both the selection of a suitable behavior as well as which sensory inputs the system should attend to. The lower-level activation and incoming sensory inputs influence, on the one hand, the adaptive execution of the behavior as such. On the other hand, the sensory input can inform the higher level in bottom-up fashion and might indirectly trigger changes on this higher-level, too. The activation on the higher level will, however, be in general more stable on a temporal scale and will reflect a specific context as well as relate to specific goals. For example, in the case of Navinet, there are different possible goals, such as food sources or the nest, which are represented in the higher-level network. Selecting one of these as a goal will guide the overall function of the system, as its behavior is directed towards approaching that location, while the sensory system will attend only to the specific (expected) sensory stimuli. Therefore, reaCog can be assumed to show goal-directed behavior and intentions.

10 Consciousness

In this section we would like to discuss to what extent properties of consciousness might be found in our system. Even though we start from a common notion of how consciousness can be viewed as consisting of separable domains, we are well aware that this approach is not the only or ultimate solution for approaching this question. But such a differentiation appears well-suited for our bottom-up approach.

Overall, many authors contribute to the view in which consciousness is broken down into a set of properties. We start from a review by Cleeremans (2005), who gives a good overview on the diverse philosophical views on consciousness and tries to integrate them into one framework. While there is disagreement in general and also on the details (see also Vision 2011), Cleeremans interestingly finds a common denominator between the different opinions that characterize possible computational correlates of consciousness. He introduced a differentiation of consciousness into three domains: phenomenal consciousness, access consciousness, and metacognition (or in other contexts referred to as reflexive consciousness). There is disagreement on the phenomenal aspect, as it is seen by one group of philosophers to be an independent domain. In contrast, there is also a view in which phenomenality cannot be separated from metacognition and access consciousness, but must be seen in relation to those (see review Cleeremans 2005).

We have argued in section 7, that the phenomenal aspect as such, i.e., the property of some neuronal structures that are equipped with subjective experience, has *per se* no function, but is, nonetheless, not separable from the functional properties. Therefore, we see the phenomenal aspect not as a separate type of consciousness, but as a property of both access consciousness and metacognition. This view has convincingly been supported by Kouider et al. (2010) as well as, in a recent review, by Cohen & Dennett (2011). Therefore, we will compare properties of reaCog with current definitions found in the literature concerning the phenomena of access consciousness and metacognition, abstracting from the phenomenal aspect.

While other philosophers require metacognition or reflexive consciousness in a system in order to attribute consciousness (see for example Rosenthal 2002 or Lau & Rosenthal 2011 for a recent review defending this view), we do not want and cannot get into this discussion as it is not our goal to review the different types of taxonomies. We basically follow one valid and common perspective, as presented by Cleeremans, and apply it to our system in order to analyze functions of our system that can match the different phenomena described. We do not aim with this approach to give a rigorous definition of consciousness (which does not seem suitable at this point, see also Holland & Goodman 2003). Instead, applying our approach, we aim to provide insight into specific functions of our system that are connected to the phenomena discussed.

10.1 Access consciousness

In this section we want to focus on the aspects of access consciousness that can be found in reaCog. Following Cleeremans, access consciousness of a system is defined by the ability to plan ahead, to guide actions, and to reason, as well as to report verbally on the content of internal representations. In contrast, non-conscious representations cannot be used this way. Selecting behaviors, planning ahead, and guiding actions are the central tasks of reaCog (see section 4, Planning ahead).

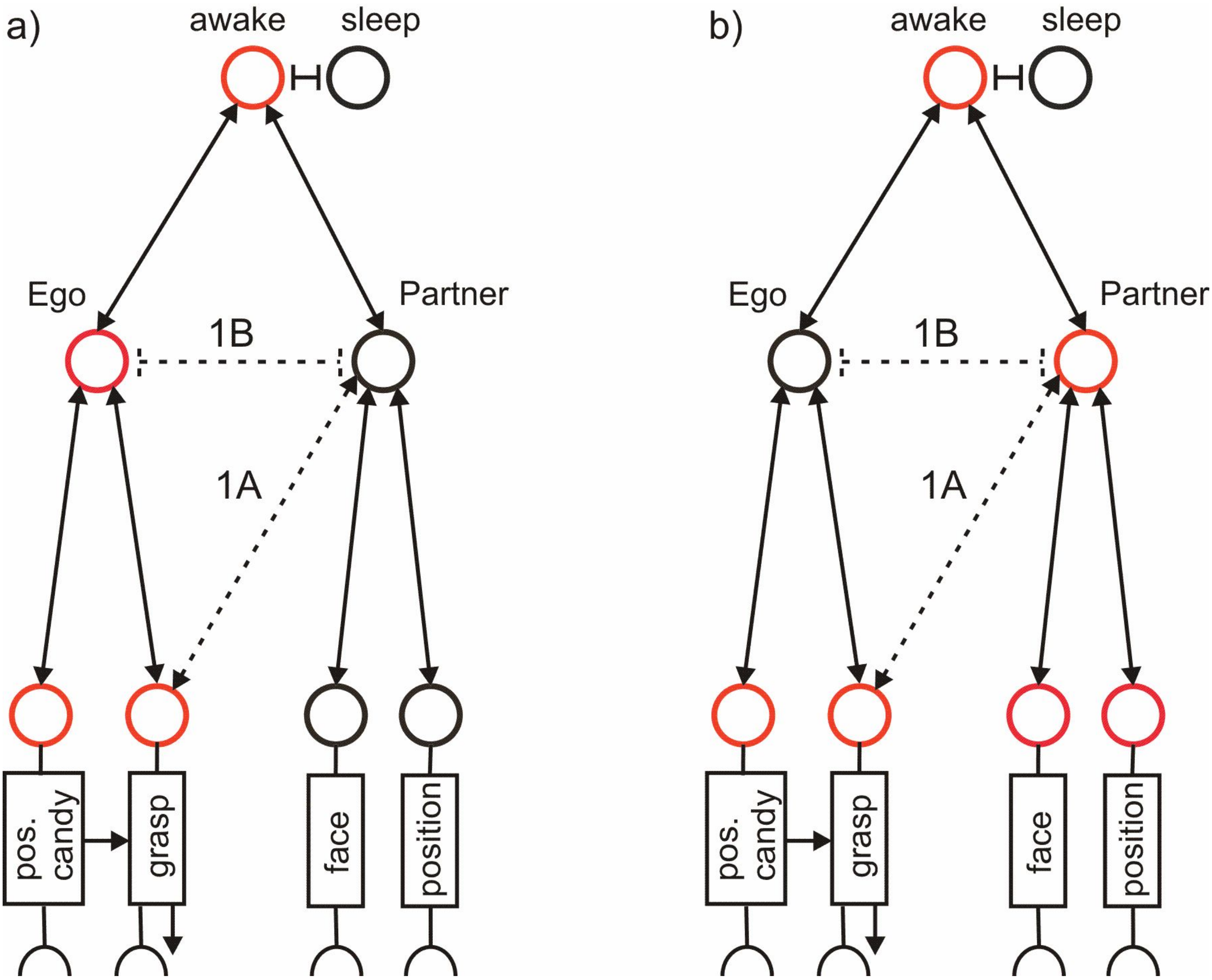

Figure 9: A possible expansion of reaCog. Without the connections 1A and 1B the network enables the agent to represent its own actions, as is already possible for the network shown in figure 2, and figure 6. After introduction of connections 1A and 1B the network is also able to represent the actions of a partner using the now shared procedure "grasp". (a) and (b) show two attractor states where active motivation units are depicted in red, whereas inactive motivation units are shown in black. Half circles indicate sensory input.

Being able to use internal representations for verbal report is currently not a part of reaCog. However, the internal representation of reaCog is already suited to allow for accessing internal representations (section 5 and figure 7). The simple solution proposed allows for communication using one-word sentences only, but provides a way, within the framework of reaCog, for the symbol-grounding problem to be addressed. Steels (2007; Steels & Belpaeme 2005) and Narayanan (1997) have already studied in detail how more complex sentences may be grounded in simple reactive systems. Thus, there already exists work on similar systems that shows how the ability to report by using more complex language structures could be implemented in a reactive system. Therefore, at least in principle, this property could be realized in reaCog, too.

The last property describing access consciousness, symbolic reasoning, is not addressed by reaCog. In the symbolic domain, there are, however, many interesting approaches in the literature that might be connected to a system like reaCog after the symbolic level has been implemented.

Concerning related work, Dehaene & Changeux (2011) review relevant network models that are supposed to simulate consciousness, including their own approach, which is termed global neural workspace theory (GNW) (see also Seth 2007 for a systematic summary). A comparison of reaCog with these approaches can be found in Cruse & Schilling (2013)). Here we will only refer to one important notion, "global availability" as used by several authors to represent a crucial property of access consciousness (e.g., Dehaene & Changeux 2011; Dehaene & Naccache 2001; Baars & Franklin 2007; Cleeremans 2005). Global availability describes the notion that many representations of the system can potentially become conscious. These representations can be selected to solve a current problem (as described for reaCog) or could be selected in a task (see GNW).

Are the representations used in reaCog globally accessible? During execution of a form of behavior the reactive system simply reacts to sensory inputs. Single local modules of the procedural memory are activated by the context, for example, the walking behavior that can execute walking even in a cluttered environment. While the behavior is driven by sensory stimuli, it is not "cognitively attended" and runs automatically in response to direct interaction with the environment. In this case, the representations are not attended by cognitive expansion and are clearly not a part of access consciousness. But, importantly, this can change whenever a problem is detected and the reactive (automatic) system is not sufficient anymore. In such a case, the WTA-net of the attention controller is activated and has to select one of the elements of the procedural memory. During planning, these elements become accessible to the attention system (Norman & Shallice 1986). The WTA-net, which constitutes the essential part of the attention controller, projects directly back to the motivation units of the procedural memories (figure 6, dashed arrows) and thereby selects just one of the possible behaviors (due to the characteristics of a Winner-Take-All network). Therefore, all the procedural modules that could be activated by the attention controller are "globally available" and form possible elements of access consciousness.

10.2 Further relations between reaCog and access consciousness

Another interesting property of reaCog and findings in psychology concern the relation between conscious and automatic procedures. It is well known that humans are able to learn a new behavior by consciously attending to that behavior. Over time, this can change and the execution of the behavior becomes more and more automatic, i.e., it is no longer necessary to be consciously aware of the exact execution of the behavior. A similar shift of attention can be found when reaCog is planning new behaviors. Triggered by the activation of a problem detector, reaCog has to shift its attention towards the new behavior during planning and the following execution of a behavior. As long as the problem-detector is still active, the reactive system is basically suspended (by switching off the loop through the body), and instead the planning system tries out new procedures that have to be attended to. After the successful execution the new solution can also be stored as a procedural memory and become part of the reactive system; it does not require cognitive attention anymore (the procedure how to store this information has not yet been implemented in reaCog). An advantage of this integration into the reactive system is that access to reactive procedures is faster than using the cognitive process, which agrees with the findings mentioned above.

There are other experimental findings highlighting the relation between conscious and non-conscious access to procedural elements. Beilock et al. (2002) found that athletes who have learned a behavior so that it can be performed automatically perform worse when they concentrate on the behavior compared to when performing the behavior while being distracted. In the attention controller of reaCog we can observe a similar phenomenon. If the attention controller is externally activated by a higher-level unit while the connected behavior is performed, this could possibly activate learning.

Such an influence would change the underlying neuronal module and could worsen the result. In contrast, without attention the behavior would be performed as it had been learned earlier.

ReaCog differs in an important aspect from the simulation studies conducted by Dehaene and colleagues, as well as from those conducted by Baars and colleagues. While the latter approaches aim to relate conscious functioning to individual brain areas or brain circuits, reaCog is not intended at all as a model of the human brain or any of its areas. Instead, it is envisioned as a reductionist approach that focuses only on function. From the bottom-up development of more and more higher-level function we offer a post-hoc discussion of the question of to what extent reaCog shows aspects of access consciousness. This approach seems particularly suitable for addressing access consciousness, as it turns out that there is no single identifiable part of reaCog that might be attributed the property of access consciousness. Instead, access consciousness appears to be an emergent property constituted by the complete system. Attention controller, procedural memory, and the connections between those two parts, as well as the internal model and the ability to use it in internal simulation, seem to be the required structures that allow access consciousness, or, in other words, together constitute the "neural workspace." The dynamics of the neural workspace as defined by Dehaene & Naccache (2001) are given through the WTA-net. But, and this is an important difference, there no re-representation in this neural workspace is necessary. The already-present representations can be reused in novel contexts. The existing modules of procedural memory are recruited in the internal simulation when planning ahead. The only difference is that the body is decoupled from the control loop and instead the loop through the world is replaced by a loop using internal models and their predictions as feedback. Together, these representations form the global workspace (this notion of internal models has been termed "second-order embodiment," c.f. Metzinger 2014).

Koch & Tsuchiya (2007) differentiate attention and consciousness, as both can be present individually and independently of each other. They conclude that different mechanisms are responsible for attention and consciousness. While such a differentiation is of course based on basic definitions, we can indeed identify different mechanisms related to these two phenomena, even though they seem to be related. In reaCog, attending to a specific stimulus is modelled as a specific activation of motivation units. Only if this activation is strong enough and/or active for enough time, can the procedure enter the phenomenal state (section 7, figure 8). Therefore, both attention and the phenomenal aspect of consciousness refer to different, but tightly coupled properties of our system.

10.3 Metacognition

Although in this article we use the term cognition in the strict sense as proposed by McFarland & Bösser (1993), when dealing with metacognition, this definition is no longer generally applicable. Therefore, in this section the term cognition is used in the usual, more qualitatively-defined way. We will describe how the motivation unit network could be expanded to allow our agent to be endowed with different aspects of metacognition. These expansions, however, have not yet been simulated by being implemented into the complete network.

Metacognition, or reflexive consciousness (sometimes called metarepresentation), the second essential domain of consciousness, according to Block (1995, 2001) and Cleeremans (2005), is characterized by Lau & Rosenthal (2011) as "cognition that is about another cognitive process as opposed to about objects in the world" (p. 365).

While the selection of procedures for control of behavior may occur on the reactive level or by application of access consciousness, metacognition in addition is able to exploit information concerning a subject's own internal states. As a further property, a metacognitive system, when selecting behavior, can represent itself *as* selecting this behavior ("I make the decision"). Metzinger (2014) classifies this ability as third-order embodiment, where the subject's own body is "explicitly represented as existing" (p. 274) and the "body as a whole" can turn "into an object of self-directed attention" (p. 275). Thus, metacognition is about monitoring internal states in order to exploit this knowledge for the control of behavior. According to Cleeremans (2005), metacognition may also be used for inferring knowledge about the internal states of other agents from observing their behavior and for communicating a subject's own states to others.

Let us first focus on the individual agent. What kind of information might be used by a metacognitive system? A typical case discussed in the literature concerns some quality measure of the procedure to be selected. During decision-making, a person, when relying on own knowledge, needs to be able to access his or her own internal state in order to estimate how sure he or she is about the specific piece of knowledge. Cleeremans et al. (2007) use as an illustrative example a system consisting of two artificial neural networks. While the first network learns an input-output mapping of the task, the other network, as a second-order network, learns to estimate a quality measure describing the performance of the first-order network. As the combination of the two networks does not only store information in the complete system, but also contains information about and for the system, the authors conclude that such a system already shows a limited form of metacognition. Such a network, using an additional second-order subnet, might be implemented in our system, too. For example, motivation units could be activated by confidence, or quality values estimated by such a second-order network. Such a situation can indeed be found in the network Navinet. Navinet is used for navigation control tasks and is inspired by work on navigation in ants. In this system, the salience of a stored stimulus guides memory retrieval (Cruse & Wehner 2011; Hoinville et al. 2012). For instance, the decision to choose one of many different food sources is influenced by the internal representation of the learned food quality (Hoinville et al. 2012). As another example, the confidence value of a visual landmark that is to be followed or not might depend on the salience of the visual stimulus, similar to the implementation of a Bayesian-like system. A different example is given by reaCog, which, by exploiting its internal body model, is capable of representing its own body for internal simulation as well as for control of behavior. Thus, at least some basic requirements for metacognition, such as being able to use own internal representations for the control of behavior, are fulfilled, if we, again, leave the phenomenal aspect aside. Below we will, in addition, briefly address the ability of the agent to represent itself.

How may metacognition be suited to support information transfer between different agents? We will not refer to communication using verbal or gestural symbols here. Instead, we want to start with the ability to identify oneself with another agent, or, in other words, to be able to "step into the shoes of the other." This faculty has been referred to as Theory of Mind (ToM). Central is the notion of being able to attribute mental states to other agents (Premack & Woodruff 1978). A classical example is the "Sally–Anne task". In this experiment, two subjects observe how a cover hides a piece of candy lying on a table. While one subject, Sally, is outside of the room, the other subject, Anne, is able to observe how the hidden candy is moved to a new location. After the change the candy lies underneath a white cover and not under the black cover, which it did to start with. The crucial test question is put to Anne: where does she think Sally will search for the candy? If Anne points to the white cover she only uses her own current beliefs about the situation, but does not apply a ToM, i.e., she does not take into account what Sally believes—since Sally has not observed the switch. But if Anne points to the original location, the black cover, she is assumed to have a Theory of Mind as she operates on a set of mental states that she ascribes to Sally.

ToM is crucial when an agent needs to capture not only physical objects, but in addition represent other agents. It becomes necessary to explicitly keep track of others' observations, plans, and intentions. Only such agents that can attribute mental states to other agents can successfully predict their behavior. There are two common explanations to account for how ToM is realized. First, the so-called theory–theory (Carruthers 1996) assumes that there are dedicated, innate, or learned procedures that allow for prediction of internal states and therefore the behavior of others. We want to concentrate on the second main approach, namely simulation theory (Goldman 2005).

Central to simulation theory is the already introduced notion of an internal simulation. As a prerequisite an agent needs an internal model of him or herself. This model can be used (as explained) for planning ahead using internal simulation. But in the same way this model can also be recruited in order to represent another agent. Thereby, other agents may be mapped onto the own internal model that allows simulating the behavior of the other agent. This faculty would enable the agent to derive all sorts of conclusions based on its own representations, such as, for example, current goals or intentions.

In the case of reaCog, we envision an extension that allows mapping another agent onto the already existing internal model. Internal simulation could be used in this context, too. Therefore, the application of such an internal simulation of another agent could lead to an interpretation of the behavior of the other. However, the two theories mentioned do not necessarily exclude each other, as can be shown when regarding the properties of the cognitive expansion further. If the interpretation found via an internal simulation of another agent is new and succeeds in simulating its behavior, the result could be stored in the procedural memory in a similar way as described for reaCog, when coming up with a new solution to a given problem. In this way, a new procedure has been learned that allows for prediction of the behavior of the other agent. As such, application of simulation theory might in the end lead to results that are described as characterizing theory-theory. The faculty of applying a ToM is currently beyond the ability of reaCog as described above, which allows for an egocentric view only. In the following, we will, however, sketch a way in which such a network may be implemented into the architecture of reaCog (for more details see Cruse & Schilling 2011).

Figure 9 shows a possible expansion of reaCog. Two motivation units represent the state "awake" and the state "sleep", respectively. In the awake state, several sensory and/or motor elements can be activated. These elements may form different contextual groups. To simplify matters, here we focus on two such groups only. One group contains the procedure "grasp" and a memory element representing the visually-given input "position of an object" (relative to the agent), in this case the position of a piece of candy (pos.candy), which is hidden under a cover. We further assume that the agent can also recognize, as a specific kind of object, a conspecific ("partner"), (see Steels & Spranger 2008 and Spranger et al. 2009 for solutions), to whom the agent can attribute properties. These are, in our example, the memory elements "face" and "position", which stand for the visual appearance and spatial location of the partner to be recognized. Together with the unit "partner" these motivation units form an excitatory network (the dashed connections marked 1A and 1B will be treated later). The procedure "grasp" contains a body-model consisting of an RNN (Schilling 2011) that contains information on the arm used for grasping. This network can be applied to both motor control and recognition of the arm. The former function is symbolized by the output arrow. Concerning the latter function, the body-model is used to minimize errors between the position of the internal model of an arm and the (underspecified) visual input of the arm (e.g., Schilling 2011). If the error could be made small enough, the visual input can be interpreted so as to match the morphology and the specific spatial configuration of the model arm. To symbolize this capability, in figure 9 the procedure "grasp" is also equipped with sensory (visual) input.

The network depicted in figure 9 (disregarding connections 1A and 1B) enables the agent to recognize the position of the candy and to grasp it ("Ego grasp candy"), as indicated by the motivation units marked red in figure 9a. It further allows recognition of the face and the position of the partner. But it does not enable the agent to "put itself into the partner's shoes". In other words, the agent is not able to realize that the partner may have his/her own representation of the world. Thus, the capability of a ToM is lacking.

The motivation unit connecting the agent-related elements "pos.candy" and "grasp" has been called "Ego" in the figures. Although not required for the functioning of this network as shown in figure 9a (disregarding connections 1A and 1B), the application of the unit Ego would allow the introduction of a Word-net representing the word "I". Thus, with this expansion the concept of "I", as opposed to other agents (e.g., a partner), can be used by our agent, allowing for internal states like "I grasp candy", and therefore for self-representation.

Unit Ego is, however, necessary in our framework when two units (here "Ego" and "Partner") share elements, as will be the case in the following example, where we will enable the agent to represent the partner performing a grasping movement. To this end, we introduce mutual excitatory connections between the unit representing the partner and the procedural element "grasp" (dashed excitatory connection 1A, figure 9). In addition, Unit "Ego" and unit "Partner" have to be connected via mutual inhibition (dashed inhibitory connection 1B, figure 9). This inhibitory connection has the effect that only one of the units—either unit "Ego" or unit "Partner"—can be activated at a given moment in time. With these additional connections 1A and 1B, the network can adopt the internal state "Partner grasp candy". This situation can be represented in the agent's memory by activation of the motivation units illustrated in figure 9b, highlighted in red. Note that the introduction of connections 1A and 1B does not alter the ability of the agent to represent the situation "Ego grasp candy" addressed above.

The architecture depicted in figure 9, including connections 1A and 1B, has eventually been termed the application of "shared circuits", since the procedure "grasp" can be addressed by both unit "Ego" and unit "Partner", which strongly reminds us of properties characterizing mirror neurons. Therefore, application of such shared circuits has been described as "mirroring" (Keysers & Gazzola 2007). Units of the grasp-net (including the target pos.candy) represent the movement and its goal, and thus correspond to representations of a motor act, such as has been attributed to mirror neurons (Rizzolatti & Luppino 2001). The grasping movement in both cases (figure 9a, b) is represented as being viewed by the agent ("Ego grasp candy", figure 9a) or by the partner ("Partner grasp candy", figure 9b). This means that there is still no ToM possible for the agent. To enable the agent to develop a ToM, we need another expansion.

To explain this, we will present a simple simulation of the Sally–Anne task mentioned above. Both protagonists, Sally and Anne, may have different memory contents concerning the position of the candy. This means that the agent, in this case Anne, needs to be able to represent some aspects of the memory of her partner, too. Therefore, the memory section representing her partner will be equipped with a memory element representing the position of the candy as viewed by her partner Sally, who left the room (figure 10, connection 2). Both memory elements that have possible access to the procedure "grasp" have to be connected by mutual inhibition, so that only one of these elements can address the procedure at a given time in order to allow for sensible representation of the situation. Now imagine that the subject Anne is either equipped with a network as depicted in figure 9, or that depicted in figure 10. Application of a system as shown in figure 9 means that the agent (Anne) has only one representation of the candy's position, namely the one seen last. Therefore only this, correct, position can be activated and it is imagined that the partner grasps the correct position—this kind of prediction is observed in children younger than about four years. Anne cannot take into account the likely assumption her partner will make about the location of the candy. In contrast, in a system as presented in figure 10, there is a difference in thinking of oneself grasping the candy or the partner grasping it. When the agent, Anne, imagines herself grasping the candy, she would grasp its position as under the correct cover (figure 10a). If asked

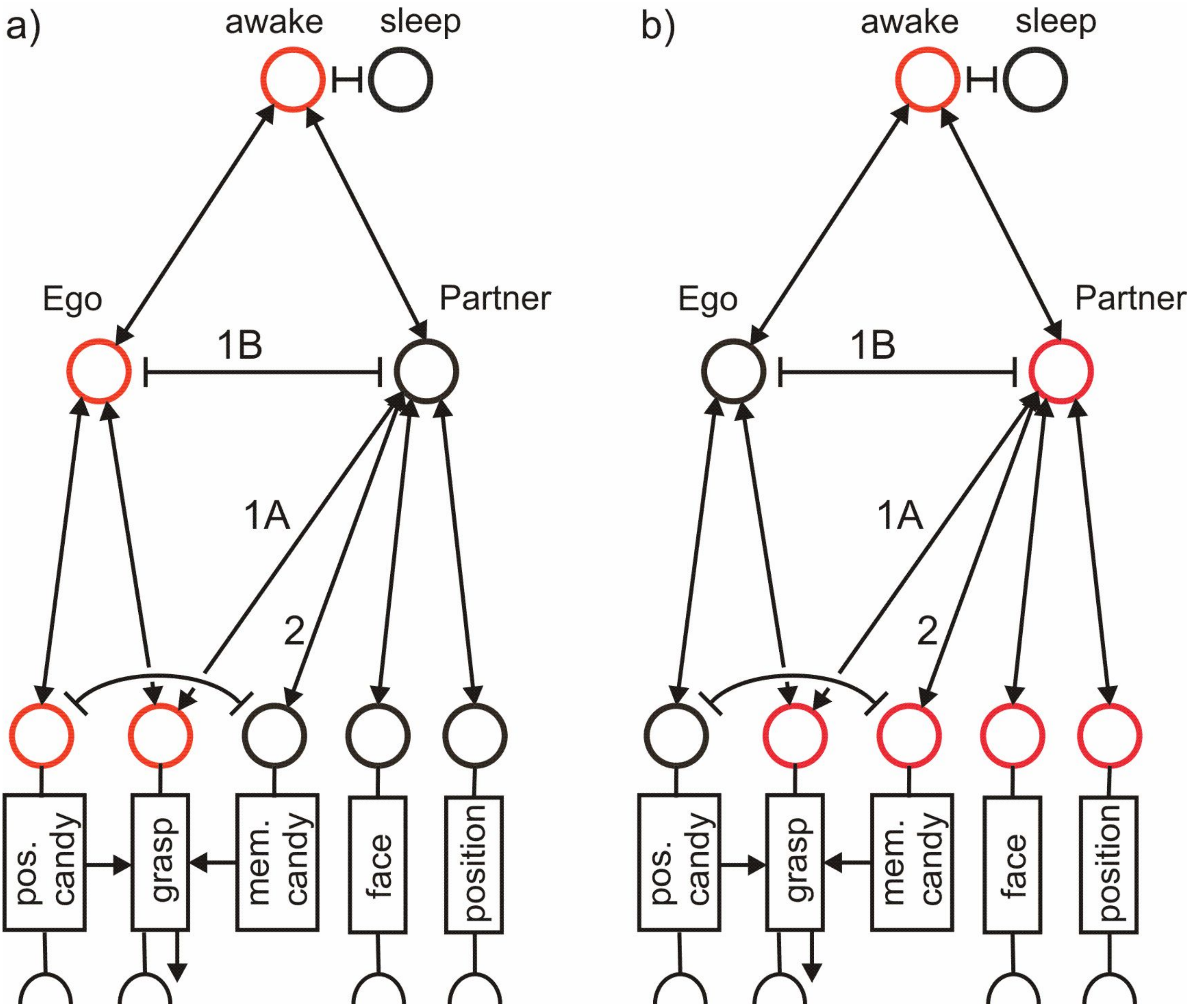

Figure 10: An expansion of the network shown in figure 9, allowing the agent to apply ToM. The expansion concerns the introduction of a new memory element ("mem.candy") plus connection # 2, which enable the agent to represent the assumed content of the partner's memory.

to simulate the internal state of her partner, as is required in the case of the Sally–Anne test (figure 10b), the position connected to her partner Sally is used and the agent will rightfully deduct that her partner's grasp would be directed towards this position—which is wrong, but this fact is not known by her partner. Therefore, the network shown in figure 10 allows for ToM, in contrast to the network shown in figure 9. The critical difference between both networks is that the network shown in figure 10 contains a separate representation of (a part of) the partner's memory. This means that a comparatively simple expansion of our network shows how the agent could be equipped with the ability to apply ToM.

11 Discussion

Consciousness and the relation of the outside world to mental representation are central to philosophy of mind, and have led to many diverse views (Vision 2011). While many of those views appear plausible in themselves, especially from a non-philosopher's perspective, there appears to be much disagreement among philosophers. Many of the positions are based on high-level views approaching consciousness in a top-down fashion. In contrast, our approach starts from a low-level control system for a behaving agent. The goal is the bottom-up development of higher-level faculties. In this way, the neural architecture implements a minimal cognitive system that can be used as a hypothesis for cognitive mechanisms and higher-level functioning, which are testable in a real-world system, for example, on a robot. This allows deriving testable and quantitative hypotheses for higher-level phenomena. In this way, a bottom-up approach can nicely complement philosophical discussions focusing mainly on higher-level aspects. In addition, such a minimal cognitive system can provide functional descriptions of higher-level properties. We briefly introduced the reaCog system in this article, following this bottom-up approach. The central concern is the emergent properties that can be identified when analyzing this system. In particular, high-level properties, such as emotions, attention, intention, volition, or consciousness have been considered here and related to the system.

From our point of view, such a bottom-up approach leads to a system that can be used to test quantitative hypotheses. Even though the system was not intended to model, for example, consciousness, the system can be thoroughly analyzed and emergent properties can be related to mental phenomena. This is particularly interesting, as high-level descriptions can leave a lot of room for interpretation. In contrast, connecting mental phenomena to mechanisms of a well-defined system allows for detailed studies and clear-cut definitions on a functional level. In this way, a system can be examined with respect to many even diverging views and may allow resolving ambiguities. Knowledge gained from analyzing the system can in this way inform philosophical theories and refine existing definitions by defining sufficient aspects as well as missing criteria.

One might ask if higher-level phenomena as considered here are not simply too far removed for such a simple system. One basic problem is represented by the frequently-formulated assumption that all these phenomena have to be tied to the notion of an internal perspective and that phenomenality has a function in and of itself. In contrast, we claim that focusing on the functional aspect is a sensible approach. It is possible because we believe that the phenomenal aspect is always coupled to specific, yet unknown, properties of the neuronal system that, at the same time, have functional effects and show subjective experience. In other words, adopting a monist view, we assume that we can circumvent the "hard" problem, i.e., the question concerning the subjective aspect of mental phenomena, without losing information concerning the function of the underlying procedures. Of course, we are not in a position to claim which of these structures, if any, are accompanied by phenomenality. If, however, the function of, for example, the artificial system indeed corresponds well enough to those of the neuronal structures that are accompanied by phenomenality, the artificial system may have this property, too.

The control network reaCog consists of local procedural modules. We have presented two subnetworks: Walknet, which aims at the control of walking, and Navinet, which deals with navigation. Both consist of a heterarchical structure of motivation units that form a recurrent neural network. This, via competition and cooperation between those units, allows for various attractor states that enforce action selection. Selection of one or a group of procedures protects a current behavioral context against non-relevant sensory input. An internal model of the body is part of the control network coordinating joint movements in walking. As this model is quite flexible and predictive, it can be used for planning ahead through internal simulation. Following the definition of McFarland & Bösser (1993), the network, since it is based on reactive procedures and is capable of planning ahead, can be termed a cognitive system, giving rise to its name: reaCog. In combination with the attention controller, the whole framework can come up with new behavioral solutions when encountering problems, i.e., behaviors that are not automatically activated by the current context. Internal simulation allows us to test these behaviors and to come up with predicted consequences, which can be used to guide the selection process for the real system. The attention controller cannot function independently. It is tightly connected to the reactive structures. The procedural memory of the reactive system is further accompanied with perceptual memory and Word-nets, a specific form of mixed procedural and perceptual memory. The latter memory elements allow the introduction of symbolic information. Symbol-grounding is realized by specific connections between the motivation unit of a Word-net and its partner motivation unit, representing the corresponding concept in the procedural (or the perceptual) memory.

Key characteristics of reaCog are modularity, heterarchy, redundancy, cross-modal influences (e.g., path integration and landmark navigation in Navinet), bottom-up and top-down attention control, i.e., the selection of relevant sensory inputs, as well as recruitment of internal models for planning. The complete control system constitutes a holistic system as the central selection control process—including the internal body-model—is implemented as an RNN. Overall, reaCog follows Anderson's massive redeployment hypothesis (Anderson 2010), since large parts of the reactive control network structure are reused in higher-level tasks (as discussed in detail in section 4 for planning ahead and in section 10.3 for Theory of Mind).

ReaCog nicely demonstrates how complex behavior can emerge from the interaction of simple control networks and coordination on a local level, as well as through the loop through the environment. Its feasibility is shown through the implementation of the system at first in dynamic simulation (for Navinet on a two DoF, wheeled robot platform; for Walknet using a hexapod, twenty-two DoF hexapod robot). Second, those control networks are currently applied to a real robot, called Hector (Schneider et al. 2011).

Emergent properties are properties that are to be addressed using levels of description other than those used to describe the properties of the elements. In the reactive part of the system (Walknet, Navinet) we have already found some emergent properties (development of different "gaits", climbing over large gaps, finding shortcuts in navigation characterized as cognitive-map-like behavior) as well as forms of bottom-up and top-down attention. With respect to the notion of access consciousness, several contributing properties are present in reaCog. Most notably, planning ahead through internal simulation is central to reaCog. New behavioral plans are tested in the internal simulation, thus exploiting the existing internal model and its predictive capabilities. Only afterwards are successful behaviors applied on the real agent. In this way, the agent can deal with novel contexts and is not restricted to the hard-wired structure of the reactive system.

Furthermore, the system shows global availability, which means that elements of the procedural memory can be addressed even if they do not belong to the current context. A third property contributing to elements forming access consciousness concerns the ability of the system to communicate with an external supervisor by following (i.e., understanding) verbal commands and by reporting on its internal states. Therefore, except for the ability of linguistic reasoning, which is clearly missing, the issues characterizing access consciousness as listed by Cleeremans (2005) are fulfilled. But there are also disadvantages: (i) First, reactive automatic control is faster. As cognitive control involves internal simulation (and probably multiple simulations) the whole process takes more time. In addition, there is an overhead of higher-level control going on in contrast to reactive control. (ii) While access consciousness enables the system to deal with novel situations and to come up with new behaviors, the same processes might interfere when they are active during processing of the reactive control level. This might lead to worse performance when both levels are active at the same time. Both mentioned drawbacks have been confirmed in psychological experiments. We have not dealt with the subjective aspect of consciousness. But leaving this aside, we have shown how reaCog shows important constituent properties of access consciousness and how it may provide, in this way, a scaffold for a more complex system that can manifest additional basic aspects of consciousness.

The property of having an internal body-model and the property of being able to internally simulate behavior have been explicitly implemented and can therefore not be considered emergent properties in our approach. However, when referring to a hypothetical evolutionary process that may have led to the development of these properties, the appearance of the body-model and of cognitive expansion might well be characterized as representing an emergent property.

We based our analysis and discussion on the perspective of Cleeremans, and used his concepts. One counter argument addressing the notion of access consciousness is that this notion is too unspecific as it does not help to distinguish between systems, and may cover "too many" systems. For instance, one may ask, following a minimalist approach, whether this notion of access consciousness might even include programs like chess-playing software. One might also ask whether there is a fundamental difference between such a system and a system like reaCog.

While both systems are able to search for the solution to a problem using internal simulation, there are indeed crucial differences. A typical chess program would be not embodied, but, obviously, today this difference can be easily overcome and the system could be realized in a robot equipped with a vision system and a hand that could move the chess figures.

However, more importantly, the basic difference between such a chess player and reaCog would be their flexibility in using internal models. A chess-playing robot always operates within the same context, which is stored in a separate memory-domain, for example in a list of symbolic rules. In contrast, reaCog basically operates with a reactive system, but can also switch to the state of internal simulation when a problem occurs. It then searches for a solution by testing memory elements not belonging to the actual context. In other words, reaCog is able to exchange information between different contextual domains. Such a switch is not available to a chess-playing program at all. Such a program cannot distinguish between different contexts. In other words, there is no global accessibility in the sense described for systems showing access consciousness. As a consequence, the discussion of drawbacks connected with access consciousness as mentioned in the above paragraph on emergent properties, that is, issues (i), and (ii), is not applicable to such a chess-playing system, and nor are the dynamical effects observed in the experiments of Beilock et al. (2002) 1 (section 10.2).

The same holds for the phenomena of a psychological refractory period, attentional blink, and the masking experiments discussed earlier in section 7. None of these phenomena can be addressed by a classical chess-player system, first, because due to the different architectures, no search of a domain belonging to a different context is possible. A chess player does not meet the requirements of access consciousness as listed by Cleeremans 2005 and represented by reaCog. Second, no specific dynamics can be found in such a chess-player system that could be made responsible for the dynamical effects mentioned above and which may provide the substrate for the occurrence of phenomenal experience. Therefore, both systems are qualitatively different. If at all, the chess player may correspond to a subsection of the symbolic domain of access consciousness, which has not yet been explicitly addressed in this article.

In an earlier paper (Cruse & Schilling 2013), taking a conservative position, we argued that properties of metacognition could not be found in the earlier version of reaCog. We have now provided some new arguments that permit a different position concerning this matter. Using this architecture, the agent is able to monitor internal states and use this information to control its behavior. Internal states may also be able to represent the agent itself. A first expansion allows representation of the activations of a partner by using the same procedure as is used for controlling the agent's own behavior (application of "shared" circuits, "mirroring"). Furthermore, using an expansion proposed by Cruse & Schilling (2011), the agent is also able to exploit and represent knowledge about the internal states of others, specifically by applying ToM.

Cruse & Schilling (2011) have further shown how this network can be expanded to represent the discrimination between subject and object (e.g., Ego push Partner) and to attribute subjective experience (e.g., pain) to the partner using a shared body-model. A further expansion that allows for mutualism—two agents cooperate to reach a common goal ("shared intention", Tomasello 2009)—requires two body-models, corresponding to what Tomasello calls a we-model.

In the remainder of this section we briefly mention some aspects not addressed by reaCog. First, not all combinations of the elements explained for our network have been tested within the complete system. For example, Walknet and Navinet have been tested in separate software and hardware simulations. Second, we concentrated on solving motor problems alone, and did not deal with how this system could solve problems in the symbolic domain at all. From an embodied point of view, this restriction is not as problematic as it might initially seem, as the solution process for many problems can be traced back to abilities that are based on solving motor tasks (Glenberg & Gallese 2011); for example this holds true even for abstract domains such as mathematical problem-solving (Lakoff & Nunez 2000).

Finally, an important aspect not addressed here in any detail concerns how learning of the memory elements, including the weight of the motivation unit network, is possible. Examples of learning position and quality of new food sources in Navinet are given by Hoinville et al. (2012), examples of learning perceptual networks, including the heterarchical arrangement of concepts, are given by Cruse & Schilling (2010a), but introduction of the ability to learn such properties within the complete system has not yet been introduced.

12 Conclusion

We describe a way to construct an artificial agent whose architecture is characterized by a number of local, reactive procedures controlled by an RNN, termed motivation unit network. This network is able to adopt various attractor states, or internal states, which are able to protect the complete system from sensory input not belonging to the current internal state. No strict hierarchy can be observed in this network. Instead, internal states may be represented by partly overlapping state vectors.

Where required, further procedures have been introduced that can be interpreted as forming explicit representations of parts of the environment. Specifically, an internal model of the agent's own body is introduced that can, as a "manipulable" body-model, be used for planning new behaviors via internal simulation. Internal manipulation is possible because the body-model, like a marionette puppet, able to adopt all configurations the real body can assume. This expansion allows the agent to switch between reactive control and cognitive control (in the sense of McFarland & Bösser 1993).

When aiming to study higher mental properties, at least in human beings, we have to deal with the phenomenal aspect of these properties. A number of experimental results suggest that, i) some, but not all neuronal activities are, under specific—and unknown in any detail—conditions equipped with a phenomenal aspect, i.e., show subjective experience, but that ii) there is no specific function of this phenomenal aspect apart from the functions that can be ascribed to the physical properties of the system. Note that this does not mean that the phenomenal aspect has no function. Rather, a network adopts the function only when, at the same time, the phenomenal aspect is given. This view allows us to focus the analysis on the functional aspect of the procedure (see section 7). However, due to our lack of knowledge, as an external observer we cannot decide whether a given internal state is a mental state or not (if mental states are understood as internal states that are equipped with a phenomenal aspect).

The complete network represents a collection of hypotheses that can be tested by comparing their properties with experimental data and by trying to match them with theoretical concepts. Examples studied in this article concern behaviors that, for an external observer, may be conceptualized as various gait patterns, or navigation using an internal map, on the "lower" level. On a higher level, we deal with inventing new behaviors and planning ahead, as well as phenomena attributed to mental states like emotions, attention, intention, and volition. Last but not least we compare the properties of our approach with different aspects of consciousness, such as access consciousness (including global accessibility) and metacognition. We claim that, at least in their basic form, these phenomena can be attributed to internal states emerging from the cooperation of decentralized elements of our network.

References

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, *33* (4), 254-313. 10.1017/S0140525X10000853

Baars, B. J. & Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: Global workspace theory and IDA. *Neural Networks*, *20* (9), 955 - 961. 10.1016/j.neunet.2007.09.013

Beilock, S. L., Carr, T. H., MacMahon, C. & Starkes, J. L. (2002). When paying attention becomes counter-productive: Impact of divided versus skill-focussed attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied*, *8* (1), 6-16. 10.1037/1076-898X.8.1.6

Bloch, A. M. (1885). Expérience sur la vision. *Comptes Rendus de Séances de la Société de Biologie*, *37*, 493-495.

Block, N. (1995). On a confusion about a function of consciousness. *The Behavioral and Brain Sciences*, *18* (2), 227-287. 10.1017/S0140525X00038188

——— (2001). Paradox and cross purposes in recent work on consciousness. *Cognition*, *79* (1-2), 197-219. 10.1016/S0010-0277(00)00129-3

Bläsing, B. (2006). Crossing large gaps: A simulation study of stick insect behavior. *Adaptive Behavior*, *14* (3), 265-285. 10.1177/105971230601400307

Bratman, M. E. (1987). *Intention, plans and practical reason.* Cambridge, MA: Harvard University Press.

Carruthers, P. (1996). Simulation and self-knowledge: A defence of the theory-theory. In Carruthers, P. and Smith, P.K. (Eds.) *Theories of theories of mind.* Cambridge, UK: Cambridge University Press.

Chalmers, D. (1997). *The conscious mind : In search of a fundamental theory.* New York, NY: Oxford University Press.

Chittka, L. & Niven, J. (2009). Are bigger brains better? *Current Biology* (19), R995-R1008. 10.1016/j.cub.2009.08.023

Cleeremans, A. (2005). Computational correlates of consciousness. *Progress in Brain Research*, *150*, 81-98. 10.1016/S0079-6123(05)50007-4

Cleeremans, A., Timmermans, B. & Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. *Neural Networks*, *20* (9), 1032-1039. 10.1016/j.neunet.2007.09.011

Cohen, M. A. & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15* (8), 358-64. 10.1016/j.tics.2011.06.008

Cruse, H. (2003). The evolution of cognition: A hypothesis. *Cognitive Science*, *27* (1), 135-155. 10.1207/s15516709cog2701_5

Cruse, H. & Schilling, M. (2010a). Learning and retrieval of hierarchically organized information in a simple, one-layered RNN. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010 at WCCI 2010 IEEE World Congress on Computational Intelligence), Barcelona, Spain* (pp. 521-528). 10.1109/IJCNN.2010.5596804

——— (2010b). Getting cognitive. In Bläsing, P., Puttke, M. and Schack, T. (Eds.) *The Neurocognition of Dance* (pp. 53-74). Psychology Press.

——— (2011). From egocentric systems to systems allowing for theory of mind and mutualism. In R. Doursat (Ed.) *Proceedings of the ECAL 2011, Paris* (pp. 184-191). MIT Press.

——— (2013). How and to what end may consciousness contribute to action? Attributing properties of consciousness to an embodied, minimally cognitive artificial neural network. *Frontiers in Psychology*, *4* (324). 10.3389/fpsyg.2013.00324

——— (2014). Action selection within short time windows. *Biomimetic and Biohybrid Systems*, *8608*, 47-58. 10.1007/978-3-319-09435-9_5

Cruse, H. & Wehner, R. (2011). No need for a cognitive map: Decentralized memory for insect navigation. *PLoS Computational Biology*, *7* (3), e1002009. 10.1371/journal.pcbi.1002009

Dehaene, S. & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70* (2), 200-227. 10.1016/j.neuron.2011.03.018

Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79* (1-2), 1-37. 10.1016/S0010-0277(00)00123-2

Dennett, D. C. (1991). *Consciousness explained.* Boston, MA: Little, Brown & Co.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193-222. 10.1146/annurev.ne.18.030195.001205

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10* (6), 732-739. 10.1016/S0959-4388(00)00153-7

——— (2002). Metalearning and neuromodulation. *Neural Networks*, *15* (4-6), 495-506. 10.3410/f.1001684.173108

Dreisbach, G. & Goschke, T. (2004). How positive affect modulates cognitive control: reduced perseveration at the cost of increased distractibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30* (2), 343-53. 10.1111/j.1460-9568.2007.05949.x

Dürr V., Schmitz, J. & Cruse, H. (2004). Behaviour-based modelling of hexapod locomotion: Linking biology and technical application. *Arthropod Structure & Development*, *33* (3), 237-250. 10.1016/j.asd.2004.05.004

Ekman, P. (1999). Basic emotions. In Dalgleish, T. and Power, M. J. (Eds.) *Basic emotions* (pp. 45-60). New York, NY: John Wiley & Sons Ltd..

Fehrer, E. & Raab, D. (1962). Reaction time to stimuli masked by metacontrast. *Journal of Experimental Psychology*, *62* (2), 143-147. 10.1037/h0040795

Fossat, P., Bacqué-Cazenave, J., De Deurwaerdère, P., Delbecque, J.-P. & Cattaert, D. (2014). Comparative behavior. Anxiety-like behavior in crayfish is controlled by serotonin. *Science*, *344* (6189), 1293-1297. 10.1126/science.1248811

Gibson, J. J. (1979). *The ecological approach to visual perception.* New Jersey: Lawrence Erlbaum Associates.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20* (1), 1-55.

Glenberg, A. M. & Gallese, V. (2011). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, *48* (7), 905-922. 10.1016/j.cortex.2011.04.010

Goldman, A. (2005). Imitation, mind reading, and simulation. In Hurley, S. and Chater, N. (Eds.) *Perspectives on imitation II* (pp. 80-81). Cambridge, MA: MIT Press.

Goschke, T. (2013). Volition in action: intentions, control dilemmas, and the dynamic regulation of cognitive control. In Prinz, W., Beisert, M. and Herwig, A. (Eds.) *Action science: Foundations of an emerging discipline* (pp. 409-434). Cambridge, MA: MIT Press.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences*, *6* (6), 242-247. 10.1016/s1364-6613(02)01913-7

Hoinville, T., Wehner, R. & Cruse, H. (2012). Learning and retrieval of memory elements in a navigation task. In Prescott, T.J., Lepora, N.F., Mura, A. and Verschure, P.F.M.J. (Eds.) *Biomimetic and Biohybrid Systems* (pp. 120-131). 10.1007/978-3-642-31525-1_11

Holland, O. & Goodman, R. (2003). Robots with internal models: A route to machine consciousness? *Journal of Consciousness Studies, Special Issue on Machine Consciousness*, *10* (4-5), 77-109.

Keysers, C. & Gazzola, V. (2007). Integrating simulation and theory of mind: From self to social cognition. *Trends in Cognitive Sciences*, *11* (5), 194-196. 10.1016/j.tics.2007.02.002

Koch, C. & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences*, *11* (1), 16-22. 10.1016/j.tics.2006.10.012

Kouider, S., de Gardelle, V., Sackur, J. & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, *14* (7), 301-307. 10.1016/j.tics.2010.04.006

Kugler, P. N., Shaw, R. E., Vicente, K. J. & Kinsella-Shaw, J. (1990). Inquiry into intentional systems I: Issues in ecological physics. *Psychol. Res.*, *52* (2), 98-121. 10.1007/BF00877518

Lakoff, G. & Nunez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York, NY: Basic Books.

Laughlin, R. B. & Pines, D. (2000). The theory of everything. *Proceedings of the National Academy of Sciences of the United States of America*, *97* (1), 28-31. 10.1073/pnas.97.1.28

Lau, H. & Rosenthal, D. M. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15* (8), 365-373. 10.1016/j.tics.2011.05.009

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*, 354-361.

Libet, B., Alberts, W. W., Wright, E. W., Delattre, L. D., Levin, G. & Feinstein, B. (1964). Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex. *Journal of Neurophysiology*, *27* (4), 546-578. 10.1007/978-1-4612-0355-1_1

Loeb, G. E. (2001). Learning from the spinal cord. *Journal of Physiology*, *533*, 111-117. 10.1111/j.1469-7793.2001.0111b.x

Maes, P. (1991). A bottom-up mechanism for behavior selection in an artificial creature. In Meyer, J.-A. and Wilson, S.W. (Eds.) (pp. 238-246). Cambridge, MA: MIT Press.

McFarland, D. & Bösser, T. (1993). *Intelligent behavior in animals and robots.* Cambridge, MA: MIT Press.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *4*, 261-292.

Menzel, R., Brembs, B. & Giurfa, M. (2007). Cognition in invertebrates. In Kaas, J.H. (Ed.) *Evolution of nervous systems in invertebrates* (pp. 403-442). Oxford, UK: Oxford University Press.

Metzinger, T. (2006). Different conceptions of embodiment. *Psyche*, *12* (4)

——— (2009). *The ego tunnel - The science of the mind and the myth of the self.* New York, NY: Basic Books.

——— (2013). Two principles for robot ethics. In Hilgendorf, E. and Günther, J.-P. (Eds.) *Robotik und Gesetzgebung* (pp. 263-302). Baden-Baden: Nomos.

——— (2014). First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood. In Shapiro, L.A. (Ed.) *The Routledge Handbook of Embodied Cognition* (pp. 272-286). London, UK: Routledge.

Mussa-Ivaldi, F. A., Morasso, P. & Zaccaria, R. (1988). Kinematic networks distributed model for representing and regularizing motor redundancy. *Biological Cybernetics*, *60* (1), 1-16. 10.1007/BF00205967

Narayanan, S. (1997). Talking the talk is like walking the walk: A computational model of verbal aspect. *COGSCI-97* (pp. 548-553). Stanford, CA. 10.1.1.35.1211

Neisser, U. (1967). *Cognitive psychology.* New York, NY: Appleton-Century-Crofts.

Neumann, O. & Klotz, W. (1994). Motor responses to non-reportable, masked stimuli: Where is the limit of direct parameter specification? In Umiltà, C. and Moscovitch, M. (Eds.) *Attention and performance XV* (pp. 123-150). Cambridge, MA: MIT Press.

Norman, D. A. & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In Davidson, R.J., Schwartx, G.E. and Shapiro, D. (Eds.) *Consciousness and self-regulation: Advances in research and theory* (pp. 1-18). New York, NY: Plenum.

O'Connor, C. M., Cree, G.S. & McRae, K. (2009). Conceptual hierarchies in a flat attractor network: Dynamics of learning and computations. *Cognitive Science*, *33* (4), 665-708. 10.1111/j.1551-6709.2009.01024.x

Pacherie, E. (2006). Toward a dynamic theory of intentions. In Pockett, S., Banks W.P. and Gallagher, S. (Eds.) *Does consciousness cause behavior? An investigation of the nature of volition* (pp. 145-167). Cambridge, MA: MIT Press.

Parisi, D. & Petrosino, G. (2010). Robots that have emotions. *Adaptive Behavior*, *18* (6), 453-469. 10.1177/1059712310388528

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Plutchik, R. and Kellerman, H. (Eds.) *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion* (pp. 3-23). New York, NY: Academic.

Premack, D. G. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral Brain Sciences*, *1* (4), 515-526. 10.1017/S0140525X00076512

Pérez, C. H., Escribano, G. S. & Sanz, R. (2012). The morphological approach to emotion modelling in robotics. *Adaptive Behavior*, *20* (5), 388-404. 10.1177/1059712312451604

Rizzolatti, G. & Luppino, G. (2001). The cortical motor system. *Neuron*, *31* (6), 889 - 901. 10.1016/S0896-6273(01)00423-8

Rosenthal, D. M. (2002). How many kinds of consciousness? *Consciousness and Cognition*, *11* (4), 653-665. 10.1016/S1053-8100(02)00017-X

Russell, S. & Norvig, P. (2003). *Artificial intelligence: A modern approach.* Englewood Cliffs, NJ: Prentice-Hall.

Schier, E. (2009). Identifying phenomenal consciousness. *Consciousness and Cognition*, *18* (1), 216-222. 10.1016/j.concog.2008.04.001

Schilling, M. (2011). Universally manipulable body models - Dual quaternion representations in layered and dynamic (MMCs). *Autonomous Robots*, *30* (4), 399-425. 10.1007/s10514-011-9226-3

Schilling, M. & Cruse, H. (2007). Hierarchical MMC networks as a manipulable body model. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2007), Orlando, FL* (pp. 2141-2146). Orlando, FL. 10.1109/IJCNN.2007.4371289

——— (2008). The evolution of cognition - From first order to second order embodiment. In Wachsmuth, I. and Knoblich, G. (Eds.) *Modeling Communication with Robots and Virtual Humans* (pp. 77-108). Springer, GER: Springer.

——— (2012). What's next: Recruitment of a grounded predictive body model for planning a robot's actions. *Frontiers in Cognition*, *3* (383), 10.3389/fpsyg.2012.00383

Schilling, M. & Cruse, H. (submitted). reaCog, a minimal cognitive controller based on recruitment of reactive systems.

Schilling, M., Schneider, A., Cruse, H. & Schmitz, J. (2008). Local control mechanisms in six-legged walking. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2008* (pp. 2655-2660). 10.1109/iros.2008.4650591

Schilling, M., Paskarbeit, J., Schmitz, J., Schneider, A. & Cruse, H. (2012). Grounding an internal body model of a hexapod walker - Control of curve walking in a biological inspired robot. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012* (pp. 2762-2768). 10.1109/iros.2012.6385709

Schilling, M., Hoinville, T., Schmitz, J. & Cruse, H. (2013a). Walknet, a bio-inspired controller for hexapod walking. *Biological Cybernetics*, *107* (4), 397-419. 10.1007/s00422-013-0563-5

Schilling, M., Paskarbeit, J., Hoinville, T., Hüffmeier, A., Schneider, A., Schmitz, J. & Cruse, H. (2013b). A hexapod walker using a heterarchical architecture for action selection. *Frontiers in Computational Neuroscience*, *7*. 10.3389/fncom.2013.00126

Schmitz, J., Schneider, A., Schilling, M. & Cruse, H. (2008). No need for a body model: Positive velocity feedback for the control of an 18DOF robot walker. *Applied Bionics and Biomechanics*, *5* (3), 135-147. 10.1080/11762320802221074

Schneider, W. (2013). Selective visual processing across competition episodes: a theory of task-driven visual attention and working memory. *Philosophical Transansaction of the Royal Society of London B: Biological Sciences*, *368* (1628), 20130060. 10.1098/rstb.2013.0060

Schneider, A., Paskarbeit, J., Schäffersmann, M. & Schmitz, J. (2011). Biomechatronics for of embodied intelligence an insectoid robot. *Proc. ICRA 2* (pp. 1-11). 10.1007/978-3-642-25489-5_1

Seth, A. (2007). Models of consciousness. *Scholarpedia 2, 1328*, *2* (1), 1328. 10.4249/scholarpedia.1328

Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, *18*, 555-586. 10.1146/annurev.ne.18.030195.003011

Spranger, M., Höfer, S. & Hild, M. (2009). Biologically inspired posture recognition and posture change detection for humanoid robots. *Proceedings of ROBIO'09: IEEE International Conference on Robotics and Biomimetics.* (pp. 562-567). 10.1109/ROBIO.2009.5420708

Steels, L. (1995). Intelligence - Dynamics and Representations. In Steels, L. (Ed.) (pp. 72-89). New York, NY: Springer. 10.1007/978-3-642-79629-6_4

——— (2003). Intelligence with representation. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 361* (1811), 2381-2395. 10.1098/rsta.2003.1257

——— (2007). The symbol grounding problem is solved, so what's next? In De Vega, M., Glennberg, G. and Graesser, G. (Eds.) *Symbols, embodiment and meaning.* New Haven, CT: Academic Press.

Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28* (04), 469-489. 10.1017/S0140525X05000087

Steels, L. & Spranger, M. (2008). The robot in the mirror. *Connection Science*, *20* (4), 337-358. 10.1080/09540090802413186

Tomasello, M. (2009). *Why we cooperate.* Cambridge, MA: MIT Press.

Valdez, P. & Mehrabian, A. (1994). Effects of color on emotions. *Journal of Experimental Psychology: General*, *123* (4), 394-409. 10.1037/a0031821

Vision, G. (2011). *Re-emergence. Locating conscious properties in a material world.* Cambridge, MA: MIT Press.

von Kleist, H. (1810). Über das Marionettentheater. In Sembdner, H. (Ed.) *Heinrich von Kleist, Sämtliche Werke und Briefe, Bd. 2* (p. 345). München, GER: Deutscher Taschenbuch Verlag (originally appeared in Berliner Abendblättern, 1. Jg., 1810).

Wundt, W. (1863). *Vorlesung über die Menschen- und Tierseele.* Leipzig, GER: Voss Verlag.

Yang, Z., Bertolucci, F., Wolf R. & Heisenberg M. (2014). Flies cope with uncontrollable stress by learned helplessness. *Current Biology*, *23* (9), 799-803. 10.1016/j.cub.2013.03.054

Zylberberg, A., Dehaene, S., Roelfsema, P. R. & Sigman, M. (2011). The human turing machine: A neural framework for mental programs. *Trends in Cognitive Sciences*, *15* (7), 293-300. 10.1016/j.tics.2011.05.007

The “Bottom-Up” Approach to Mental Life

A Commentary on Holk Cruse & Malte Schilling

Aaron Gutknecht

With their “bottom-up” approach, Holk Cruse and Malte Schilling present a highly intriguing perspective on those mental phenomena that have fascinated humankind since ancient times. Among them are those aspects of our inner lives that are at the same time most salient and yet most elusive: we are conscious beings with complex emotions, thinking and acting in pursuit of various goals. Starting with, from a biological point of view, very basic abilities, such as the ability to move and navigate in an unpredictable environment, Cruse & Schilling have developed, step-by-step, a robotic system with the ability to plan future actions and, to a limited extent, to verbally report on its own internal states. The authors then offer a compelling argument that their system exhibits aspects of various higher-level mental phenomena such as emotion, attention, intention, volition, and even consciousness.

The scientific investigation of the mind is faced with intricate problems at a very fundamental, methodological level. Not only is there a good deal of conceptual vagueness and uncertainty as to what the explananda precisely are, but it is also unclear what the best strategy might be for addressing the phenomena of interest. Cruse & Schilling’s bio-robotic “bottom-up” approach is designed to provide answers to such questions. In this commentary, I begin, in the first section, by presenting the main ideas behind this approach as I understand them. In the second section, I turn to an examination of its scope and limits. Specifically, I will suggest a set of constraints on good explanations based on the bottom-up approach. What criteria do such explanations have to meet in order to be of real scientific value? I maintain that there are essentially three such criteria: biological plausibility, adequate matching criteria, and transparency. Finally, in the third section, I offer directions for future research, as Cruse & Schilling’s bottom-up approach is well suited to provide new insights in the domain of social cognition and to explain its relation to phenomena such as language, emotion, and self.

Keywords
Bio-robotics | Bottom-up approach | Emergence | Evolution | Explanation | Mechanisms | Robotics | Social cognition

Commentator

Aaron Gutknecht
aaron-gutknecht@gmx.de
Johann Wolfgang Goethe-Universität
Frankfurt a. M., Germany

Target Authors

Holk Cruse
holk.cruse@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Malte Schilling
malte.schilling@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Biorobotics and the bottom-up approach to mental life

From my perspective, there are two basic ideas underlying the overall research strategy entertained by Cruse and Schilling. The first is that in order to understand a system and its properties, it has to be *reinvented* or *reconstructed* by the researcher. The second is that mental phenomena may arise as *emergent* properties via the interaction of low-level components of a system. I’d like to first provide an outline of these basic ideas and the underlying strategy as I understand them. In the next section, I will critically evaluate what types of questions the approach is best suited to answer, and what kind of problems it will likely face.

The first of the two main ideas is central to the research area of bio-robotics. If we are able to create an artificial system that exhibits the phenomena of interest, we should be a great deal closer to understanding how these phenomena come about in nature. In order for this approach to lead to valid conclusions, however, the process of reconstruction has to do justice to the systems we are seeking to understand. In the present context we are concerned, above all, with humans and other animals. This means that the way the artificial system achieves the desired results has to be biologically plausible, i.e., it has to be reasonable to assume that the capacities of the organism that we are trying to understand are really based on similar mechanisms. In this vein, Cruse & Schilling (this collection) are realising the basic reactive modules of their system in form of artificial neural networks that were inspired by biological research on, for instance, stick insects (Walknet) and desert ants (Navinet).

The second of the basic ideas derives its plausibility from an evolutionary perspective on psychological faculties. Emotion, attention, the ability to plan future actions, and any other "higher-level" capacities, including consciousness, did not arise suddenly from one generation to the next and independently of pre-existing, more fundamental abilities, such as the ability to control one's own body and respond adaptively to environmental stimuli. Rather these latter abilities and the interactions between the mechanisms responsible for them might well have been crucial for mental properties to evolve. From this perspective, the idea of reconstructing the evolutionary process by starting with basic reactive structures and examining how through the interaction of these structures unexpected properties might *emerge* seems very promising. Since humans also gradually evolved from simpler organisms, it is natural to assume that the same dependence between reactive structures and "higher-level" phenomena is present in our case as well. The investigation of this dependence might thus provide new insights into the mechanisms underlying human psychology.

But what does it mean exactly to say that a property *emerges* from basic structures? What is an emergent property? The philosophical controversies surrounding the concept of emergence date back over a hundred years and although usage of the term has become increasingly popular in recent years, among both philosophers and scientist, it can hardly be said to have one universal definition. Rather, there are numerous and varied interpretations, a fact which inevitably leads to confusion and misunderstanding (for a good overview see O'Connor & Wong 2012). It is thus vital to identify precisely what is meant by emergence in any particular case. Notwithstanding this inherent ambiguity, there seems, however, to be a shared idea behind much talk of emergent properties: this is the idea that as systems become increasingly complex they tend to exhibit certain higher-level properties, which are novel or unexpected given their simpler, lower-level, components.

Depending on how this claim is interpreted it can have more or less serious implications regarding the fundamental structure of nature, as well as the structure of science. In order to obtain a particularly strong and at the same time highly influential reading, it must be understood in a two-fold sense. First, as meaning that these properties cannot *even in principle* be predicted or explained on the basis of the lower-level properties of the system and, second, as indicating that such properties are associated with genuinely *new causal powers*, i.e., they make a real difference to the run of events and are not mere epiphenomena (for discussion see Kim 1999, 2006).[1] This kind of emergence could be called *strong emergence.*[2] Central to this conception is that emergent properties causally influence the simpler entities from whose organisation they emerge. This sort of causal influence is called "downward causation", as emergent properties are conceived as

[1] Such conceptions go back to thinkers such as Samuel Alexander, C. L. Morgan, and C. D. Broad, prominent figures in a philosophical movement, which came to be known as "British emergentism". The following discussion is, however, intended to illustrate the problematic nature of the concept of emergence and not to offer an analysis of the ideas of a particular philosophical school.

[2] It should be noted that the there is no universal definition of the term "strong emergence" in the current literature (for some alternative characterisations see Chalmers 2006; Bedau 1997; Yates 2013).

higher-level properties arising from certain lower-level properties and relations. Typically, it is assumed that what we find at the lowest level of this hierarchy are the properties and relations of fundamental physical particles. Given this assumption, the existence of emergent properties would entail that a complete description of the fundamental physical organisation of a system might still leave something out. The system might still have some properties that could not be predicted on the basis of such a description and could not be explained in terms of the organisation of its basic physical constituents. Moreover, because emergent properties are causally efficacious, knowledge of the basic physical components of a system and their behaviour may not be sufficient to predict the future evolution of the system. These considerations seem to lead to the conclusion that the meta-scientific thesis, according to which all phenomena can ultimately be explained by the fundamental laws of physics, would turn out to be false. If certain properties belonging to the domains of psychology, biology, or chemistry were emergent properties, these could not even in principle be captured by basic physics alone. All sciences dealing with genuinely emergent properties would remain completely autonomous, positing their own independent laws and explanations. Furthermore, since emergent properties have the ability to causally influence lower-level entities, the fundamental laws of physics would not even suffice to explain processes taking place at the *physical* level (see also Chalmers 2006).

These are substantial conclusions that could be met with some scepticism. They are also one of the reasons for the fierce controversy surrounding the concept of emergence. Furthermore, the condition that emergent properties are themselves causally efficacious and the general idea of "downward causation" leads to problems in and of itself. This is because there has to be a systematic relationship between emergent and lower-level properties, even though they are conceived as being distinct from another. Often this is expressed by saying that emergent properties are completely determined by lower-level properties and require them for their existence. In other words, if all lower-level properties of a system are fixed, its emergent properties are also fixed; and without any appropriate lower-level properties, a system cannot have emergent properties. If this weren't the case, it would be unclear in what sense emergent properties *emerge from* lower-level ones (Kim 2006). If their relationship were completely coincidental, this would surely be an inappropriate description.

Based on this requirement, Kim (1999, 2006) has put forth an influential argument that the idea of "downward causation" is untenable. In summary, Kim's basic argument is this: suppose an emergent property (let's say a feeling of thirst) causes a lower-level property (e.g., a certain activation pattern N in the brain). If feeling thirsty is an emergent property, there have to be appropriate lower-level properties from which it emerges. Let's call these the "emergence base" of feeling thirsty. Now, that feeling thirsty causes N means that there is a natural law that occurrences of feeling thirsty are always followed by occurrences of N (feeling thirsty is nomologically sufficient for N). But since occurrences of feeling thirsty are always accompanied by occurrences of its emergence base, it must also be true that occurrences of its emergence base are followed by occurrences of N. Therefore, if feeling thirsty causes N, its emergence base also causes N. But this makes feeling thirsty completely redundant as a cause of N. Its emergence base is completely sufficient to explain Ns occurrence, leaving the feeling of thirst as a mere epiphenomenon. Since this example can easily be generalised, one can conclude that there are no genuine cases of downward causation and hence no genuine emergent properties of the type presently under consideration.

In summary, it can be stated that emergence is a highly controversial concept—not only because of its inherent ambiguity, but also on account of certain varieties of emergentism that have substantial metaphysical and meta-scientific implications as well as a commitment to the problematic idea of downward causation. The crucial questions remaining now are whether Cruse & Schilling (this collection) provide a clear interpretation of the concept of emergence and whether it provokes the kind of controversy and criticism outlined above. What kind of emergence is involved in their claim that mental states might be construed as emergent properties? In fact, they provide two slightly different characterisations. According to the first, an emergent property is to be understood as a property of a whole system that cannot, *at first sight,* be traced back to the interactions of the systems components. Alternatively, one might say that we cannot, at first sight, predict the emergent properties of a complex system based on our knowledge of its parts and their interaction. Thus, we might be genuinely surprised that the system in question exhibits such properties. Emergence in this sense is sometimes called *weak emergence* (Chalmers 2006). If this is all that it means for a system to have emergent properties, few would raise serious objections. This sort of emergence is just a consequence of our limited knowledge and cognitive capacities and is relative to the judging subject: what might not be immediately predictable for one person might be just so for another. Emergentism, in this sense, has no far-reaching metaphysical or meta-scientific implications and is not committed to any sort of "downward causation".

Cruse & Schilling (this collection) provide a second, and equally unproblematic, definition of emergence that is specifically tailored for application in the context of robotics. According to that definition, a property of an artificially constructed system is emergent if it was not explicitly implemented by its designers. We might call this *implementational emergence.* This appears to be relatively independent of the sort of "weak" emergence I've just described. Even a property not explicitly implemented might be predictable without too much effort, whereas a property deliberately implemented might not be predictable, at least by persons lacking experience or competence. I think that most of the emergent properties Cruse & Schilling (this collection) attribute to their artificial system, reaCog, match both characterisations: they were neither explicitly implemented nor would we immediately expect or predict that reaCog would exhibit them. At the same time, the properties in question are highly interesting and are not simply insignificant side effects. This is important since, according to the definitions provided by Cruse & Schilling, the claim that an artificial system exhibits emergent properties is, *in and of itself,* not particularly notable. But this depends entirely on what the emergent properties in question precisely are. The finding that reaCog exhibits, in this way, aspects of psychological characteristics, such as emotion or attention and the ability to perform non-trivial body movements, are most certainly of considerable scientific significance. In conclusion, we may say that although the kind of emergentism advocated by Cruse & Schilling does not have the same far-reaching implications as the particularly demanding conception outlined above, it is nonetheless useful and philosophically interesting. This is because it functions as the basis of an intriguing approach to the study of psychological properties, which I shall now endeavour to describe.

Combining the idea of emergence with the idea, outlined above, that in order to understand a system and its properties, it has to be reinvented or reconstructed, we arrive at a fascinating research strategy. The first step consists in observing the behaviour of animals that, although lacking many of the sophisticated abilities with which humans are endowed, are nonetheless capable of flexibly controlling their bodies in order to cope with an unpredictable environment (such as stick insects, desert ants, and honey bees). Based on these observations one then develops a neural network model (e.g., Walknet or Navinet) designed to produce the behaviour observed in the first step. Next, this model is realised in an artificial system (either virtual or robotic) in order to examine to what extent the behaviour produced by the model matches the behaviour of the biological organism on which it is based. If it resembles it to a great extent, this can be taken as *prima facie* evidence that the mechanisms underlying the behaviour are the same for the animal and the robot. Different modules that are constructed in this way are then integrated into a holistic system. Further modules might be added step-by-step (e.g., Body Model, Attention-Controller, Word-Nets). The result is a complex system (in the present case “reaCog”) the behaviour and properties of which cannot be easily predicted even by its very own designers. The last, and most important step consists of examining whether the system shows characteristics that were not explicitly implemented but instead arise from the dynamic interactions of the system’s components. The most intriguing question in this context is, of course, whether the final system shows aspects of those phenomena that are constitutive of *having a mind*.

Although this is only a rough sketch of the methodology entertained by Cruse & Schilling (this collection), I hope I have captured the essential points sufficiently to proceed with an evaluation of its scope and the possible problems it might face. What kind of questions is the bottom-up approach best suited to answer? Which phenomena or processes can be addressed by research based on this approach? What considerations have to be taken into account in order for the presented research strategy to be successful? Are there any general constraints bio-robotic bottom-up explanations have to meet? As we shall see, the answers to these last two question are directly connected to two characteristics of the research strategy outlined in the previous paragraph: first, that it involves, at two points, a comparison of the behaviour of significantly different systems and, second, that it is specifically designed to discover emergent properties.

2 The bottom-up approach: Objectives, benefits and constraints

2.1 Mechanisms and the evolution of the mind

The most important aspect of the proposed approach is that it helps to elucidate the *mechanisms* underlying various mental properties. This is possible because many of the basic features of the control system reaCog are known. Using the words of Cruse & Schilling (this collection), it constitutes a “quantitatively defined system”. As all components are realised as artificial neural networks, all information about the number of neurons, the connection weights between them, and the way individual neurons process information is available. More importantly, however, the basic functional architecture of the system is well understood. Which modules are connected in which ways to other modules, how they receive their input, and what other parts of the system might be affected by their outputs does not have to be figured out by painstaking investigation—as is the case in biological research. Because these facts about reaCog are known, it is possible to provide detailed mechanism descriptions. In this way, reaCog’s ability to plan its future actions by internal simulation can be explained by reference to the interaction of its various sub-modules: a problem detector is activated when sensory input indicates that current behaviour will lead, if continued, to adverse effects for the system (e.g., falling over). This leads to the abortion of current behaviour and activation in the Spreading Activation Layer, which randomly excites the Winner-Takes-All network (WTA-net). After some time, the WTA-net adopts a relaxed state in which only one of its units is active. This active unit in turn stimulates its counterpart in the Motivation Unit Network, leading to activity of the corresponding reactive procedures. These provide motor output that can be redirected to the body model, which then simulates the execution of the proposed behaviour and predicts its likely consequences. If the system predicts that the problem will persist, the process of internal simulation goes on until a solution is found, which can then be used to control the actual movements of the system.

Explanations like these contain a lot of information about which functional subparts of a system are engaged during the exercise of the ability in question. In this particular case it makes clear how the ability to plan ahead, a cognitive ability, depends heavily on basic reactive structures that are designed to control specific leg movements as well as an internal model of the body. The same is true for various other capacities like attention and Theory of Mind. Thus, new insights into the mechanisms responsible for those phenomena in humans could be gained by considering how body models and motor control mechanisms are realised in our case and how these systems interact. In other words, the bottom-up approach may lead to new directions for future research concerning human psychology by suggesting how specific functional modules interact in order to bring about a particular target phenomenon. Whether this approach is tenable depends on the degree to which findings pertaining to the artificial system might legitimately be used to draw conclusions about human beings. I will propose a number of constraints to ensure that this condition is fulfilled below.

Another class of questions that a bottom-up strategy is well designed to answer has to do with the evolution of cognitive capacities: how did cognitive systems evolve from purely reactive systems? How did emotions, attention, or even consciousness arise? What are the natural precursors of these phenomena? Cruse & Schilling (this collection) show convincingly that no completely new neural modules are needed in order for such properties to occur. Rather, minor changes in the basic architecture might suffice to generate radical extensions of the abilities of a system. In this way, a reactive system with a body model can acquire the ability to plan ahead if it is able to disconnect its motor system from the physical body and instead send the motor signals to its internal body model. No novel "planning module" is needed. Already existing modules just have to become dissociable and can thus acquire new functions (Cruse 2003). In addition, the target paper suggests an answer to the question of the evolutionary function of cognition understood as the ability to plan ahead: it was the necessity of being able to control a complex body in a complex environment that made this ability highly valuable. Detecting problems by perception, finding innovative solutions by internal simulation and acting on them are capacities that are extremely advantageous for any organism possessing a body with a high number of redundant degrees of freedom (see Cruse 2003). This is in line with, and actually extends, the widespread assumption that the evolutionary function of cognition is to deal with environmental complexity (Godfrey-Smith 2002).

2.2 Constraints on bio-robotic bottom-up explanations

In the previous paragraph we saw that the framework Cruse & Schilling (this collection) present is well-equipped to give new insights into the underlying mechanisms of psychological phenomena and the evolution of cognition, as well as a promising approach to creating highly flexible and intelligent robots. There are, however, some problems the proposed strategy has to face, especially if the control structures become increasingly complex. I therefore want to suggest a set of three constraints on good bottom-up explanations of biological/psychological phenomena.

1. *Adequate matching criteria:*[3] At two points the research strategy described in section 1 involves a comparison between the behaviour of an artificial system on the one hand and a biological system on the other. First, this is the case in the development of neural network models of animal behaviour. In this context, the comparison is used to ascertain whether the proposed model of the mechanisms underlying certain capacities (e.g., walking) really reproduces the original behaviour of the animal (e.g., a stick insect). Second, there is a similar process of comparison involved in the application of psychological concepts to the complete system. At different points in their discussion, Cruse & Schilling (this collection) argue that their system has certain mental capacities because it exhibits behaviour (or would exhibit it if certain extensions were implemented) connected to those mental capacities in humans. So, for example, just as the performance of athletes might worsen if they consciously attend to what they are doing, the activation of the attention controller in reaCog can lead to poorer results compared to unimpeded execution of the reactive procedures.

   Both processes of comparison require criteria to identify when the behaviour of the artificial system and that of the biological system

[3] I credit this term to Datteri & Tamburrini 2007.

are relevantly similar, i.e., similar enough in order to provide evidence for the claim that similar mechanisms are at work in both cases or that the artificial system and the biological system share certain psychological characteristics (Datteri & Tamburrini 2007). The difficulty of finding such criteria increases the more the bodies of the compared systems differ. In some cases they might nonetheless be easy to find and relatively uncontroversial. This, however, is not always the case. For instance, in their discussion of emotions—and more specifically the emotion of happiness—, Cruse & Schilling (this collection) suggest that by increasing the threshold of the problem detector reaCog would take more risks, thus behaving similarly to humans when they are happy. Now, the question is whether the kind of risky behaviour exhibited by reaCog when the threshold of its problem detector is increased is the same kind of risky behaviour humans exhibit when they are happy. Only if this condition is fulfilled can the similarity be taken as evidence that reaCog shows aspects of the emotion of happiness.

2. *Biological plausibility*: Any proposed mechanism should be biologically plausible, i.e., it has to be reasonable to assume that the capacities of the organism that we are trying to understand are really based on such a mechanism. This can, at least to some degree, be ensured by trying to create similarities between the artificial and the biological organism on a basic structural level, for example by using artificial neural networks. Furthermore, it is necessary to decide how fine-grained the model should be. Should the model take brain structures, neurons, or subcellular elements as its basic building blocks? Should intracellular processes be neglected or are they important? The answer will of course always be relative to our particular epistemic goals. Finally, there are different options regarding the way artificial neurons process information, i.e., how they calculate their output value depending on the weighted sum of their inputs. All these factors might turn out to be important if the results are to be used to infer biological mechanisms. The requirement of biological plausibility shouldn’t, however, be overemphasised. Cruse & Schilling (this collection) stress that they are not trying to present a realistic model of neuronal activity in living organisms. Hence, they are using biologically implausible, non-spiking artificial neurons as the basic elements of their architecture, while noting that some authors (referring to Singer 1995) have located the neural basis of consciousness in synchronously oscillating spikes. This, however, is not a weighty objection to the proposed approach since it is designed as a *functional approach*. The question is: how do different functional subsystems like a system for controlling the swing-movement of a leg, a system modelling the robot’s body, and a system allowing for the selection of different internal states interact in order to produce certain emergent phenomena? Therefore, the concrete physical realisation of these subsystems is of only secondary importance.

3. *Transparency:*[4] Doubts about the strategy of using artificial systems in order to understand biological systems arise because even if we were to create an extremely intelligent robot, it would not necessarily help us to understand the mechanisms underlying its intelligence. Rather, we might be faced with yet another complex system whose workings we do not understand (Holland & Goodman 2003). Now, the approach Cruse & Schilling (this collection) present is specifically designed to discover emergent properties, i.e., properties that were not explicitly implemented. This means that there will be a high risk of finding properties in the complete system that cannot be readily provided with a clear-cut mechanistic explanation involving the cooperation of the system’s components. Although the explanations of the occurrence of various psychological properties presented in the present paper are quite convincing, the

[4] The concept of transparency has a number of other well-established interpretations in the literature that should not be confused with the one at issue in the present context. These include, for example, “semantic” (Clark 1989) and “phenomenal” (Metzinger 2004) transparency.

bottom-up strategy might eventually exhaust its potential if the complexity of the system is further increased.

3 Future perspectives: The social insect

I would like to conclude by briefly proposing a perspective for future research based on the system reaCog. As presented, its ability to interact and cooperate with other agents is fairly restricted. At the same time, the pre-requisites of a broader social extension of the system seem to be in place. The present paper already shows how reaCog could be equipped with the capacities to recognize the behaviour of others and apply a Theory of Mind. In their 2011 paper, Cruse & Schilling further propose that by implementing a two-body model (a "We-model") reaCog might be capable of cooperative behaviour using shared goals. Integration and further expansion of such social capacities, and their application in an actual robot, seems promising considering the importance of social interaction in processes such as language acquisition and emotional regulation. Some have even suggested that the presence of other agents in the environment, or, in other words, dealing with social complexity, was a dominant factor in the evolution of sophisticated cognitive abilities (Humphrey 1976). Thus bio-robotic research in this direction might provide new insights into the mechanisms underlying such developmental and evolutionary processes. Moreover, a social extension of reaCog might eventually shed light on potential emergent phenomena *on a group level*, such as labour division, collective planning, social hierarchies and, most fundamentally, joint action coordination. What high-level social phenomena emerge when multiple bio-robotic systems like reaCog interact with each other?

Cruse and Schilling's system seems particularly well-suited to further illuminate motor theories of social cognition. According to such theories, the important social cognitive capacity of understanding another's actions is directly linked to mechanisms that are active when the observer performs similar actions Gallese et al. 2004; for criticism see Jacob & Jeannerod 2005). The underlying neural mechanism has come to be known as the mirror-neuron system. Furthermore, there is evidence that the mirror-neuron system plays a role in certain aspects of self-consciousness. For instance, Uddin (2007; see also Molnar-Szakacs & Uddin 2013) suggests that this is the case for representations of the physical self, and ascribes frontoparietal mirror-neuron areas an important function for self-recognition (especially the recognition of one's own face). As mirroring mechanisms can be integrated in reaCog as well, this opens the possibility of further investigating motor theories of social cognition and the relation between internal motor simulation and the self in a quantitatively defined system.

An ability that is highly important for human social interaction is the ability to communicate using language. At this point, the linguistic capacities of reaCog still seem quite inflexible and limited in scope. A highly interesting extension of this system would be to provide it with the means to learn words and their meanings by interaction with other agents. Some of the pre-requisites, like the ability to internally simulate the behaviour of others, could, as Cruse and Schilling argue, be implemented in reaCog by using its internal body-model to represent another agent. Robotic research in this direction was performed by Steels & Spranger (2009). Their artificial systems are capable of autonomously acquiring a simple language consisting of words for specific body postures. After learning is complete, the artificial agents are able to reliably assume body postures on verbal command by other agents. Since social learning has also been implicated in the process of concept formation (Steels 2002), the proposed extension might also foster our understanding of this intriguing phenomenon.

4 Conclusion

In conclusion it can be stated that Cruse & Schilling (this collection) present a highly fascinating research strategy that is well worth pursuing. The bottom-up approach can provide us with new insights regarding the functional mechanisms underlying psychological phenomena and their evolution. Although the notion of emergence is central to it, Cruse & Schilling (this collection) avoid the philosophical controversies surrounding this concept by interpreting it in a less demanding, yet interesting and useful way. There are, however, a number of constraints that explanations based on the bottom-up approach have to meet. First, since Cruse & Schilling's (this collection) strategy involves, at two points, a comparison between markedly different systems, criteria are needed according to which we can determine whether the two systems exhibit relevantly similar behaviour. Second, the structural architecture of the artificial system must have an adequate degree of biological plausibility. And finally, it has to be ensured that increasing the complexity of the system does not lead to the practical impossibility of elucidating the mechanisms underlying its emergent properties.

A promising next step for bottom-up research as presented by Cruse & Schilling (this collection) would be to take it to the level of social interaction. An extensive social extension of their system could shed light on a wide range of intriguing phenomena. Is it possible to discover emergent properties on a group level? In what precise way are mirroring mechanisms involved in social cognition? What role do such mechanisms play for the phenomenon of self-consciousness? What role do reactive structures and internal body-models play in the processes of language acquisition and comprehension? Of course this is only a small selection of the questions further bio-robotic research might contribute to answering. Cruse & Schilling (this collection) made clear that starting from the bottom is a strategy with enormous scientific significance. There is no doubt that this work will make an important contribution to a plethora of research projects in the future.

References

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, *11* (s11), 375-399. 10.1111/0029-4624.31.s11.17

Chalmers, D. J. (2006). Strong and weak emergence. In P. Davies & P. Clayton (Ed.) *The re-emergence of emergence* (pp. 244-256). Oxford, UK: Oxford University Press.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing.* Cambridge, MA: MIT Press.

Cruse, H. (2003). The evolution of cognition - a hypothesis. *Cognitive Science*, *27* (1), 135-155. 10.1207/s15516709cog2701_5

Cruse, H. & Schilling, M. (2011). From egocentric systems to systems allowing for theory of mind and mutualism. In T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo & R. Doursat (Ed.) *Advances in artificial life, ECAL 2011. Proceedings of the eleventh european conference on the synthesis and simulation of living systems* (pp. 185-192). Cambridge, MA: MIT Press.

Clark, A. & Schilling M. (2015). Mental states as emergent properties. In T. Metzinger & J. M. Windt (Ed.) *Open MIND* (pp. 1-39). Frankfurt a. M., GER: MIND Group.

Datteri, E. & Tamburrini, G. (2007). Biorobotic experiments for the discovery of biological mechanisms. *Philosophy of Science*, *74* (3), 409-430. 10.1073/pnas.1015390108

Gallese, V., Keysers, C. & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8* (9), 396-403. 10.1016/j.tics.2004.07.002

Godfrey-Smith, P. (2002). Environmental complexity and the evolution of cognition. In R. Sternberg & J. Kaufman (Ed.) *The evolution of intelligence* (pp. 233-249). Hove, UK: Psychology Press.

Holland, O. & Goodman, R. (2003). Robots with internal models a route to machine consciousness? *Journal of Consciousness Studies*, *10* (4-5), 77-109.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Ed.) *Growing point in ethology* (pp. 303-317). Cambridge, UK: Cambridge University Press.

Jacob, P. & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences*, *9* (1), 21-25.

Kim, J. (1999). Making sense of emergence. *Philosophical studies*, *95* (1), 3-36. 10.1023/A:1004563122154

——— (2006). Emergence: Core ideas and issues. *Synthese*, *151* (3), 547-559. 10.1093/acprof:oso/9780199585878.001.0001

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity.* MIT Press.

Molnar-Szakacs, I. & Uddin, L. Q. (2013). The emergent self: How distributed neural networks support self-representation. *Handbook of neurosociology* (pp. 167-182). Dordrecht, NL: Springer.

O'Connor, T. & Wong, H. Y. (2012). Emergent properties. *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/properties-emergent/

Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, *18* (1), 555-586. 10.1146/annurev.ne.18.030195.003011

Steels, L. & Kaplan, F. (2002). AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, *4* (1), 3-32. 10.1075/eoc.4.1.03ste

Steels, L. & Spranger, M. (2009). How experience of the body shapes language about space. In M. Kaufmann (Ed.) *IJCAI'09: Proceedings of the 21st international joint conference on Artifical intelligence*. San Francisco, CA: Morgan Kaufmann.

Uddin, L. Q., Iacoboni, M., Lange, C. & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, *11* (4), 153-157. 10.1016/j.tics.2007.01.001

Yates, D. (2013). Emergence. In H. Pashler (Ed.) *Encyclopaedia of the Mind* (pp. 283-287). San Diego, CA: SAGE Reference.

The Bottom-Up Approach: Benefits and Limits

A Reply to Aaron Gutknecht

Holk Cruse & Malte Schilling

Aaron Gutknecht supports our bottom-up approach, specifies possible limits and highlights interesting future aspects. His added perspective is valuable and interesting to us. As we fully agree with his view, we only add some complementary remarks.

Keywords
Bottom-up approach | Concept clarifying machine | Emergence | Emotions in arthropods | We-model

Authors

Holk Cruse
holk.cruse@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Malte Schilling
malte.schilling@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Commentator

Aaron Gutknecht
aaron-gutknecht@gmx.de
Johann Wolfgang Goethe-Universität
Frankfurt a. M., Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

We appreciate the comments given by Aaron Gutknecht very much, in particular his discussion and clarification of the term "emergence" and its philosophical background. This discussion comprises a sensible completion of our article going beyond the scope of our expertise. In this context, Aaron Gutknecht correctly states that our way of using the term "emergence" may cover two aspects, one called "weak emergence", the other he addressed as "implementational emergence". We have – possibly forming some kind of common denominator - a third characterization in mind, one that covers different description levels: a phenomenon is considered emergent if it turns out

Cruse, H. & Schilling, M. (2015). The Bottom-Up Approach: Benefits and Limits - A Reply to Aaron Gutknecht.
In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 9(R). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570931

that known properties of the network could also be characterized on a different level of description than the one currently used. On this different level the phenomenon conceptually constitutes a term or definition. If we, for example, describe the structure and function of reaCog on the neuronal level, we may realize at some point that there are behavioral aspects which could, by an outside observer, be characterized by a term that is not defined at a neuronal level of description, such as, for example, “intention”.

2 The bottom-up approach

This way of using the term emergence is directly related to the bottom-up approach applied here. This approach is inspired by Feynman, who stated that we understand a system only when we are able to construct it (in Hawking 2001) and may be even dated back to Giambattista Vico (1710). The bottom-up approach allows us to study the extent to which linguistic concepts proposed in the literature may correspond to properties realized by our artificial system. If one was not prepared to accept that a specific concept would correspond to selected properties of the artificial system, either the linguistic concepts might be adapted accordingly, or the artificial system might be judged as to show deficits. The latter case could then give rise to adapt the current simulation model to better match the verbal proposal given. This capability of the bottom-up approach led Manuela Lenzen (2014) to characterize reaCog as a “concept clarifying machine” (“Begriffspräzisierungsmaschine”).

3 Possible limits of the bottom-up approach

Aaron Gutknecht further proposes a well-chosen list of issues that should be taken into account when following a bottom-up approach as proposed here, namely “adequate matching criteria”, “biological plausibility” and “transparency”.

Concerning the first issue, “adequate matching criteria”, Aaron Gutknecht addresses a possibly critical point. In section 8 (*Emotions*), we characterize happiness by the property that risky decisions are made more probable. We admit that our example is formulated in a sketchy way, only addressing one basic aspect for illustration. There are, however, more deeply founded examples that have been briefly referred to in the main text and will be explained in more detail here. Two recent studies, one in crayfish, the other in the fruitfly, provide strong hints that emotion-like states can be found in simple organisms as arthropods or, more specifically in the latter-case, insects. In crayfish, Fossat et al. (2014) have convincingly shown that context-independent, anxiety-like behavior can be induced by experimentally applied stress or by application of serotonin. Both methods lead the animals to avoid illuminated sections of their environment which they are normally interested to explore. Anxiety is related to fear but considered a secondary emotion that occurs after the stressing signal has disappeared. Thus, the probability of selecting specific behaviors, in this case exploration of illuminated places, is decreased. This avoidance behavior could be abolished after application of drugs that are known to have anxiolytic effects in mammals. Applied to reaCog, these results could be interpreted in the following way. Emotion-like states would not only influence the global WTA net, but also thresholds of local, lower level WTA networks that are responsible for switching between different procedures.

Another interesting case has been reported by Yang et al. (2014) in Drosophila. These animals learnt that various behaviours selected in trying to avoid a problem, in this case escape from a heated ground, were not successful. As a consequence, they ended up in a state of passivity. This result has been discussed as an example of “learned helplessness”, which is considered an animal model of depression. In our framework, this could simply be realized by freezing activity in the Spreading Activation Layer network that provides input to the WTA net (section 4).

Concerning the second issue, “biological plausibility”, we fully support Gutknecht’s perspective and have only a minor aside. Application of non-spiking neurons is not necessarily biologically implausible. Rather, non-spiking neurons do exist in invertebrate and in vertebrate brains. They play important functional roles, but are generally less well-known, mainly because investigating them involves methodological problems that are more difficult than those of spiking neurons. The third issue, “transparency”, addresses the view that the bottom-up strategy may eventually exhaust its potential when the complexity of the system is further increased. Although we agree with Gutknecht here, we would like to add that the bottom-up approach still bears the advantage that, as the details of such a system are known, its properties can be thoroughly analyzed by physical and/or mathematical methods. This ability, of course, does not guarantee that one will find answers in such a hypothetical case, but there are various methods available to address such questions. Further, we believe that the problem of lacking transparency may not happen to occur too often. This belief is supported by the observation that already our simple system, reaCog, appears to be able to reach integration levels characterized by terms such as intention, volition and consciousness.

4 What should be done next?

Aaron Gutknecht closes his comments by considering future aspects. Again, we agree with his recommendations and have, partly, indeed started with two of the aspects addressed. We applied the internal model in a cooperative scenario in which the visual impression of another agent performing an action was mapped onto the system’s own internal body model. In this way the internal model was driven by the visual input and the internal model reenacted what the other agent was doing. This mapping allows one to connect the experiences of somebody else to one’s own action repertoire as one steps into the shoes of the other (Schilling 2011; see also Gallese & Cuccio this collection). Second, as mentioned in the main text, shared circuits are required for an agent to represent the action of a partner (Cruse & Schilling this collection, figure 9). In order to allow for ToM, an additional separate representation of the partner’s memory is required (figure 10). To be able to apply a supermodel (or we-model, Tomasello 2009), a more complex model is required (see Cruse & Schilling 2011, figure 6).

5 Conclusion

The bottom-up approach advocated here to understand higher-level phenomena may be considered a non-Platonic approach that aims to construct artificial, but strongly biologically inspired systems. These systems should be able to simulate complex behavioral tasks, but do so by application of simple elements, artificial neurons, and a simple decentralized neuronal architecture. If successful one could then study whether more abstract concepts introduced in psychology or philosophy, for example, could sensibly be applied to such a system. We claim to have shown an example supporting this approach.

References

Cruse, H. & Schilling, M. (2011). From egocentric systems to systems allowing for theory of mind and mutualism. In R. Doursat (Ed.) *Proceedings of the ECAL 2011, Paris* (pp. 731-738). Cambridge, MA: MIT Press.

——— (2015). Mental states as emergent properties. In T. Metzinger & J. M. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Fossat, P., Bacqué-Cazenave, J., De Deurwaerdère, P., Delbecque, J.-P. & Cattaert, D. (2014). Anxiety-like behavior in crayfish is controlled by serotonin. *Science, 344* (6189), 1293-1297. 10.1126/science.1248811

Gallese, V. & Cuccio, V. (2015). The paradigmatic body. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-23). Frankfurt a. M., GER: MIND Group.

Hawking, S. (2001). *The universe in a nutshell.* London, UK: Bantam Press.

Lenzen, M. (2014). Der sensible Hector - Interaktion mit Robotern. *Frankfurter Allgemeine Zeitung* (2043.9.2014, p.N3)

Schilling, M. (2011). Learning by seeing - Associative learning of visual features through mental simulation of observed action. In R. Doursat (Ed.) *Proc. of the ECAL 2011, Paris* (pp. 731-738). Cambridge, MA: MIT Press.

Tomasello, M. (2009). *Why we cooperate.* Cambridge, MA: MIT Press.

Vico, G. (1710). De antiquissima italorum sapientia. In R. Parenti (Ed.) *Opere.* Naples, I: F. Rossi.

Yang, Z., Bertolucci, F., Wolf , R. & Heisenberg , M. (2014). Flies cope with uncontrollable stress by learned helplessness. *Current Biology, 23* (9), 799-803. 10.1016/j.cub.2013.03.054